PLAY IT BACK: ITERATIVE ATTENTION FOR AUDIO RECOGNITION

Alexandros Stergiou[1,2,], Dima Damen[3]*

[1]Vrije University of Brussels, Belgium   [2]Interuniversity Microelectronics Centre, Leuven, Belgium
[3]University of Bristol, United Kingdom

ABSTRACT

A key function of auditory cognition is the association of characteristic sounds with their corresponding semantics over time. Humans attempting to discriminate between fine-grained audio categories, often replay the same discriminative sounds to increase their prediction confidence. We propose an end-to-end attention-based architecture that through selective repetition attends over the most discriminative sounds across the audio sequence. Our model initially uses the full audio sequence and iteratively refines the temporal segments replayed based on slot attention. At each playback, the selected segments are replayed using a smaller hop length which represents higher resolution features within these segments. We show that our method can consistently achieve state-of-the-art performance across three audio-classification benchmarks: AudioSet, VGG-Sound, and EPIC-KITCHENS-100. [1]

Index Terms— Audio classification, playback, attention

1. INTRODUCTION

Audio recognition is the task of categorizing audio with discrete labels that semantically represent the emitted sounds. This includes significant challenges considering the similarity in object sounds (e.g. boat motors and road vehicles), musical instruments (e.g. guitar, banjo, and ukulele), human (e.g. wail and groan), or animal (e.g. yip and growl) sounds.

In everyday life, we repeat parts of songs or ask for someone to repeat themselves to better understand audio. This relates to the development of echoic memory which is responsible for the memorization of sounds [1, 2]. Therefore, repeated listens and replays of sound stimulants [3] are an essential part of learning and associating sound patterns.

Driven by the perception of sound through echoic memory and the recent success of Vision Transformers (ViT) [4] at utilizing global context information, we propose an end-to-end attention-based model that recognizes sounds through discovering and playing back the most informative sounds from the audio sequence, as shown in Figure 1. We use slots [5] to attend to category-relevant sounds in the input sequence. These slots select the time segments to be replayed.

Fig. 1: **Playback of discriminative sounds**. Given an audio sequence, the most relevant sounds are selected and played back at reduced hop length. The generated playbacks attend solely informative sounds at a higher temporal resolution.

Coarser features from earlier playbacks are memorized alongside finer (i.e. higher-temporal resolution) features from later playbacks with the use of a transformer decoder.

Our contributions are as follows: i) We propose to select and replay relevant audio features with decreased hop lengths, slowing down relevant parts of the audio. ii) We propose an end-to-end transformer architecture for audio recognition that jointly selects and attends to multiple audio replays, and refines the final class predictions. iii) Our method achieves state-of-the-art performance on AudioSet [6], VGG-Sound [7], and EPIC-KITCHENS-100 [8].

2. RELATED WORK

Audio recognition. A popular approach for audio classification has been the use of convolutional networks, previously used for image-based object recognition [9, 10, 11] or video classification [12] tasks, to learn features from audio spectrograms. The introduction of Transformer-based architectures has further given rise to their adaptation for audio recognition by works relying on hybrid architectures [13, 14, 15]. Similar attempts have also built on image-pretrained Transformer models for attending audio spectrograms [16, 17]. [18] incor-

---

Fig. 2: **PlayItBack architecture**. The spectrogram of the full audio sequence (top) is replayed by focusing on discriminative features and reducing the hop length to capture finer temporal details (bottom). During each playback, spectrogram patches are tokenized $\mathbf{x}_i$ and appended patch (frequency and temporal) positional encodings (P). Several multi-head attention layers are used to encode features $\mathbf{z}_i$. Slot attention $\mathcal{G}(\mathbf{z}_i)$ then discovers discriminative temporal segments. These are considered input to the next playback. To combine decisions between playbacks, a recurrent Transformer Decoder $\mathcal{D}(\mathbf{z}_i)$ takes previously decoded features from i and the encoded features in $i + 1$ playback appended with patch encodings (PP). PlayItBack is trained by classification loss $\mathcal{L}_{CLS_i}$, regularized by the weighted sum of ranking losses $\mathcal{L}_{rank_i}$ between i and $\{1, ..., i-1\}$ playbacks.

porated an additional video modality to improve performance. Recently, [19] have also studied the effects of different hop lengths on the temporal resolution of spectrograms.

In contrast to the majority of previous works, we focus on identifying relevant and irrelevant sounds. The irrelevant sounds are removed, while relevant segments are slowed and replayed with predictions calculated across *playbacks*.

