



# Multi-view Fusion and Parameter Perturbation for Few-Shot Class-Incremental Audio Classification

Yulu Fang<sup>1\*</sup>, Mingyue He<sup>2\*</sup>, Qisheng Xu<sup>3†</sup>, Jianqiao Zhao<sup>4</sup>, Cheng Yang<sup>3</sup>, Kele Xu<sup>3†</sup>, Yong Dou<sup>3</sup>

<sup>1</sup>Zhengzhou university, China

<sup>2</sup>Wuhan University, China

<sup>3</sup>College of Computer Science and Technology, National University of Defense Technology, China

<sup>4</sup>Durham university, United Kingdom

fang-yulu@163.com, scouthe@163.com, qishengxu@nudt.edu.cn, kele.xu@ieee.org

## Abstract

Audio classification tasks typically assume a fixed number of classes, which is often unrealistic in real-world applications where the target class vocabulary is dynamic or unknown in advance. A significant challenge arises when models must adapt to new classes incrementally, as this process is prone to catastrophic forgetting—a sharp decline in performance on previously learned classes, especially in data-scarce scenarios. While dynamic network-based methods and prototype refinement-based methods have been proposed to address these challenges, they overlook two critical issues: (1) inadequate representation of raw audio samples, which limits generalization, and (2) the risk of overfitting, which limits adaptivity. In this paper, we propose **Multi-View Fusion and Parameter Perturbation (MVFP2P)**, a novel framework that leverages the complementary learning system to enhance generalizability and adaptivity within a unified incremental learning framework. MVFP2P addresses the limitations of existing methods by integrating multi-view learning to enrich feature representation and a parameter perturbation mechanism to reduce overfitting. Extensive evaluations on two widely-used audio datasets, NS-100 and LS-100, demonstrate that MVFP2P outperforms state-of-the-art methods in terms of average accuracy and performance drop rate. Notably, MVFP2P not only mitigates catastrophic forgetting more effectively but also enhances the model’s adaptability to new classes, making it a robust solution for dynamic audio classification tasks.

**Index Terms:** incremental learning, few-shot learning, audio classification, multi-view learning, parameter perturbation

## 1. Introduction

Audio classification aims to recognize distinct audio signals by identifying their key patterns, serving as a fundamental component of many audio-based tasks, such as audio scene classification [1], audio analysis [2], and video understanding [3]. Thanks to the powerful automatic feature extraction capabilities of deep neural networks, data-driven audio classification methods have become the mainstream approach. Despite their success, these methods face two significant limitations: (1) Dependence on labeled data: Deep learning models require substantial labeled data to effectively learn the key patterns in audio signals. However, annotating audio data is expensive, making it challenging to obtain sufficient labeled samples. (2) Fixed vocabulary assumption: Existing methods typically assume a fixed set of audio classes, conflicting with real-world scenarios where new classes frequently emerge.

To reduce reliance on labeled data, few-shot learning [4] has emerged as a promising paradigm in audio classification,

enabling models to recognize new classes with only a small number of labeled samples. Pons et al. [5] systematically evaluated mainstream few-shot learning methods, including naive regularization, prototypical networks, transfer learning, and their combinations, to enhance the use of limited training data. Their results showed that metric-based prototypical networks delivered the most promising performance. However, few-shot learning methods have a critical limitation: they focus exclusively on new classes during deployment and fail to retain knowledge of the original training classes. This makes them unsuitable for systems requiring consistent performance across both old and new classes.

Incremental learning is a paradigm designed to enable systems to classify new classes while retaining knowledge of old ones, addressing catastrophic forgetting [6]. This paradigm aligns with the dynamic nature of real-world audio classification, where new classes frequently emerge. Wang et al. [7] introduced a generative replay strategy to synthesize training data for old classes, mitigating catastrophic forgetting. Similarly, Mulimani et al. [8] proposed a class-incremental framework for multi-label audio classification, enabling learning of new classes without forgetting old ones. However, these methods face two key challenges: (1) they require large amounts of labeled data for new classes, and (2) they often need retraining when new classes are introduced. These limitations make them impractical for resource-constrained environments or domains with scarce labeled data.

