Team koi: Technical Report for EPIC-SOUNDS Audio-Based Interaction Recognition Challenge 2025

Xinglin Xie, Kehuan Song, Licheng Jiao, Wenping Ma, Xiaoqiang Lu, Dan Zhang
School of Artificial Intelligence, Xidian Universitly

Abstract

Epic-Kitchen EPIC-SOUNDS Audio-Based Interaction Recognition Challenge 2025 involves learning the mapping from audio samples to their corresponding action labels. On the CodaLab leaderboard, our team "koi" secured a Top-1 accuracy of 57.51%, placing first. In this report, we will go into the technical specifics of how we tackled this task.Our solution leverages refined tuning of AudioInceptionNeXt [1] and cross-architecture fusion with UniRepLKNet [2]. A dynamic weight allocation strategy is introduced to integrate feature representations from different architectures, enhancing classification accuracy while maintaining computational efficiency. Experimental results demonstrate that the proposed approach achieves a Top-1 accuracy of 57.51 % on the EPIC-SOUNDS dataset in the challenge.

1. Introduction

EPIC-SOUNDS Audio-Based Interaction Recognition is one of the core events in the EPIC-KITCHENS 2025 Challenge, aiming to identify interaction behavior categories in videos through audio signals. Relying on the EPIC-SOUNDS dataset, the competition requires participants to assign 44 pre-defined interaction category labels (e.g., "open bag", "chop vegetables") to trimmed audio clips. Derived from the audio component of EPIC-KITCHENS-100, the dataset focuses on unscripted daily activities in kitchen scenarios, covering sounds generated by object interactions (e.g., metal collisions, water flow) and wearer actions (e.g., footsteps, tableware operations). Unlike traditional datasets, EPIC-SOUNDS addresses the issues of temporal misalignment between visual and audio events and cross-modal single-label annotation in EPIC-KITCHENS-100, while introducing challenges such as variable audio lengths, background noise interference, and diversity of homogenous sounds (e.g., subtle acoustic differences when chopping different vegetables or acoustic feature variations of the same action in different environments).

In terms of methodologies, previous research on audio interaction recognition has primarily adopted two technical pathways: One is based on Convolutional Neural Network (CNN) architectures, which leverage their local feature extraction capabilities to process the time-frequency domain representations of audio. For example, the Slow-Fast [3] two-stream model captures global frequency semantics and long-term activities (e.g., continuous stirring) through a "slow stream" with low temporal resolution, while a high-resolution "fast stream" analyzes short-term transient features and local details (e.g., the moment a knife cuts into ingredients), enhancing the richness of feature representation through cross-stream interaction. Subsequent studies, such as AudioInceptionNeXt, further combined self-supervised contrastive learning to achieve superior performance on datasets like EPIC-SOUNDS compared to Transformer models, verifying the advantages of CNNs in computational efficiency and feature generalization. The other pathway relies on Transformer-based methods [4] [5], which model the long-range dependency of audio sequences through self-attention mechanisms. For instance, some studies use supervised learning to directly optimize classification loss or employ self-supervised pre-training (e.g., masked audio modeling) to mine latent semantic structures, particularly demonstrating stronger global contextual modeling capabilities in processing long-duration audio clips. Additionally, cross-architecture fusion strategies (such as feature concatenation or dynamic weight fusion of CNNs and Transformers) have gradually become a trend, improving classification accuracy by integrating multi-scale features while mitigating computational costs through lightweight designs.

These approaches all revolve around the time-frequency characteristics of audio data, aiming to solve the problem of distinguishing interaction categories in the complex acoustic environment of EPIC-SOUNDS and providing technical support for intelligent perception in kitchen scenarios. EPIC-SOUNDS Audio-Based Interaction Recognition faces notable challenges in multi-scale time-frequency feature modeling, cross-scenario generalization, and balancing efficiency with accuracy, where audio signals from diverse interactions differ significantly in temporal duration (short transients vs. long-term activities) and frequency distribution (high-frequency details vs. low-frequency fundamentals), while the dataset's complex kitchen background noise requires robust feature extraction to distinguish target sounds from clutter, and the imbalance between Transformer-based models' high performance but high computational cost and CNNs' efficiency but limited long-term context capture demands innovative lightweight or hybrid architectures.