Selecting discriminative features. Modeling discriminative features has been a central focus of image recognition methods. [20] generated features from multiple scales selecting the best-suited features per scale. [21] proposed the aggregation of features from image regions cropped based on class saliencies. [22] applied distortion grids on the cropped regions. Most similar to our work, [23] used a recurrent CNN to select image regions that are attended to in follow-up scales. In their work, only a single region was selected per scale. Instead, we identify multiple discriminative sound segments which we combine to form the next playback.

We describe our PlayItBack method next.

3. METHOD

In this section, we describe the proposed PlayItBack architecture, depicted in Figure 2. We compute the mel-spectrogram for a given audio sequence resulting in an $F \times T$ representation of frequency F and time T, and extract k non-overlapping patches. We project the patches to feature tokens $\mathbf{x}_i \in \mathbb{R}^D$, where $D = FT$. A transformer encoder $\mathcal{B}$ is used to encode these into features $\mathbf{z}_i$. Slot attention $\mathcal{G}$ is applied to $\mathbf{z}_i$ to select the discriminative regions which, at the next playback, will be

slowed by decreasing the hop length in the spectrogram. The Decoder $\mathcal{D}$ then relates features across playbacks.

Transformer Encoder. Given linear projections $\mathbf{x}_i$, we use frequency and temporal patch positional encodings P. The encoder network $\mathcal{B}$ extracts representations for each playback, $\mathbf{z}_i = \mathcal{B}(\mathbf{x}_i) \in \mathbb{R}^{d \times C}$, with $d < FT$ resolution and C channels.

Slot attention. We use slot attention [5] $\mathcal{G}$ to iteratively map the resulting feature vectors $\mathbf{z}_i$ from each playback to two slot vectors $\mathbf{s}_{lj}$ corresponding to the informative $\mathbf{s}_{1j}$ and uninformative $\mathbf{s}_{2j}$ temporal segments of the audio input respectively. We use $j \in \{1, ..., J\}$ to denote slot iterations. The query $\mathbf{Q}_{lj} = MLP(LN(\mathbf{s}_{lj-1}))$, key $\mathbf{K}_{lj} = MLP(LN(\mathbf{z}_i))$ and value $\mathbf{V}_{lj} = MLP(LN(\mathbf{z}_i))$ use Layer Normalization $LN(\cdot)$ followed by Multi-Layer Perceptron $MLP(\cdot)$ to map the features $\mathbf{z}_i$ and slots $\mathbf{s}_{lj}$ vectors to a common dimension d. We set the softmax temperature based on a fixed value $\sqrt{d}$.

$$\mathbf{h}_{lj} = GRU\left(\frac{\mathbf{a}_{lj}\,\mathbf{V}_{lj}}{\sum\limits_{m \in \{1,2\}} \mathbf{a}_{mj}}\right), \text{ where } \mathbf{a}_{lj} = Attn\left(\frac{\mathbf{K}_{lj}\,\mathbf{Q}_{lj}^T}{\sqrt{d}}\right) \quad (1)$$

A Gated Recurrent Unit (GRU) with two hidden units is used at each slot iteration updating the slot hidden states $\mathbf{h}_{lj}$ as in [5]. A linear transformation alongside a residual connection is used for the slots $\mathbf{s}_{lj} = \mathbf{s}_{lj-1} + MLP(LN(\mathbf{h}_{lj}))$.

We train the two slots so $\mathbf{s}_1$ attends to informative audio while $\mathbf{s}_2$ captures the remaining audio. This is achieved by combining $\mathcal{G}(\mathbf{z}_i)_1 = \mathbf{s}_1$ and the *inverse* of the uninformative slot $\mathcal{G}(\mathbf{z}_i)_2 = \mathbf{s}_2$ to create the attention matrix: $\mathbf{M} = Attn(\mathcal{G}(\mathbf{z}_i)_1^T \mathcal{G}(\mathbf{z}_i)_2^{-1})$. We normalize and rescale the

Model	Backbone	Train set	mAP
Audio-only models			
MAE-AST [24]	ViT-B [4]	mini-AS	30.6
Perceiver [25]	Perceiver	AS-2M	38.4
Conformer [14]	Conformer	AS-2M	41.1
PANN [10]	ResNet38 [26]	AS-2M	43.4
MBT [18]	ViT-B	AS-500K	44.3
PSLA [9]	EffNet-B2 [27]	AS-2M	44.4
PaSST [17]	DeiT-B [28]	AS-2M	47.1
HTS-AT [16]	Swin-T [29]	AS-2M	47.1
MaskSpec [30]	ViT-B	AS-2M	47.1
Audio-MAE [31]	ViT-B	AS-2M	47.3
PlayItBackX3	MViTv2-B [32]	AS-500K	**47.7**

Table 1: **Comparisons to state-of-the-art audio-only models on AudioSet**. We report the mean average precision (mAP) alongside the backbone and training set used.