As an alternative, Few-Shot Class-Incremental Audio Classification (FCAC) [9] has been proposed to address these limitations. FCAC combines the strengths of few-shot and incremental learning, enabling models to recognize new classes with limited audio samples. Wang et al. [10] pioneered dynamic network-based FCAC for audio classification, progressively expanding a base classifier to incorporate novel classes. Recent studies [11, 12, 13, 14] have further refined classifier prototypes and enhanced their distinctiveness. However, these methods overlook two critical issues: (1) inadequate representation of raw audio samples, which limits generalization, and (2) potential overfitting. Specifically, they rely on log-mel spectrograms, which compress high-frequency information and lose discriminative details. Additionally, feature extractors trained on large base class datasets risk overfitting to base classes, hindering adaptation to new classes.

In this paper, we propose **Multi-View Fusion and Parameter Perturbation (MVFP2P)**, a novel framework that leverages the complementary learning system (CLS) to enhance generalizability and adaptivity within a unified incremental learning framework. Specifically, we use multi-view learning to aggregate multiple hand-crafted features, enabling richer representations of raw audio samples and mitigating the limi-

\* Equal contribution.

† Corresponding author.

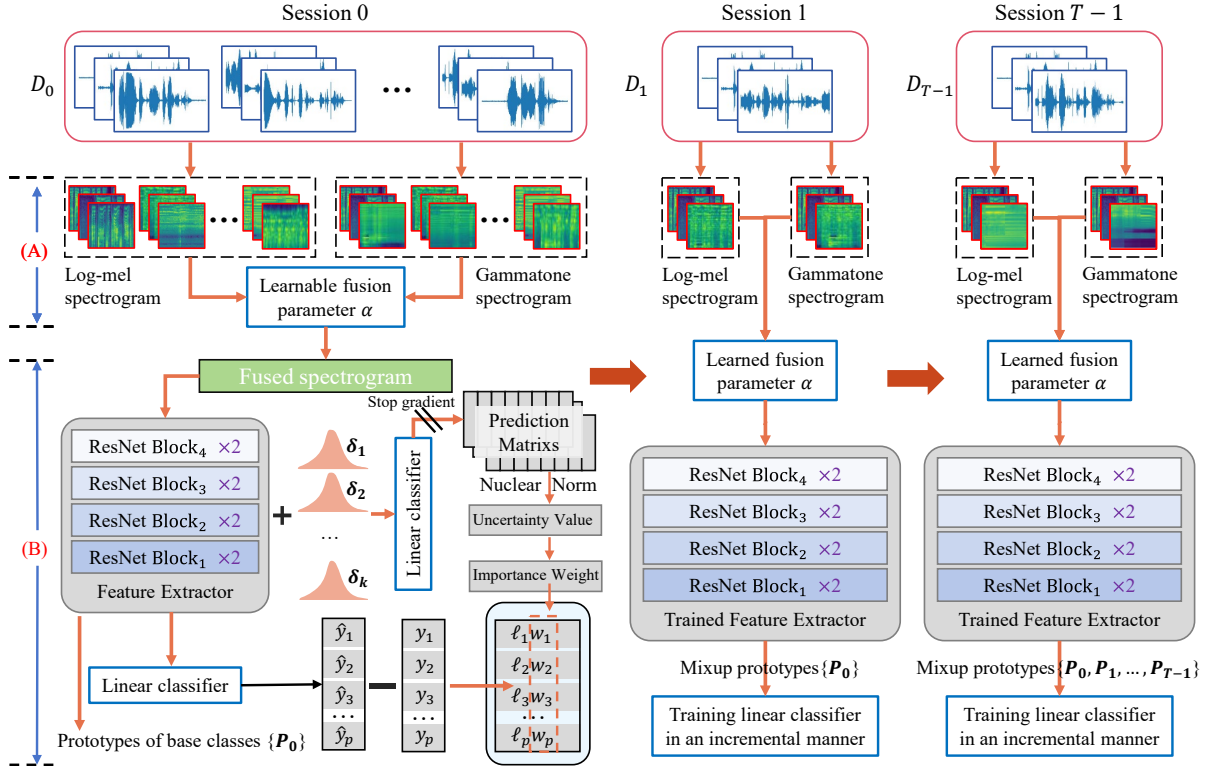


Figure 1: The MVF2P framework comprises two key components: (A) Generalizability through Multi-View Learning and (B) Adaptivity via Parameter Perturbation. First, a multi-view spectrogram fusion module aggregates log-mel and gammatone spectrograms to create a richer representation of raw audio, enhancing model generalizability. Second, a parameter perturbation mechanism follows a Gaussian probability density function and reduces overfitting by performing  $k$  times forward passes with perturbations, generating a  $k \times B \times N_0$  matrix (where  $B$  is batch size and  $N_0$  is base session classes). Using the nuclear norm, uncertainty values are computed to weight training losses for new samples, enabling targeted learning and improving adaptability.

tations of single-view approaches. Additionally, we introduce a parameter-perturbation mechanism for the feature extractor, which applies controlled perturbations during training to prevent overfitting to base classes and ensure robust performance on unseen data.