To tackle the challenges of multi-scale feature modeling, limited long-range dependency capture, and model efficiency-accuracy trade-offs, our solution is structured as follows:

- **Multi-branch Optimization of AudioInceptionNeXt**: Parallel depthwise separable convolutions with diverse kernel sizes in a single-stream architecture enable simultaneous extraction of short-term transient details (via small kernels) and long-term global frequency semantics (via large kernels), resolving the temporal-frequency variation challenge in interactions.
- **Global Dependency Modeling with UniRepLKNet**: UniRepLKNet's large convolutional kernels model long-range temporal dependencies in audio sequences, supplementing CNNs' local perception to enhance robustness against background clutter and contextual ambiguity.
- **Dynamic Model Fusion**: Adaptive weight adjustment based on validation performance fuses AudioInceptionNeXt (high efficiency in local features) and UniRepLKNet (strong global modeling) outputs, optimizing ensemble decisions without excessive computational overhead, thus addressing the trade-off between Transformer-like performance and CNN-like efficiency.

2. Methodology

2.1. AudioInceptionNeXt

AudioInceptionNeXt is designed to effectively extract multi-scale time-frequency features from audio signals. It uses parallel multi-scale depthwise separable convolutional kernels in its core block. The large-scale kernels capture long-duration activities and global frequency semantics, while small-scale ones focus on short-duration activities and local frequency details. For instance, a 3×3 kernel can quickly detect sharp transient sounds, like the click of a utensil, while an 11×11 kernel is better at analyzing more extended and continuous audio patterns, such as the steady noise of a running blender.

The overall structure of AudioInceptionNeXt is based on a modified ResNet50 architecture. The input stem processes the log-mel spectrogram, reducing its size while increasing the number of channels. The model is organized into four feature stages, with each stage having a specific number of AudioInceptionNeXt blocks and channel adjustments. The classification head at the end is composed of a global average pooling layer followed by a fully connected layer, which outputs predictions for the 44 interaction categories in the dataset.

The AudioInceptionNeXt block can be formalized as follows:

$$\text{Output} = \text{Concat}(\text{Conv}_{3\times3}(x), \text{Conv}_{k\times k}(x)) + x \quad (1)$$

where x is the input feature map and $\text{Conv}_{k\times k}$ denotes a depthwise separable convolution with kernel size k.

During training, data augmentation techniques are used to improve the model's generalization ability. Temporal augmentations include random cropping with a certain duration jitter, and adding Gaussian noise to simulate real-world noisy environments. Spectral augmentations involve frequency masking and amplitude scaling. The Stochastic Gradient Descent (SGD) optimizer with appropriate momentum and weight decay is used, along with a Cosine annealing learning rate scheduler. Multi-GPU training with SyncBN is employed to speed up the training process.

2.2. UniRepLKNet

Inspired by RepLKNet [6], UniRepLKNet is crafted to model long-range temporal dependencies in audio sequences, which is vital for grasping complex interactions that occur over extended periods. It uses 31×3 depthwise separable convolutions, decomposed into 1×31 (temporal) and 31×1 (frequency) kernels. This decomposition allows the model to capture temporal dependencies over 310ms and significantly reduces the number of parameters, leading to a substantial reduction in computation. For example, in a long sequence of sounds during a cooking process that involves multiple steps and continuous audio streams, UniRepLKNet can effectively analyze the long-term patterns and context within these audio signals.

Dilation rates are applied in different stages, expanding the receptive field up to 1.24s. This enlarged receptive field enables the model to better understand the overall context of the audio, which is particularly useful in scenarios where the relationship between distant audio events matters, such as in a sequence of actions in a kitchen where one action's sound might influence the perception of a later action.

The receptive field calculation for the dilated convolution is given by:

$$RF = 1 + \sum_{i=1}^{n} (k_i - 1) \times \prod_{j=1}^{i-1} s_j \times d_i \quad (2)$$

where k_i is the kernel size, s_j is the stride, and d_i is the dilation rate at layer i.

2.3. Multi-Scale Temporal-Semantic Fusion

As shown in Fig, 1, we employ a late-stage ensemble strategy by integrating the inference results of AudioInceptionNeXt and UniRepLKNet through adaptive mean fusion.

