ICASSP 2023

Feature-Rich Audio Model Inversion for Data-Free Knowledge Distillation Towards General Sound Classification

平安科技
PING AN TECHNOLOGY

Zuheng Kang **Yayun He** Jianzong Wang Junqing Peng Xiaoyang Qu Jing Xiao

Ping An Technology (Shenzhen) Co., Ltd.

Abstract

Data-Free Knowledge Distillation (DFKD) has recently attracted growing attention in the academic community, especially with major breakthroughs in computer vision. Despite promising results, the technique has not been well applied to audio and signal processing. Due to the variable duration of audio signals, it has its own unique way of modeling. In this work, we propose feature-rich audio model inversion (FRAMI), a data-free knowledge distillation framework for general sound classification tasks. It first generates high-quality and feature-rich Mel-spectrograms through a feature-invariant contrastive loss. Then, the hidden states before and after the statistics pooling layer are reused when knowledge distillation is performed on these feature-rich samples. Experimental results on the Urbansound8k, ESC-50, and audioMNIST datasets demonstrate that FRAMI can generate feature-rich samples. Meanwhile, the accuracy of the student model is further improved by reusing the hidden state and significantly outperforms the baseline method.

Feature Invariance Contrastive Inversion

Audio and image processing are intrinsically different, shown in Table 1. When generating large spectra in a traditional model inversion, the feature distribution in the time dimension is usually very sparse. This means that only a few features in a short period of time are enough to determine its category.

To improve recognition accuracy, we updated the method of contrastive inversion method [1] that can guarantee the feature richness of the generated audio samples in the time dimension.

