The 39th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

西安交通大學 XI'AN JIAOTONG UNIVERSITY

蘭州大學 LANZHOU UNIVERSITY

极韵 | DREAM

SMILES LAB 域览九州

M3Net: Efficient Time-Frequency Integration Network with Mirror Attention for Audio Classification on Edge

Presenter: Xuanming Jiang **2025 | Pennsylvania, USA**

Xuanming Jiang[1,3], Baoyi An[2,3], Guoshuai Zhao[1,4*], Xueming Qian[1,4]

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China [2]School of Physical Science and Technology, Lanzhou University, Lanzhou, China
[3]Xi'an Jiyun Technology Co., Ltd., Xi'an, China [4]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd., Xi'an, China

jiangxm24@stu.xjtu.edu.cn, anby20@lzu.edu.cn, {guoshuai.zhao, qianxm}@mail.xjtu.edu.cn

Outline

1 Introduction

1 Introduction: Research Background

Applications:

- ➢ Human-machine interaction
- ➢ Intelligent robotic
- ➢ Wise information technology of med

Requirement

High Real-Time Performance

▼

Move the data processing procedure from the cloud to the edge for reducing latency

Challenges：

- ➢ Place higher demands on model performance
- ➢ Impose greater computation cost on MCUs with limited resources

Solution

Apply CNNs for addressing multiple audio classification scenarios based on edge devices

Higher Model Complexity → **Generally** → **Better Performance**

Challenges：

- ➢ SOTA methods currently struggle to achieve a required balance between performance and complexity
- ➢ CNNs are originally designed for image-based tasks, which may limit the performance of audio classification models based on CNNs

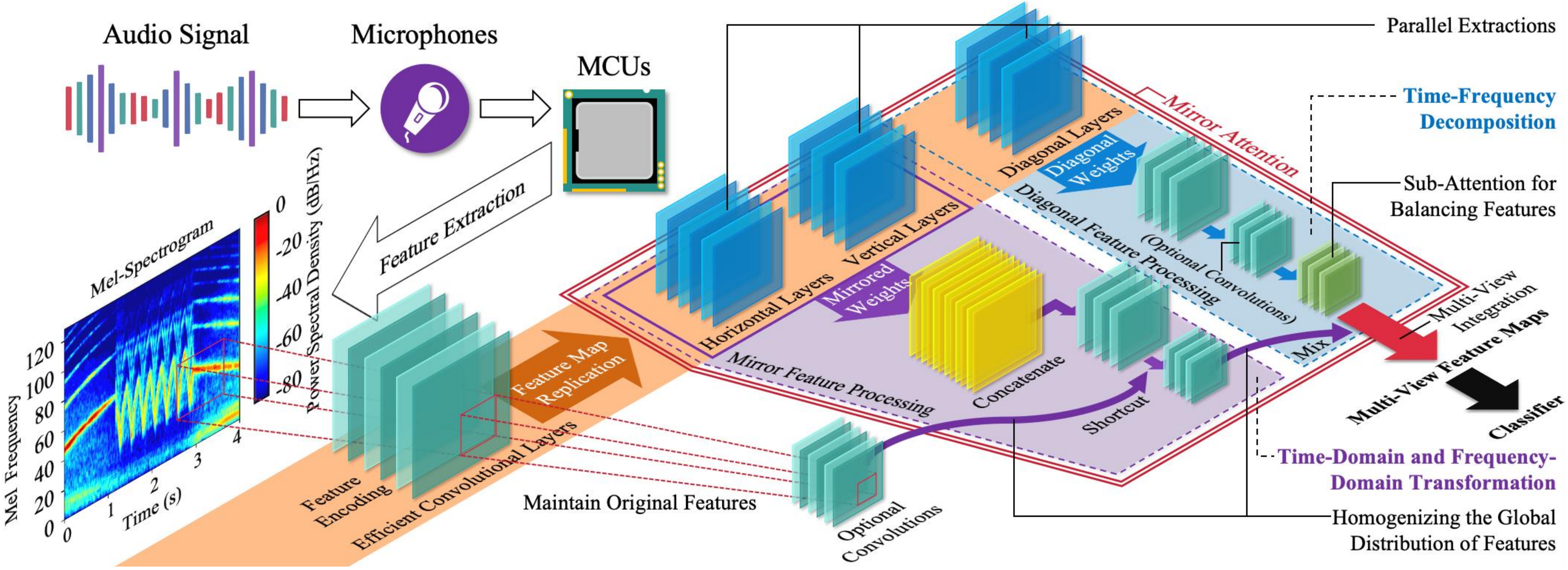
Existing Methods:

- ➢ Model compression and pruning
- ➢ Multiple attention mechanisms
- ➢ Transfer learning

2 Methodology

2 Methodology: Emulation of Human Counterintuitive Learning

Mini Mirror Multi-View Network (M3Net)