## 2. Method

### 2.1. Preliminary

FCAC represents a significant challenge in audio classification, requiring models to learn incrementally from a stream of limited audio samples while balancing the retention of old knowledge and the acquisition of new information. We assume  $T$  training tasks (referred to as sessions), including one base session (session 0) and  $T - 1$  incremental sessions (sessions 1 to  $T - 1$ ). Mathematically, the training data for each session is denoted as  $\mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_{T-1}$ , where  $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$  for  $i \neq j$ . For the  $i$ -th session's training data  $\mathcal{D}_i$ , we represent it as  $(X_i, Y_i) = (x_j, y_j)_{j=1}^{N_i}$ , where  $x_j \in \mathbb{R}^D$  is a training sample belonging to class  $y_j \in Y_i$ , and  $Y_i$  is the label space for session  $i$ . During session  $i$ , only  $\mathcal{D}_i$  is used for training, and the trained model is evaluated on all prior sessions,  $\mathcal{Y}_i = Y_1 \cup Y_2 \cup \dots \cup Y_i$ . The base session dataset  $\mathcal{D}_0$  is large-scale, with sufficient samples per classes for training. In contrast, dataset  $\mathcal{D}_1$  to  $\mathcal{D}_{T-1}$  are small-scale and follow an  $N$ -way  $K$ -shot setup: each incremental session contains  $N$  classes, with only  $K$  samples per class.

This setup underscores the core challenge of FCAC: learning effectively from extremely limited data while avoiding catastrophic forgetting of previously learned classes.

### 2.2. Generalizability through multi-view learning

Typically, FCAC methods rely on the log-mel spectrogram as the sole input for model training. However, the log-mel spectrogram is generated using Mel-scale-based bandpass filters, which are highly sensitive to low-frequency information but compress high-frequency information. This inherent limitation hinders performance in audio classification scenarios involving high-frequency signals, such as instrumental sounds.

To address this, we propose a multi-view spectrogram fusion module. While log-mel spectrograms excel at discriminating low-frequency information, we introduce gammatone spectrograms as a complementary view to capture high-frequency details. Derived from a gammatone filter bank that mimics the human cochlea's frequency response, gammatone spectrograms use frequency-selective filters to represent audio signal across different bands, capturing fine-grained temporal and spectral information and mitigating high-frequency compression. Mathematically, the fusion module is defined as:

$$Spec_f = \alpha \times \text{log-mel} + (1 - \alpha) \times \text{gammatone}, \quad (1)$$

where  $Spec_f$  denotes the fused spectrogram,  $\alpha$  is a learnable parameter optimized during training.

Table 1: Comparison Results from Various Methods on LS-100 Dataset.

