TFFR-Net: Resolving Time-Frequency Ridge Crossings via Frequency-Rate Lifting

NeurIPS 2026

Pingping Pan[1] You Li[1]* Yunjian Zhang[2]* Mujiangshan Wang[2]

[1] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China

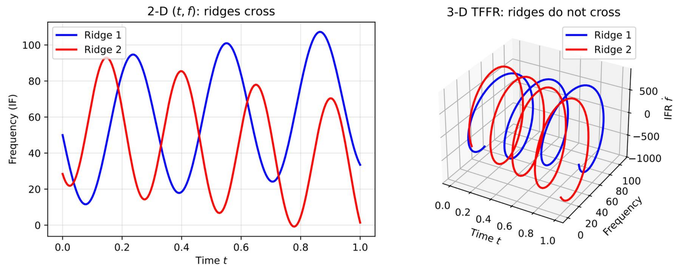
[2] Shenzhen Kaihong Digital Industry Development Co., Ltd. and Harbin Institute of Technology (Chongqing), China

1 Motivation & frequency-rate lifting

Ridge crossings create ambiguity in the 2-D time-frequency plane: different components share the same (t, f) pixel. Peak tracking and 2-D segmentation may swap or merge their identities.

$$\mathrm{IF}(t) = d\varphi(t)/dt \qquad \mathrm{IFR}(t) = d\mathrm{IF}(t)/dt$$

Lift each ridge to Γ(t) = (t, IF(t), IFR(t)). At the same IF, different frequency rates separate overlapping components.



Why a third coordinate helps

For generic chirps, simultaneous IF-and-IFR coincidence adds an independent constraint. Under continuous parameter distributions, this exceptional set has measure zero. Finite grids can still leave near-coincident ridges unresolved.

69.00% in 2-D → 6.12% in 3-D

Unresolved trials: 3,449/5,000 vs. 306/5,000. Monte Carlo: K = 2–6, N = 256. Resolution thresholds: εf = fs/N and εr = fs²/N².

2 Benchmark results

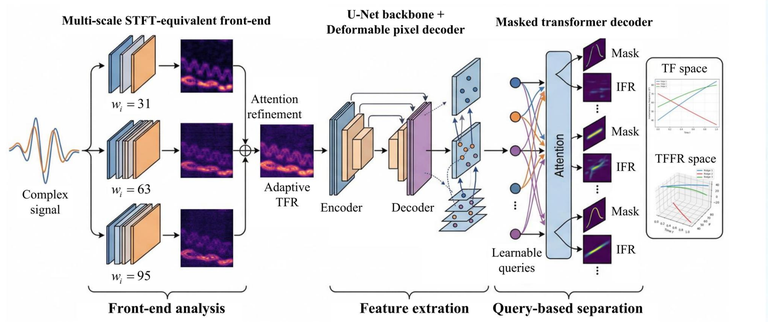
Synthetic chirps: K = 1–6 components, 0–25 dB SNR.
Errors use Hungarian matching. Lower is better.

IF MAE (Hz)						
SNR	ACMD	VNCMD	IVGNMD	RPRG	TFD-Net	Ours
0 dB	20.22	19.55	31.54	19.33	9.45	**8.87**
5 dB	18.12	15.01	34.07	14.86	6.05	**5.70**
10 dB	15.06	12.91	25.75	12.77	4.03	**3.62**
15 dB	12.52	11.55	24.74	11.76	3.14	**2.56**
20 dB	10.93	11.13	23.44	11.32	2.40	**2.02**
25 dB	10.00	10.98	23.14	11.17	1.96	**1.66**

RPRG denotes RPRG+ICCD.

At 25 dB vs. TFD-Net: IF MAE 1.66 Hz (15.3% lower), Dice loss 0.198 (20.2% lower), IF Std 1.97 Hz (26.8% lower).

3 TFFR-Net architecture



- Learnable analysis: Three complex convolutions (31, 63, 95) adapt the window. Fourier-constrained kernels recover STFT magnitudes as a special case.

  Temporal features: A U-Net fuses B with its temporal gradient. A 1×1 convolution and deformable pixel decoder fuse four scales.
- Instance decoding: Q = 10 queries and L = 4 transformer layers predict a ridge mask, IFR map and activity score.