Methodology: Time-Frequency Integration

(1) Define feature element:

$$\boldsymbol{X}=\{X(w,h,c)|w\in[0,W\text{-}1],h\in[0,H\text{-}1],c\in[0,C\text{-}1]\}$$

(2) Time-domain transformation:

$$X_{T,c\in[0,C\text{-}1]}(w,h)=M_{HMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{HMF}=\begin{bmatrix}-1&0&W\\0&1&0\\0&0&1\end{bmatrix}$$

(3) Frequency-domain transformation:

$$X_{F,c\in[0,C\text{-}1]}(w,h)=M_{VMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{VMF}=\begin{bmatrix}1&0&0\\0&-1&H\\0&0&1\end{bmatrix}$$

(4) Time-frequency decomposition:

$$X_{D,c\in[0,C\text{-}1]}(w,h)=M_{CSF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{CSF}=\begin{bmatrix}-1&0&W\\0&-1&H\\0&0&1\end{bmatrix}$$

Mirror Attention

$$\boldsymbol{X}_{MA}(w,h,c)=\boldsymbol{X}_{MFP}(w,h,c)\odot\boldsymbol{X}_{DFP}(w,h,c)$$

Mirror features between time and frequency

$$\boldsymbol{X}_{MFP}=F_M(M_{HMF}\boldsymbol{X},M_{VMF}\boldsymbol{X}),\quad \boldsymbol{X}\in R^{W\times H\times C}$$

Separated time and frequency features

$$\boldsymbol{X}_{DFP}=F_D(M_{CSF}\boldsymbol{X})=F_D(M_{HMF}M_{VMF}\boldsymbol{X})$$

Methodology: Time-Frequency Integration

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

(1) Define feature element:

$$\boldsymbol{X}=\{X(w,h,c)|w\in[0,W\text{-}1],h\in[0,H\text{-}1],c\in[0,C\text{-}1]\}$$

(2) Time-domain transformation:

$$X_{T,c\in[0,C\text{-}1]}(w,h)=M_{HMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{HMF}=\begin{bmatrix}-1&0&W\\0&1&0\\0&0&1\end{bmatrix}$$

(3) Frequency-domain transformation:

$$X_{F,c\in[0,C\text{-}1]}(w,h)=M_{VMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{VMF}=\begin{bmatrix}1&0&0\\0&-1&H\\0&0&1\end{bmatrix}$$

(4) Time-frequency decomposition:

$$X_{D,c\in[0,C\text{-}1]}(w,h)=M_{CSF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\quad M_{CSF}=\begin{bmatrix}-1&0&W\\0&-1&H\\0&0&1\end{bmatrix}$$

Mirror Attention

$$\boldsymbol{X}_{MA}(w,h,c)=\boldsymbol{X}_{MFP}(w,h,c)\odot\boldsymbol{X}_{DFP}(w,h,c)$$

Mirror features between time and frequency

$$\boldsymbol{X}_{MFP}=F_M(M_{HMF}\boldsymbol{X},M_{VMF}\boldsymbol{X}),\quad \boldsymbol{X}\in R^{W\times H\times C}$$

Separated time and frequency features

$$\boldsymbol{X}_{DFP}=F_D(M_{CSF}\boldsymbol{X})=F_D(M_{HMF}M_{VMF}\boldsymbol{X})$$

3 Experimental Results

3 Experimental Results: SOTA Comparison

UrbanSound 8K

Volume: ~8,700 audio clips **Content:** Typical urban noises **Duration:** <4s

Details: 10 classes in total, including air conditioners, car horns, children playing, dogs barking, drilling, engine idling, gunshots, jackhammers, sirens, and street music.

Dataset: UrbanSound8K		Accuracy	# Param	Wilcoxon	Data	Pre-training
Methods	Features	(%)	($\times 10^6$)	p-value	augmentation	
AemNet-DW	Log-Mel	82.25	0.9	< 5.0e-2		✓
ULSED	Log-Mel	83.5	0.34	< 5.0e-2	✓	
2D CNN	GFCC	89	1.8	< 5.0e-2	✓	
PhiNets M40	Mel-Spectrogram	76.3	**0.027**	< 5.0e-2	✓	
SE-TCAM 1D CNN	Raw audio	94.04	0.81	< 5.0e-2		
M3Net (Ours)	Mel-Spectrogram	**97.44**	0.029	baseline		