Methods	Venue	Session	Accuracy in various sessions(%)								AA (%)	PD (%)	
			0(0-59)	1(60-64)	2(65-69)	3(70-74)	4(75-79)	5(80-84)	6(85-89)	7(90-94)			8(95-99)
Finetune	\	Base	92.02	72.90	37.03	28.12	20.75	14.45	5.70	3.23	0.27	30.50	91.75
		Novel	-	86.60	31.50	28.87	25.45	24.24	18.17	13.46	11.80	30.01	74.80
		Both	92.02	73.95	36.24	28.27	21.93	17.33	9.86	7.00	4.88	32.39	87.14
iCaRL	CVPR 2017	Base	92.02	80.80	73.18	58.45	26.95	16.93	32.58	29.53	26.38	48.54	65.64
		Novel	-	58.00	67.10	57.40	20.05	16.48	30.33	26.83	28.95	38.14	29.05
		Both	92.02	79.05	72.31	58.24	25.23	16.80	31.83	28.54	27.41	47.94	64.61
DFSL	ICASSP 2021	Base	91.93	91.93	91.88	91.85	91.83	91.86	91.85	91.85	91.84	91.87	<b>0.09</b>
		Novel	-	53.60	61.90	50.67	48.90	51.56	47.97	44.11	45.38	50.51	<b>8.22</b>
		Both	91.93	88.97	87.60	83.61	81.11	80.01	77.22	74.26	73.25	81.99	18.68
CEC	CVPR 2023	Base	91.72	91.67	91.25	91.14	91.10	91.07	90.97	<b>90.66</b>	<b>90.72</b>	91.14	1.00
		Novel	-	86.30	82.76	69.67	68.25	67.06	66.03	60.35	60.05	70.06	26.25
		Both	91.72	91.25	90.04	86.84	85.38	84.01	82.65	79.49	78.45	85.54	13.27
ARP	INTERSPEECH 2023	Base	92.35	92.22	91.85	91.66	91.60	91.56	91.47	91.43	91.31	91.72	1.04
		Novel	-	42.92	44.99	41.00	38.09	37.75	35.57	32.91	31.89	38.14	11.03
		Both	92.35	88.43	85.16	81.53	78.22	75.73	72.84	69.87	67.54	79.07	24.81
META-SC	INTERSPEECH 2023	Base	<b>92.75</b>	<b>92.73</b>	92.28	<b>92.29</b>	<b>92.28</b>	<b>92.29</b>	<b>92.25</b>	90.54	90.54	<b>91.99</b>	2.21
		Novel	-	91.93	83.54	77.73	73.97	74.21	73.68	71.52	71.75	77.29	20.18
		Both	<b>92.75</b>	<b>92.67</b>	91.03	89.37	87.70	86.97	86.06	83.53	83.02	88.12	9.73
PAN	TMM 2023	Base	91.83	91.47	91.38	91.27	91.21	91.20	90.92	90.50	90.12	91.10	1.71
		Novel	-	89.14	86.63	78.55	72.07	70.41	68.39	64.27	62.93	74.05	26.21
		Both	91.83	91.29	90.70	88.73	86.42	85.09	83.41	80.84	79.24	86.39	12.59
Ours	\	Base	92.48	92.35	<b>92.32</b>	91.93	92.15	92.06	92.05	90.25	90.38	91.78	2.10
		Novel	-	<b>95.60</b>	<b>88.90</b>	<b>80.33</b>	<b>80.60</b>	<b>81.88</b>	<b>78.73</b>	<b>77.20</b>	<b>77.20</b>	<b>82.56</b>	18.40
		Both	92.48	92.60	<b>91.83</b>	<b>89.61</b>	<b>89.26</b>	<b>89.07</b>	<b>87.61</b>	<b>85.44</b>	<b>85.11</b>	<b>89.22</b>	<b>7.37</b>

### 2.3. Adaptivity via parameter perturbation

In few-shot scenarios, base classes typically have abundant training samples, while new classes introduced during the incremental phase often have limited samples. This imbalance risks overfitting: the model may over-learn base classes during the base training phase, memorizing their features and losing adaptability to new classes. Conversely, the scarcity of new class samples may cause overfitting during the incremental phase, leading to poor generalization on unseen test samples.

To address this, we design a loss adjustment mechanism based on parameter perturbation. During training, we introduce  $k$  perturbations to the model’s parameters, generating a  $k \times B \times N_0$  matrix, where  $B$  is the batch size and  $N_0$  is the number of base classes. Using the nuclear norm, we compute uncertainty values from this matrix, reflecting the model’s prediction confidence. These values assign importance weights to samples, adjusting their contribution to the training loss. By prioritizing high-uncertainty samples, the mechanism reduces overfitting to base classes and enhances adaptability to new classes, ensuring robust performance in dynamic environments.