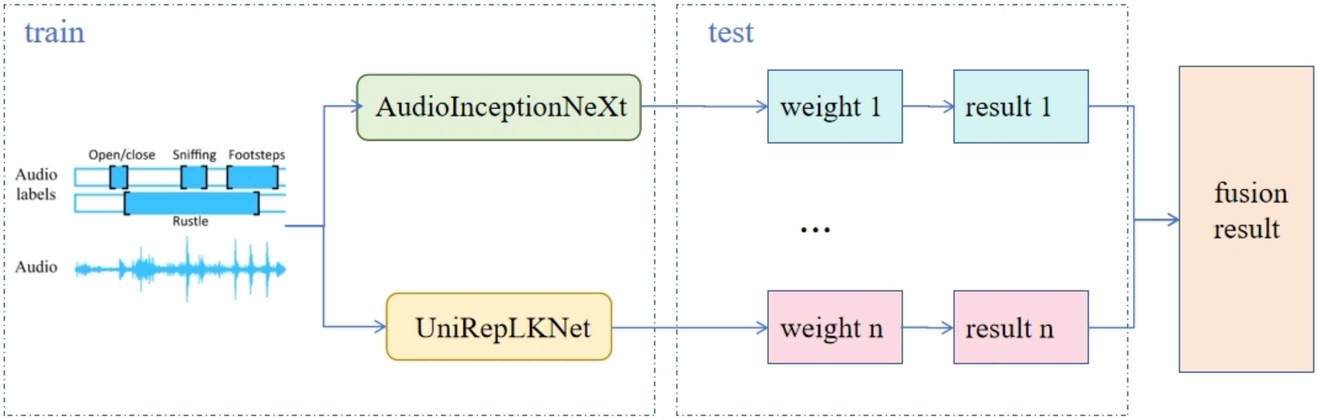

Figure 1. Proposed our approach architecture

This approach dynamically balances the complementary strengths of the two models—AudioInceptionNeXt's multi-scale local feature sensitivity and UniRepLKNet's long-range contextual modeling capability—thereby improving cross-scenario generalization and prediction consistency.

3. Experimental

3.1. Dataset and Evaluation Metrics

The EPIC-SOUNDS dataset for the Audio-Based Interaction Recognition challenge is sourced from the EPIC-KITCHENS project. It contains 100 hours of Full HD video and audio data, with recordings from 45 kitchens in 4 cities captured by head - mounted cameras. The dataset has 20 million frames and 90,000 action segments, and is annotated for 44 human - object interaction classes in kitchen scenarios, based on multi - language narrations. It is split into train, validation and test sets, where the test set includes unseen participants and rare action classes, aiming to test the generalization ability of models.

In the EPIC - SOUNDS Audio - Based Interaction Recognition competition, the Top - 1 accuracy is a crucial metric. It measures the proportion of audio samples for which the model's most likely predicted interaction class exactly matches the true class. Mathematically, if we denote the total number of test audio samples as N, and the number of samples where the top - ranked predicted class by the model is the correct one as C, then the Top - 1 accuracy A_{top-1} is calculated by the formula $A_{top-1} = \frac{C}{N} \times 100\%$. A higher Top - 1 accuracy indicates that the model is more precise in making its single, best - guess predictions for the audio - based human - object interaction classes in the dataset.

3.2. Single model performance

We tested the performance of each model under different parameter settings respectively and obtained the results shown in Table 1. In the testing of single models, the Top-1 accuracy can reach up to 55.83% at maximum, which indicates that these two models have significant advantages in the decoupled capture of long-term global semantics and short-term local details in audio signals.

Table 1. Single model performance

index	Model	ACC@1
0	AudioInceptionNeXt	55.72
1	AudioInceptionNeXt	55.70
2	AudioInceptionNeXt	55.60
3	UniRepLKNet	55.76
4	UniRepLKNet	55.66
5	UniRepLKNet	55.83

3.3. model fusion performance

We performed fusion on the models in Table 2 and used a weighted method to calculate the mean of the 44 category scores across multiple models. A progressive fusion strategy was adopted, where independent models were first preliminarily fused, followed by weighted processing of the fusion results to obtain the final output. Experimental results show that this approach increased the Top-1 accuracy to 57.51%. By introducing a dynamic weight allocation mechanism, our method effectively achieves synergistic integration of the feature expression advantages of different architectures across models.

Table 2. model fusion performance

index	fusion models	weight	ACC@1
6	1,2,3,4	[0.2 0.2 0.2 0.2 0.2]	57.27
7	6,1,2,3	[0.25 0.25 0.25 0.25]	56.44
8	6,7,1,5	[0.45 0.35 0.1 0.1]	57.19
9	2,3,4,5	[0.15 0.25 0.25 0.35]	57.36
10	6,8,9	[0.3 0.3 0.4]	57.51

4. conclusion

This report presents a methodology centered on model fusion, aiming to enhance the performance of audio classification tasks. The approach employs a late-stage ensemble strategy, adaptively integrating the inference outputs of dual models through a mean fusion mechanism to dynamically balance the sensitivity of the AudioInceptionNeXt model to local multi-scale features and the capability of the UniRepLKNet model to capture long-term contextual relationships. Ultimately, this solution achieved a score of 57.51 % in the 2025 EPIC-SOUNDS Audio-Based Interaction Recognition Challenge.

References

[1] K W Lau, Y A U Rehman, Y Xie, et al. Audioinceptionnext: Tcl ai lab submission to epic-sound audio-based interaction-recognition challenge 2023. *arXiv preprint*, arXiv:2307.07265, 2023. 1

[2] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5513–5524, 2024. 1

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2018. 1

[4] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022. 1

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020. 1

[6] Xiaohan Ding, X. Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11965, 2022. 2