Model	top-1	top-5	mAP	AUC	d-prime
Audio-only models					
McDonnell & Gao [33]	39.7	71.6	40.3	0.963	2.532
Peng et al. (A) [34]	44.3	-	48.4	-	-
ResNet-101 [12]	45.6	72.3	47.6	0.968	2.615
Chen et al. [7]	51.0	76.4	53.2	0.973	2.735
MBT (A) [18]	52.3	78.1	-	-	-
Slow-Fast [12]	52.4	78.1	54.4	0.974	2.761
PlayItBackX3	**53.7**	**79.2**	**56.1**	**0.978**	**2.846**
Models trained with additional modalities					
Peng et al. (AV) [34]	50.6	-	53.9	-	-
PolyViT [35]	51.7	-	-	-	-
MBT (AV)	64.1	85.6	-	-	-

Table 2: **Comparisons to state-of-the-art models on VGG-Sound**. We report the top-1 and top-5 accuracies (%) alongside mAP, the AUC and d-prime.

main diagonal $diag(\mathbf{M})$ by interpolation so that it matches the temporal dimension of $\mathbf{x}_i$. Activations above the normalized average (> 0.5) are selected for the segments in $\mathbf{x}_{i+1}$.

Transformer Decoder. Given the extracted encoder features $\mathbf{z}_i$, the decoder transformer $\mathcal{D}$ relates information across playbacks. Positional encodings based on patches and the playback number are added to $\mathbf{z}_i$. Considering the iterative nature of the PlayItBack model, cross-attending [25] information over playbacks enables the model to retain general features and associate patterns that are common. For the decoder, we define the query from the previous playback as $\mathbf{Q}_i = MLP(LN(\mathbf{v}_i))$, where $\mathbf{v}_1$ is initialized with a latent vector then updated at each playback $\mathbf{v}_i = \mathcal{D}(\mathbf{z}_{i-1}, \mathbf{v}_{i-1})$, where $i > 1$, key $\mathbf{K}_i = MLP(LN(\mathbf{z}_i))$ and value $\mathbf{V}_i = MLP(LN(\mathbf{z}_i))$ for the cross attention. This is followed by a self-attention block. The decoder features are then passed to a classifier shared across playbacks.

Classification and rank loss. We use an inter-playback weighted ranking loss $\mathcal{L}_{\text{rank}(i)}$ for forcing the network to attain more confident predictions in later playbacks. The ranking loss $\mathcal{L}_{\text{rank}(i)}$, uses the pair-wise class probabilities $p(\omega)_i$ and $p(\omega)_m \forall m \in \{1, ..., i-1\}$ for the correct class label ω. We compute the probability difference $\mathcal{L}_{\text{rank}(m \leftrightarrow i)}$ between the ith playback and all previous playbacks.

$$\mathcal{L}_{\text{rank}(i)} = \sum_{m=1}^{i-1} \lambda_m \, max(0, \gamma - p(\omega)_i + p(\omega)_m) \quad (2)$$

The ranking loss thus uses predictions from the previous playbacks as a reference with the expectation that $p(\omega)_i > p(\omega)_m + \gamma$, i.e. subsequent playbacks always increase confidence, where γ is the ranking loss's margin. For stability in training, we include a weight $\lambda_m = \frac{1}{i-m}$ computed based on the difference between the playback indices.

We combine $\mathcal{L}_{rank(i)}$ with an inter-playback cross-

entropy loss $\mathcal{L}_{\text{CLS}(i)}$ and define our multi-task loss as:

$$\mathcal{L} = \mathcal{L}_{\text{CLS}(1)} + \sum_{i=2}^{N} \beta \, \mathcal{L}_{\text{CLS}(i)} + (1 - \beta) \, \mathcal{L}_{\text{rank}(i)} \quad (3)$$

where β is a weighting parameter for the aggregation of the cross-entropy and ranking losses. During inference, our model uses the average of all predictions across playbacks.

4. EXPERIMENTS

Datasets We evaluate our proposed PlayItBack architecture on three large-scale datasets. **AudioSet** [6] is composed of 2M 10s audio clips from YouTube annotated with 527 classes (AS-2M). Because of the high imbalance of the dataset, we instead train with the proposed AS-500K [18]. **VGG-Sound** [7] consists of 200k clips of 10s length with 309 labels corresponding to human actions, objects and interactions. **EPIC-KITCHENS-100** [8] includes 90k clips of hand-object interactions labeled with 97 verb, 300 noun classes, and 4025 action classes. The clip length is variable and 2.6s on average.