## 3. Experiments

### 3.1. Datasets description

To evaluate the effectiveness of our proposed method, MVF2P, we conduct experiments on two publicly available datasets: Nsynth [15] and librispeech [16]. Nsynth is a large-scale, high-quality dataset containing 305,979 monophonic music clips from 1,006 instruments. Librispeech is a comprehensive English speech corpus with approximately 1,000 hours of labeled audiobook recordings, widely used for speech processing tasks. For our experiments, we use subsets from these datasets, labeled NS-100 (from Nsynth) and LS-100 (from Librispeech).

### 3.2. Implementation details

Our implementation is divided into two phases: base session and incremental session. The backbone of our method is a ResNet18 network adapted as a feature extractor, combined with a linear classifier. In base session, we train model for 200 epochs on a large-scale dataset  $\mathcal{D}_r$ . We extract and fuse log-mel and the gammatone spectrogram from the raw audio samples as described in Section 2.2, and encode the fused spectrogram using the feature extractor to obtain a general feature representation. To enhance adaptivity, we set the parameter perturbation count to 5. During incremental sessions, we refine the linear classifier by mixing samples from new classes with prototypes from old classes. Each session introduces  $N = 5$  new classes, with  $K = 5$  samples per class. Both phases are optimized using Stochastic Gradient Descent with a learning rate of 0.1. Finally, model performance is evaluated using three metrics: Accuracy, Average Accuracy (AA), and Performance Drop (PD).

### 3.3. Experimental results

In this part, we evaluated our approach MVF2P by comparing it against seven competitive baseline methods: Finetune [17], iCaRL [18], DFSL [10], CEC [19], FACT [20], CLOM [21], DSN [22], LDC [23], ARP [12], META-SC [9], PAN [13]. To ensure a fair and comprehensive comparison, we use a combination of results from reproduced open-source implementations and metrics reported in the original papers as our baseline.

As presented in Table 1, MVF2P achieves the highest AA scores on the LS-100 dataset, reaching 82.56% for novel classes and 89.22% for both novel and base classes. These results represent significant improvements of 5.27% and 1.10%, respectively, over the state-of-the-art method META-SC. Furthermore, MVF2P demonstrates the lowest PD score of 7.37% for both novel and base classes on the LS-100, outperforming all base-

Table 2: Comparison Results from Various Methods on Nsynth-100 Dataset.