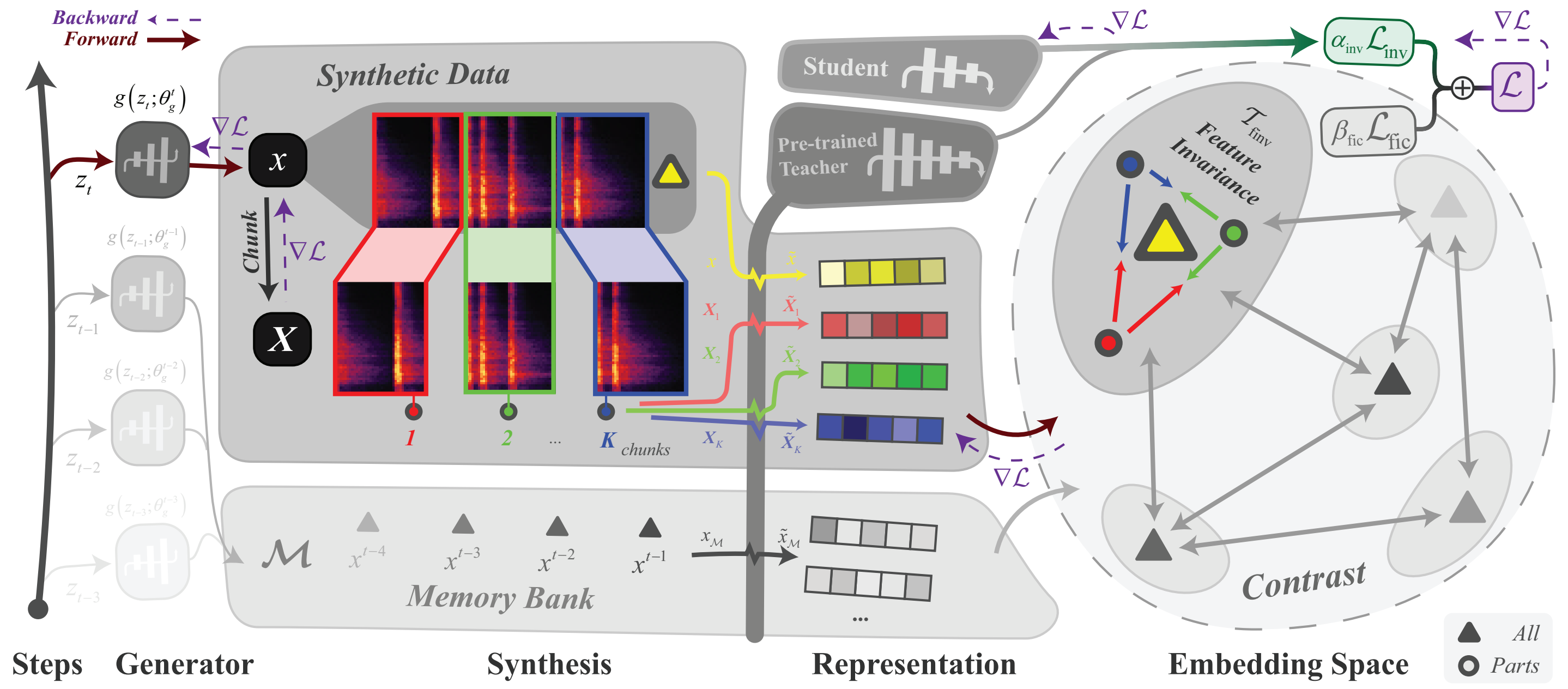

Figure 1. Feature invariance contrastive inversion overview.

We propose a feature invariance contrastive inversion method (shown in Figure 1) to ensure the feature richness of the generated samples, avoiding the problem of sparse audio features produced by traditional methods. Specifically, the currently synthesized data is first cut into K chunks in the time dimension shown in the red, green blue spectral.

Table 1. General sound characteristics.

Image/Picture		General Sound	
		Time Dependent (TD)	Time Independent (TID)
Size	*Fixed size*	*Variable length in time dimension*	
Measure	*Length and width with same measure*	*Time and feature with different measures*	
Content	*Each section with different content*	*Different content...*	*Same content at different times*
Information	*Feature-Rich*	*Feature-Rich*	*Feature-Sparse*

Then each chunk is converted into an embeddings by a pre-trained teacher model. According to the characteristics of time-independent audio, the audio features in each block should be similar. In other words, their feature embeddings should be as close to each other as possible, denoted as $\mathcal{T}_{\text{finv}}$ in Equation 1. Based on this, we construct a feature-invariant contrast inversion loss (denoted as $\mathcal{L}_{\text{fic}}$ in Equation 2) to guide the model to synthesize feature-rich and reliable audio samples and ultimately achieve higher classification accuracy.

$$\mathcal{T}_{\text{finv}}\left(\tilde{\boldsymbol{X}}\right) = \mathbb{E}_{\left(\tilde{x}_i,\tilde{x}_j\in\tilde{\boldsymbol{X}}\right)\wedge(i<j)}\left[\cos\left(\tilde{x}_i,\tilde{x}_j\right)\right] \tag{1}$$

$$\mathcal{L}_{\text{fic}}\left(\mathcal{X}'\right) = -\mathbb{E}_{x_i\in\mathcal{X}'}\left[\log\frac{\exp\left(\left(\cos\left(\tilde{x}_i,\tilde{x}_i^{+}\right)+\mathcal{T}_{\text{finv}}\left(\tilde{\boldsymbol{X}}_i\right)\right)/\tau\right)}{\sum_j\exp\left(\cos\left(\tilde{x}_i,\tilde{x}_j^{-}\right)/\tau\right)}\right] \tag{2}$$

Reused Teacher Backend Knowledge Distillation

In sound classification tasks, there is usually a statistics pooling layer to eliminate the time dimension. It can suppress irrelevant patterns while highlighting discriminative ones. We can make use of information on both frame-level hidden states and utterance-level hidden states, then calculate the mean and variance of these two types of hidden states separately to help your student model learn better from your teacher model. The procedure is shown in Figure 2.