IFR feedback is the key mechanism

$$\bar{r}_i = \frac{\sum_{f,t} \sigma(\hat{M}_i)_{f,t} \cdot \hat{R}_{i,f,t}}{\sum_{f,t} \sigma(\hat{M}_i)_{f,t} + \epsilon}, \qquad \mathbf{q}_i^{\mathrm{pos}} := \mathbf{q}_i^{(0)} + \mathrm{MLP}(\bar{r}_i),$$

Each layer projects the mask-weighted mean IFR into query position. Different rate offsets guide attention through overlapping TF masks.

Training objectives

- Hungarian matching with ±1-pixel frequency alignment assigns queries to ground-truth ridges. Unmatched queries are inactive.

$$L = 2.0\, L_{ins} + 1.0\, L_{ifr} + 0.1\, L_{cls}$$

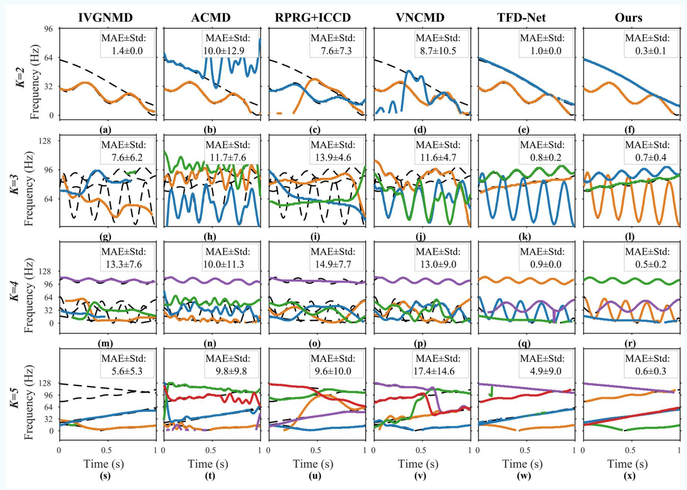
  Lins = 0.2 Dice + 0.8 BCE. Lifr uses smooth-L1 on GT ridge pixels. Lcls balances active and inactive queries.
- F = T = 256. Hidden 256, 2 heads. RMSprop, LR 1e-4, batch 8.

Ablation at 25 dB

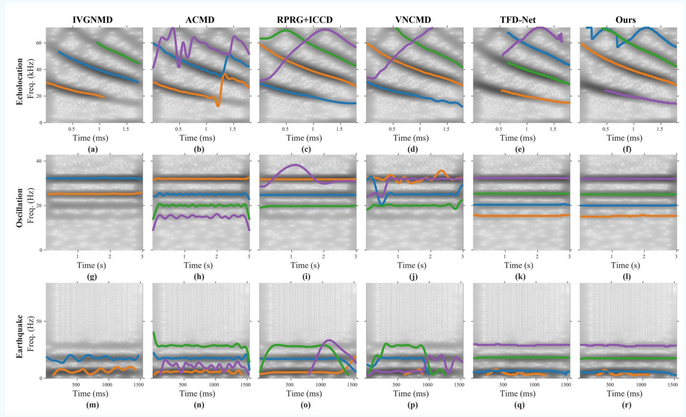
Variant	Dice loss	IF MAE	IF Std
Mask only	0.917	18.57	18.98
+ IFR prediction	0.205	1.87	2.36
+ IFR feedback	0.198	1.66	1.97
Backbone IFR	0.201	1.82	2.25
Backbone seg.	0.910	18.56	18.98

IFR supervision prevents instance collapse. Query feedback further reduces IF MAE by 11.2% over IFR prediction alone.

4 Synthetic & real-world evidence



Synthetic signals at 20 dB, K = 2–5. All six methods.
Colored: predicted IF. Black dashed: ground truth.



Bat echolocation (143 kHz sampling), power oscillation, and earthquake floor vibration. Qualitative STFT overlays. No ground-truth IF labels are available for these signals.

Queries specialize along the rate axis

500 signals, 1,750 matched pairs. Four queries prefer positive IFR and six prefer negative IFR.

Conclusions & scope

Frequency-rate lifting resolves most finite-resolution crossing ambiguities. Across all tested SNR levels, TFFR-Net improves IF MAE by at least 5.8% and Dice loss by at least 5.4% over TFD-Net.

Current output: component ridges. Future work: time-domain reconstruction and broader real-signal validation.