Methods	Venue	Session	Accuracy in various session(%)										AA (%)	PD (%)
			0	1	2	3	4	5	6	7	8	9		
Finetune	\	Base	99.96	88.91	85.41	80.36	72.51	45.24	59.31	48.53	50.68	53.28	68.42	46.68
		Novel	-	38.75	30.25	36.96	37.54	28.95	27.24	22.30	20.58	19.00	29.06	19.75
		Both	99.96	84.73	76.92	71.06	63.18	40.15	47.99	38.33	38.01	37.86	59.82	62.10
iCaRL	CVPR 2017	Base	99.98	98.42	99.25	98.40	94.56	82.36	85.09	80.59	75.78	74.53	88.90	25.45
		Novel	-	36.94	31.88	35.03	38.33	35.27	30.76	26.75	25.52	22.27	31.42	14.67
		Both	99.98	93.30	88.88	84.82	79.57	67.65	65.92	59.65	54.62	51.01	74.54	48.97
DFSL	ICASSP 2021	Base	99.93	99.11	98.83	95.83	94.84	94.81	94.39	93.76	92.06	91.61	95.52	8.32
		Novel	-	57.01	55.57	59.89	59.35	56.46	52.29	50.94	52.57	52.49	55.17	4.52
		Both	99.93	96.00	92.95	89.26	86.47	83.66	80.28	77.68	76.12	75.01	85.74	24.92
CEC	CVPR 2023	Base	99.96	<b>99.87</b>	<b>99.90</b>	<b>99.29</b>	<b>99.24</b>	<b>99.30</b>	<b>99.26</b>	<b>99.24</b>	<b>99.20</b>	<b>99.23</b>	<b>99.45</b>	<b>0.73</b>
		Novel	-	71.06	71.61	72.37	69.17	69.20	66.92	64.80	65.28	63.59	68.22	7.47
		Both	99.96	<b>97.47</b>	<b>95.56</b>	<b>93.52</b>	91.22	<b>89.90</b>	87.85	85.84	84.92	83.19	90.94	16.77
ARP	INTERSPEECH 2023	Base	99.96	99.85	99.72	99.10	98.58	98.56	98.50	98.37	98.23	98.15	98.90	1.81
		Novel	-	52.99	59.92	66.25	67.62	68.41	63.54	60.05	60.35	57.12	61.81	<b>-4.13</b>
		Both	99.96	95.95	93.60	92.06	90.32	89.14	86.16	83.47	82.28	79.69	89.26	20.27
META-SC	INTERSPEECH 2023	Base	<b>99.98</b>	98.05	98.01	95.08	93.44	94.09	92.98	92.79	91.75	91.60	94.78	8.38
		Novel	-	84.63	<b>82.37</b>	<b>83.01</b>	84.42	<b>78.68</b>	75.43	72.65	72.81	70.46	78.27	14.18
		Both	<b>99.98</b>	96.93	95.60	92.50	91.03	89.27	86.78	84.96	83.77	82.09	90.29	17.89
PAN	TMM 2023	Base	99.95	98.11	97.85	95.10	94.39	94.63	94.06	93.99	93.06	92.76	95.39	7.19
		Novel	-	83.77	79.59	82.86	83.26	77.64	74.25	71.06	71.15	68.85	76.94	14.92
		Both	99.95	96.92	95.04	92.48	91.42	89.32	87.07	85.07	83.83	82.00	90.31	17.95
Ours	\	Base	<b>99.98</b>	97.51	97.15	95.75	94.22	94.51	93.66	94.27	92.64	92.13	95.18	7.85
		Novel	-	<b>90.80</b>	80.30	82.73	<b>85.30</b>	<b>78.68</b>	<b>79.87</b>	<b>76.43</b>	<b>76.08</b>	<b>72.82</b>	<b>80.33</b>	17.98
		Both	<b>99.98</b>	96.95	94.55	92.96	<b>91.84</b>	89.56	<b>88.79</b>	<b>87.33</b>	<b>85.66</b>	<b>83.44</b>	<b>91.11</b>	<b>16.54</b>

Table 3: Ablation Study of different components in MVF2P

No.	Component		AA		PD	
	Multi-view Learning	Parameter Perturbation	Novel (%)	Both (%)	Novel (%)	Both (%)
①	✗	✗	77.41	88.39	14.20	9.28
②	✓	✗	80.52	88.80	16.52	8.39
③	✓	✓	<b>82.56</b>	<b>89.22</b>	18.40	<b>7.37</b>

line methods. Additionally, MVF2P consistently achieves superior or comparable results across all training sessions, including base classes, novel classes, and their combination. These results highlight our method’s strong capability to balance generalizability and adaptivity.

To further validate the effectiveness of our method, we conducted experiments on NS-100, with results presented in Table 2. Our approach consistently exhibits superior performance for novel classes and both novel and base classes across almost all training sessions, aligning with the trends observed on LS-100. Notably, while CEC achieves the highest accuracy for base classes, it exhibits limited adaptability to novel classes. In contrast, our method not only ranks among the top three for base classes but also achieves significant improvements for novel classes and their combination with base classes.

### 3.4. Ablation study

To evaluate the effectiveness of MVF2P’s key components, we conducted an ablation study on LS-100. As detailed in Section 2, our approach includes two core components: Generalizability through multi-view learning and Adaptivity via parameter perturbation. The results of incrementally adding these components, shown in Table 3, demonstrate that each contributes to improved AA and reduced PD. The multi-view learning enriches feature representation by combining log-mel and gammatone spectrograms, while the parameter perturbation reduces overfitting by introducing controlled noise during training. Together, these components enable the model to better balance retaining old knowledge and acquiring new information, resulting

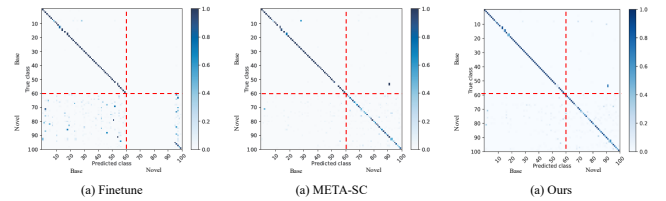


Figure 2: Confusion matrix of the last incremental session obtained by Finetune, META-SC, and Ours method on the LS-100.

in robust performance in incremental learning scenarios.