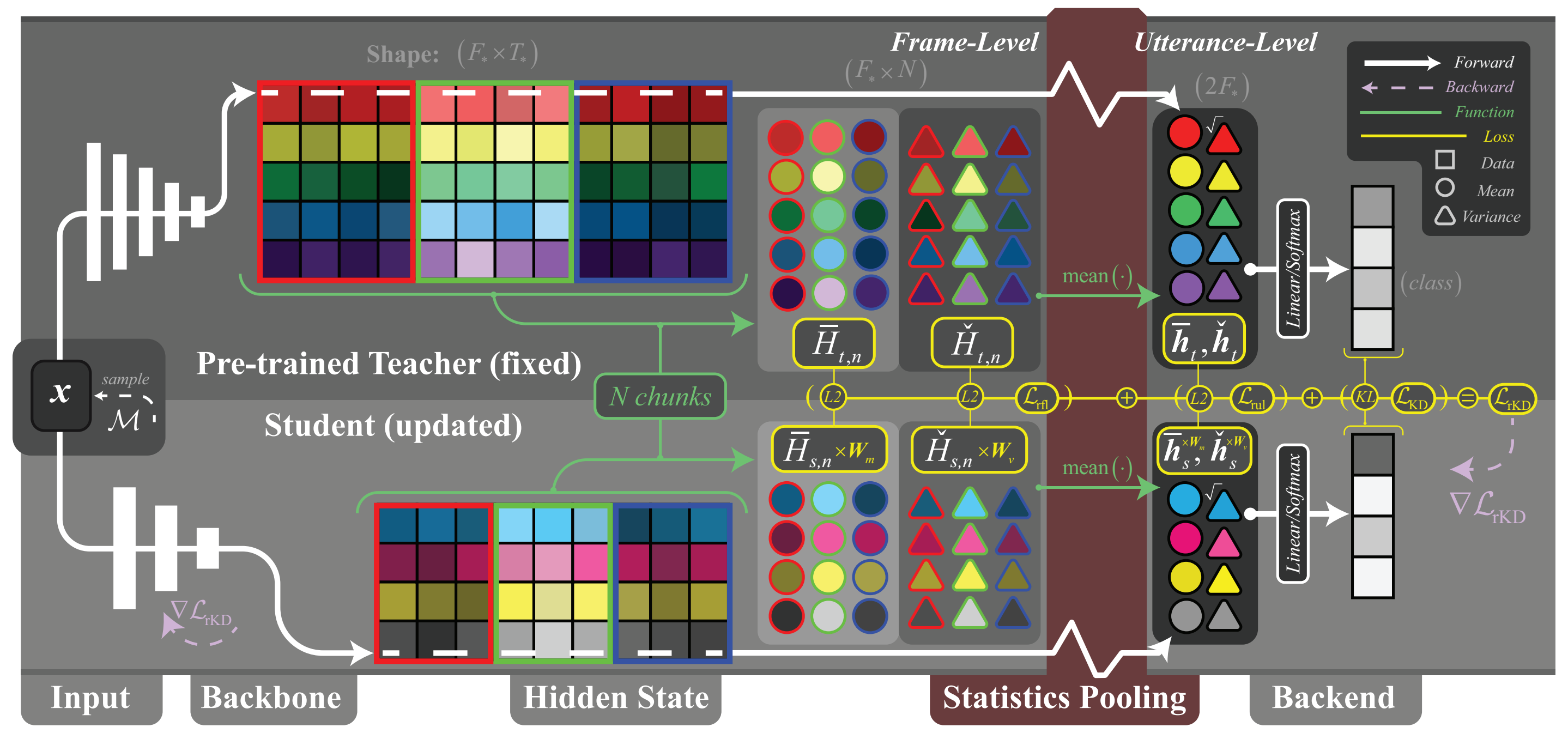

Figure 2. Reused teacher backend KD overview.

Experiments

Experimental results on Urban-sound 8k, ESC50 and Audio MNIST are reported in Table 2. The table above shows that the teacher outperforms the student and KD [2] when trained with data. However, in the case of data-free, Deep inversion (ADI) [3] performs very poor, much worse than other methods. The FRAMI framework proposed in this paper is stronger than KD, and sometimes even better than the accuracy of the teacher. In our analysis, the model inversion not only restores real samples but also mixes various features to create more samples with richer features.

In Table 3, we used a feature invariance term called $\mathcal{T}_{\text{finv}}$ to improve the accuracy of the student model. We found that this term had a greater impact on US8k than on ESC50. We also found that $\mathcal{T}_{\text{finv}}$ created more feature-rich samples, which allowed more valuable information to be captured at the frame-level, thereby improving accuracy.

Table 2. Comparison of different methods on 3 datasets.