Evaluation metrics. For AudioSet along the lines of previous works, we use the mean average precision (mAP). For VGG-Sound, as in [12], we report the top-1/5 % accuracies, mAP, AUC, and d-prime. For EPIC-KITCHENS-100 we report the top-1/5 % accuracies for the verb, noun, and action labels.

Implementation details. We use PlayItBackX3 with $N=3$ as our model for comparative evaluation, with ablations showcasing that this produces the best accuracy (top-1)/compute (GFLOPs) trade-off. We use the 24-layer MViTv2-B [32] as our default encoder[2]. We note that due to the fixed number of 2D patches used by MViTv2, the spectrogram dimensions remain constant throughout playbacks. For all experiments, we set the ranking margin $\gamma = 0.05$, $J = 3$ slot iterations,

---

[2]The flattened vector size is d=50 and the number of features is C=768

Model	GFLOPs	verb		noun		action	
		top-1	top-5	top-1	top-5	top-1	top5
Damen et al. [8]	N/A	42.6	75.8	22.3	44.6	14.5	28.2
MBT (A) [18]	34.2	44.3	-	22.4	-	13.0	-
Slow-Fast [12]	35.1	46.5	78.3	22.8	44.9	15.4	28.6
PlayItBackX3	122.8	**47.0**	**78.7**	**23.1**	**45.1**	**15.9**	**29.2**

Table 3: **Comparisons to state-of-the-art for EPIC-KITCHENS-100**. We report the top-1 and top-5 accuracies for the verb, noun, and action labels.

Model	freq.	top-1	top-5	mAP	AUC	d-prime
PlayItBackX0	32kHz	52.1	77.8	54.7	0.970	2.757
PlayItBackX0	16kHz	51.8	77.4	54.3	0.966	2.743
PlayItBackX1	16kHz	52.5	78.3	55.1	0.972	2.789
PlayItBackX2	16kHz	53.2	78.7	55.5	0.976	2.810
PlayItBackX3	16kHz	**53.7**	**79.2**	**56.1**	**0.978**	**2.846**

Table 4: **Frequency to playbacks on VGG-Sound** given top-1 and top-5 accuracies, mAP, AUC, and d-prime.

and $\beta = 0.7$. As in [12] we use spectrograms with frequency dimension of 128 corresponding to inputs of size $128 \times 100S$ for S seconds of audio. Our initial spectrograms are created based on the same hop length of 10ms, and 16kHz frequency as in [12, 18, 31]. For subsequent iterations, we reduce the hop length by 1ms at each iteration.

We train for 50 epochs with Mixup [36] ($\alpha = 0.3$) and base learning rate of 0.5 for AudioSet and 0.01 for VGG-Sound & EPIC-KITCHENS-100. We use warm-up for the first 2.5 epochs, a decayed cosine schedule, batch size of 64 with SGD, momentum set to 0.9, and $1e^{-4}$ weight decay.

Results. We compare PlayItBack to current state-of-the-art models on **AudioSet** in Table 1. PlayItBackX3 achieves the best performance in comparison to other models.

We report results on **VGG-Sound** in Table 2. PlayItBackX3 performs favorably to in-domain audio models. Compared to the previously top-performing SlowFast model [12], we observe a +1.3%p. top-1 accuracy improvement. Our model is only outperformed by the multi-modal (audio-visual) version of MBT (AV) [18]. However, PlayItBackX3 outperforms MBT (A) in the audio-only setting.

In Table 3, we compare to audio-classification methods on **EPIC-KITCHENS-100**. We observe that the relative improvement in performance varies across datasets (higher performance gains are observed in AudioSet and VGG-Sound, while somewhat smaller on EPIC-KITCHENS-100). We believe that this is due to EPIC-KITCHENS-100 containing audio segments of 2.6s in length on average, compared to 10s durations of AudioSet and VGG-Sound. As the segments are already shorter, they intuitively benefit less from further playbacks by focusing on discriminative regions. Even in such settings, PlayItBackX3 demonstrates a moderate but consistent performance improvement.

Ablations. Table 1 demonstrates that while PlayItBack uses a sampling frequency of 16kHz, it can outperform HTS-AT [16] and PaSST [17] which are trained on sampling

J	top-1	GFLOPs
1	53.3	120.4
2	53.5	121.5
3	**53.7**	122.8

Table 5: **Number of slot attention iterations** (J) with respect to the top-1 accuracy and GFLOPs.