### 3.5. Further analysis

To further evaluate the performance of our method across different classes, we conducted an error rate analysis. Specifically, after completing training for final incremental session on LS-100, we used the trained model to classify all audio classes and calculated the accuracy for each class. We compared our approach with Finetune and the state-of-the-art baseline META-SC. As illustrated in Figure 2, our method not only achieved the highest overall performance but also consistently outperformed the existing baselines across every individual class.

## 4. Conclusion

In this paper, we propose **Multi-View Fusion and Parameter Perturbation (MVF2P)**, a novel framework that leverages the complementary learning system to enhance generalizability and adaptivity within a unified incremental learning framework. MVF2P integrates two core components: Generalizability through multi-view learning and Adaptivity via parameter perturbation. Experimental results demonstrate the effectiveness of our method in balancing generalizability and adaptivity, achieving robust performance in incremental learning scenarios.

## 5. Acknowledgements

This work was supported by National Science and Technology Major Project (2023ZD0121101), National University of Defense Technology (ZZCX-ZZGC-01-04)

## 6. References

- [1] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [2] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, "Comparative analysis of audio classification with mfcc and stft features using machine learning techniques," *Discover Internet of Things*, vol. 4, no. 1, p. 1, 2024.
- [3] L. Sun, X. Xu, M. Wu, and W. Xie, "Auto-acd: A large-scale dataset for audio-language representation learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5025–5034.
- [4] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *arXiv preprint arXiv:1904.04232*, 2019.
- [5] J. Pons, J. Serrà, and X. Serra, "Training neural audio classifiers with few data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 16–20.
- [6] Z. Gao, X. Zhang, K. Xu, X. Mao, and H. Wang, "Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [7] Z. Wang, C. Subakan, E. Tzinis, P. Smaragdis, and L. Charlin, "Continual learning of new sound classes using generative replay," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 308–312.
- [8] M. Mulimani and A. Mesaros, "Class-incremental learning for multi-label audio classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 916–920.
- [9] Y. Li, W. Cao, J. Li, W. Xie, and Q. He, "Few-shot class-incremental audio classification using stochastic classifier," *arXiv preprint arXiv:2306.02053*, 2023.
- [10] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 321–325.
- [11] W. Xie, Y. Li, Q. He, and W. Cao, "Few-shot class-incremental audio classification via discriminative prototype learning," *Expert Systems with Applications*, vol. 225, p. 120044, 2023.
- [12] W. Xie, Y. Li, Q. He, W. Cao, and T. Virtanen, "Few-shot class-incremental audio classification using adaptively-refined prototypes," *arXiv preprint arXiv:2305.18045*, 2023.
- [13] Y. Li, W. Cao, W. Xie, J. Li, and E. Benetos, "Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes," *IEEE Transactions on Multimedia*, 2023.
- [14] Y. Li, J. Li, Y. Si, J. Tan, and Q. He, "Few-shot class-incremental audio classification with adaptive mitigation of forgetting and overfitting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [15] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1068–1077.
- [16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [17] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5138–5146.
- [18] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.
- [19] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-shot incremental learning with continually evolved classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 455–12 464.
- [20] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9046–9056.
- [21] Y. Zou, S. Zhang, Y. Li, and R. Li, "Margin-based few-shot class-incremental learning with class-level overfitting mitigation," *Advances in neural information processing systems*, vol. 35, pp. 27 267–27 279, 2022.
- [22] B. Yang, M. Lin, Y. Zhang, B. Liu, X. Liang, R. Ji, and Q. Ye, "Dynamic support network for few-shot class incremental learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2945–2951, 2022.
- [23] B. Liu, B. Yang, L. Xie, R. Wang, Q. Tian, and Q. Ye, "Learnable distribution calibration for few-shot class-incremental learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 699–12 706, 2023.