Experimental Results: SOTA Comparison

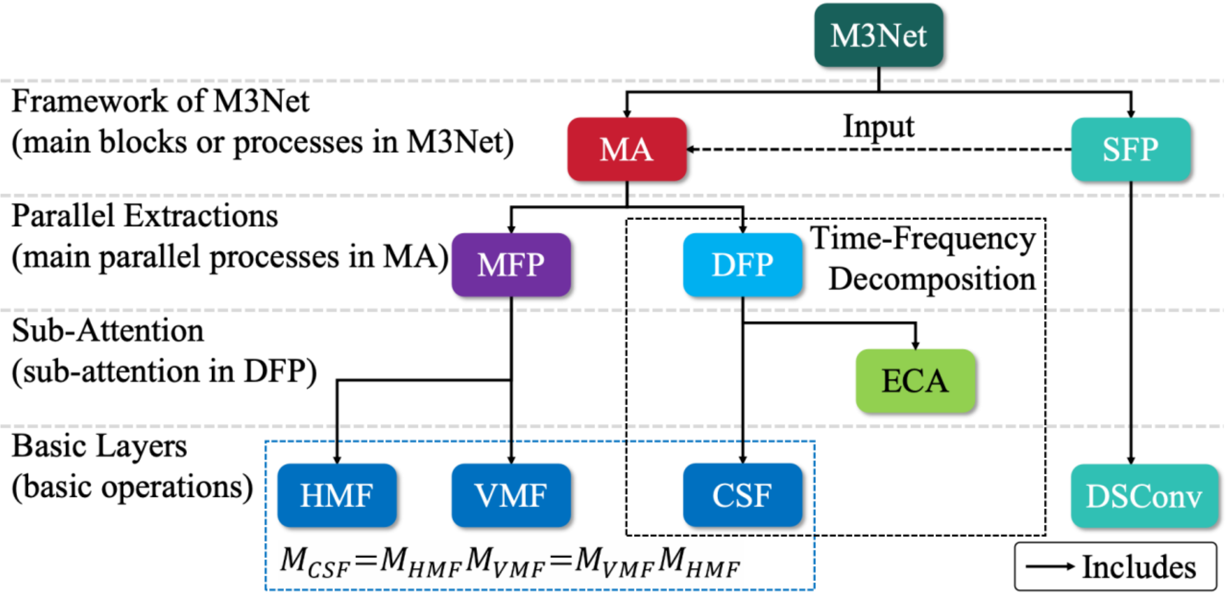
SpeechCommands V2

Volume: ~105,000 audio clips **Content:** Voice commands **Duration:** ~1s

Details: 35 classes in total, including verbal directives such as “yes”, “no”, “up”, “down”, numeric commands ranging from “zero” to “nine”, miscellaneous terms like “bed”, “bird”, “cat”, “dog”.

Dataset: SpeechCommandsV2		Accuracy	# Param	Wilcoxon	Data	Pre-training
Methods	Features	(%)	($\times 10^6$)	p-value	augmentation	
DeLoRes M	Log-Mel	89.7	5.3	< 5.0e-2	✓	✓
AdaptFormer	Log-Mel	92.3	1.43	< 5.0e-2		✓
DCLS-Delays	Mel-Spectrogram	95.35	2.5	< 5.0e-2	✓	
SeqBoat	Raw audio	97.35	0.293	2.2e-1		
EAT-S-GMME	Raw audio	**97.88**	1.54	1.0e-1	✓	
M3Net (Ours)	Mel-Spectrogram	97.03	**0.031**	baseline		

3 Experimental Results: Ablation Study



Dataset	M3Net w/o MA		M3Net w/o SFP	
	Accuracy (%)	Wilcoxon	Accuracy (%)	Wilcoxon
US8K	75.29 (-22.15)	< 5.0e-2	87.31 (-10.13)	< 5.0e-2
SCV2	68.13 (-28.90)	< 5.0e-2	84.26 (-12.77)	< 5.0e-2

Ablation target		1	2	3	4
ECA			✓	✓	✓
CSF				✓	✓
VMF					✓
HMF					
US8K	Accuracy (%)	93.88 (-3.56)	94.56 (-2.88)	95.36 (-2.08)	94.82 (-2.62)
	Wilcoxon	< 5.0e-2	< 5.0e-2	< 5.0e-2	< 5.0e-2
SCV2	Accuracy (%)	94.00 (-3.03)	94.20 (-2.76)	94.89 (-2.14)	91.82 (-5.21)
	Wilcoxon	< 5.0e-2	< 5.0e-2	< 5.0e-2	< 5.0e-2

4 Conclusion & Discussion

Conclusion & Discussion

Background

① With the advancement of AI, edge devices are taking on increasing responsibilities

② As IC enter the post-Moore's Law era, edge devices' performance limitations become increasingly significant

③ Edge devices need to meet societal demands from both hardware and algorithmic perspectives

Contributions

① An **time-frequency separation** method for audio learning

② An **mirror attention mechanism** for extending pattern set

③ The **M3Net** with high-performance but lightweight

Methodology

M3Net transcends the conventional emulation of human auditory perception by CNNs and ventures into the field of replicating human counterintuitive supervised cognitive processes.

Achievements

M3Net:

① Achieving comparable performance to SOTAs with less than 10% of the parameters

② Adapting to different kinds of audio contents

③ More sensitive to complex and dynamic audio

Thank you

2025 | PHILADELPHA, USA