Fig. 3: **VGG-Sound top-1 accuracy over different playback-numbers (N)** with respect to the compute (in GFLOPs).

frequencies of 32kHz. We confirm this by ablating the impact on PlayItBack. Table 4 demonstrates that our proposed replays at 16kHz, can be a better performing strategy than increasing the number of samples per second over the entire audio sequence as in PlayItBackX0 trained with 32kHz.

Table 5 compares the performance achieved with different numbers of slot iterations J on VGG-Sound with PlayItBackX3. In general, moderate performance improvements can be achieved by increasing the number of slot attention iterations. The added computations also remain moderate with +2.6 GFLOPs from $J = 1$ to $J = 3$. In Figure 3, we investigate the impact of the number of playbacks (N) on the model performance. We use a decoder-only model ($N = 0$), alongside PlayItBackXN. Performance improvements are shown for $1 \leq N \leq 3$. Further N increases, come with performance drops, due to increased model complexity in tandem with the challenge of discovering salient information in very deep playbacks - and thus very small hop lengths. As showcased in this figure and across results, $N = 3$ offers the best performance. The performance remains consistent over multiple runs and across datasets.

5. CONCLUSIONS

We propose an end-to-end attention-based architecture for audio recognition. Our PlayItBack model uses information from the full audio sequence to iteratively discover segments that are relevant to the sound. Audio segments are discovered with slot attention and amplified in the next iteration (playback). A transformer decoder is used to relate information across playbacks. We demonstrate the advantages of our PlayItBack approach through extensive experiments on AudioSet, VGG-Sound, and EPIC-KITCHENS-100 and ablation studies.

Future work can explore the selection strategy for the number of playbacks, which might vary per audio sample. We hope PlayItBack can trigger insights into similar approaches for other audio signals such as speech as well as audio-visual fine-grained understanding.

6. REFERENCES

[1] Terry Clark, "Echoic memory explored and applied," *Journal of services marketing*, 1987.

[2] Rael D Strous, Nelson Cowan, Walter Ritter, and Daniel C Javitt, "Auditory sensory (" echoic") memory dysfunction in schizophrenia.," *The American journal of psychiatry*, 1995.

[3] Gabriel A Radvansky, *Human Memory*, Psychology Press, 2005.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[5] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, et al., "Object-centric learning with slot attention," in *NeuIPS*, 2020.

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, et al., "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, et al., "Rescaling egocentric vision: collection, pipeline and challenges for EPIC-KITCHENS-100," *IJCV*, 2022.

[9] Yuan Gong, Yu-An Chung, and James Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM TASLP*, 2021.

[10] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, et al., "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, 2020.

[11] Yun Wang, Juncheng Li, and Florian Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP*, 2019.

[12] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Slow-fast auditory streams for audio recognition," in *ICASSP*, 2021.

[13] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM TASLP*, 2020.

[14] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[15] Koichi Miyazaki, Tatsuya Komatsu, Tomoki Hayashi, Shinji Watanabe, et al., "Convolution augmented transformer for semi-supervised sound event detection," in *DCASE*, 2020.

[16] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, et al., "Htsat: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP*, 2022.

[17] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint*, 2021.

[18] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, et al., "Attention bottlenecks for multimodal fusion," *NeurIPS*, 2021.

[19] Haohe Liu, Xubo Liu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley, "Learning the spectrogram temporal resolution for audio classification," *arXiv preprint arXiv:2210.01719*, 2022.

[20] Shenlong Wang, Linjie Luo, Ning Zhang, and Jia Li, "Autoscaler: Scale-attention networks for visual correspondence," in *BMVC*, 2017.

[21] Amir Rosenfeld and Shimon Ullman, "Visual concept recognition and localization via iterative introspection," in *ACCV*, 2016.

[22] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba, "Learning to zoom: a saliency-based sampling layer for neural networks," in *ECCV*, 2018.

[23] Jianlong Fu, Heliang Zheng, and Tao Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017.

[24] Alan Baade, Puyuan Peng, and David Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," *arXiv preprint*, 2022.

[25] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, et al., "Perceiver: General perception with iterative attention," in *ICML*, 2021.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[27] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, et al., "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[30] Dading Chong, Helin Wang, Peilin Zhou, and Qingcheng Zeng, "Masked spectrogram prediction for self-supervised audio pre-training," *arXiv preprint*, 2022.

[31] Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, et al., "Masked autoencoders that listen," in *NeurIPS*, 2022.

[32] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, et al., "Mvitv2: Improved multiscale vision transformers for classification and detection," in *CVPR*, 2022.

[33] Mark D McDonnell and Wei Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," in *ICASSP*, 2020.

[34] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022.

[35] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, et al., "Polyvit: Co-training vision transformers on images, videos and audio," *arXiv preprint*, 2021.

[36] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.