	Method		Data Driven			Data Free	
Acc(%)	Teacher	Student	Teacher	Student	KD[2]	ADI[3]	FRAMI
US8k	res34	res18	80.05	76.70	75.63	66.67	**79.93**
	wrn40	wrn16	77.65	72.63	75.27	63.44	**78.14**
ESC50	res34	res18	68.25	62.50	67.00	59.25	**67.75**
	wrn40	wrn16	64.50	60.75	63.50	58.75	**63.50**
AMnst	res34	res18	99.90	99.53	99.87	99.10	**99.80**
	wrn40	wrn16	99.83	99.47	99.80	99.03	**99.73**

Table 3. Ablation study of FRAMI with or without $\mathcal{T}_{\text{finv}}$ and Reused for WRN-based models.

Acc(%)	US8k	ESC50	AMnst
FRAMI (full)	**78.14**	**63.50**	-
w/o $\mathcal{T}_{\text{finv}}$	76.82	63.25	**99.73**
w/o Reused	77.54	62.75	-
w/o $\mathcal{T}_{\text{finv}}$ w/o Reused	76.58	63.00	99.67

The Figure 3 shows a sample of the resulting spectrum, which is intelligible to the human ear.

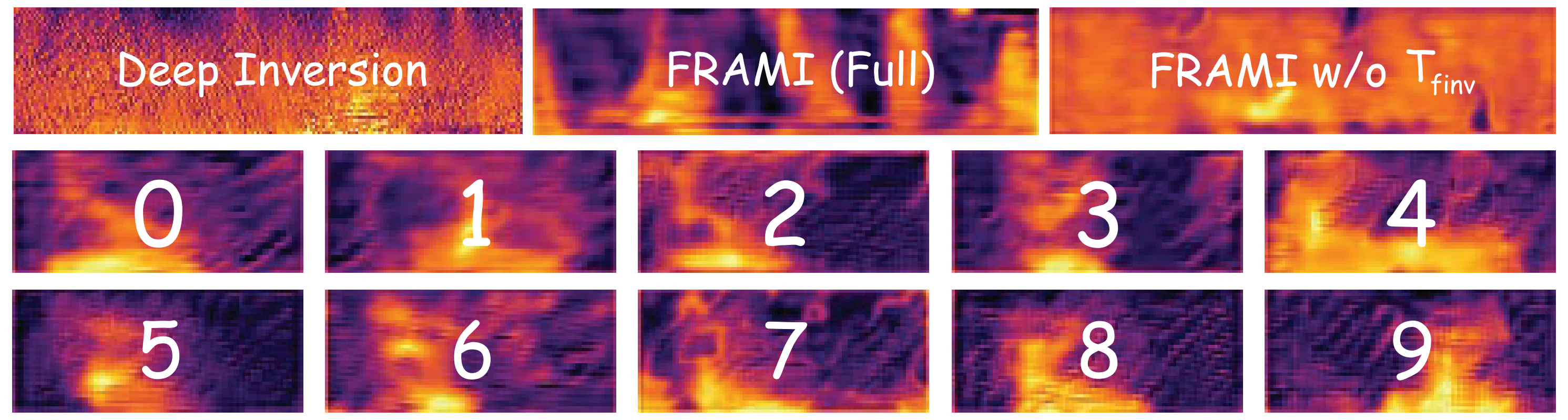

Figure 3. The generated spectral samples.

Conclusions

In this paper, we propose FRAMI, a framework for data-free knowledge distillation for general sound classification tasks. We design a feature invariance contrastive inversion to ensure the feature richness of the generated samples, avoiding the problem of sparse audio features produced by traditional methods. In knowledge distillation, the student model uses these feature-rich samples to mimic the teacher model at a deeper level by simultaneously learning the hidden states before and after the statistics pooling layer. Experimental results on Urbansound8k, ESC-50, and audioMNIST demonstrate that both methods, alone or in combination, improve the accuracy of the student model. Although this is a simple, preliminary exploration, we validate the feasibility of data-free knowledge distillation in general sound classification and are convinced that it will be extended to more audio models and more audio scenarios.

References

[1] G. Fang, J. Song, X. Wang, C. Shen, X. Wang, and M. Song, "Contrastive model inversion for data-free knowledge distillation," in *IJCAI*, 2021.

[2] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," in *NIPS 2014 Deep Learning Workshops*, 2015.

[3] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *CVPR*, 2020, pp. 8715–8724.

https://tech.pingan.com 12 Science and Technology South Rd, Nanshan District, Shenzhen, China, 510280 kangzuheng896@pingan.com.cn, heyayun097@pingan.com.cn