Masked Audio Modeling with CLAP and Multi-Objective Learning

Yifei Xin[1*], *Xiulian Peng*[2], *Yan Lu*[2]

[1]Peking University, Beijing, China
[2]Microsoft Research Asia, Beijing, China

xinyifei@stu.pku.edu.cn

Abstract

Most existing masked audio modeling (MAM) methods learn audio representations by masking and reconstructing local spectrogram patches. However, the reconstruction loss mainly accounts for the signal-level quality of the reconstructed spectrogram and is still limited in extracting high-level audio semantics. In this paper, we propose to enhance the semantic modeling of MAM by distilling cross-modality knowledge from contrastive language-audio pretraining (CLAP) representations for both masked and unmasked regions (MAM-CLAP) and leveraging a multi-objective learning strategy with a supervised classification branch (SupMAM), thereby providing more semantic knowledge for MAM and enabling it to effectively learn global features from labels. Experiments show that our methods significantly improve the performance on multiple downstream tasks. Furthermore, by combining our MAM-CLAP with SupMAM, we can achieve new state-of-the-art on various audio and speech classification tasks, exceeding previous self-supervised and supervised pretraining methods.

Index Terms: masked audio modeling, CLAP, multi-objective learning

1. Introduction

Recent years have seen considerable success in masked image modeling (MIM) [1, 2, 3] in the image domain, demonstrating promising results on a variety of downstream tasks, such as image classification, semantic segmentation, and object detection. MIM uses a pre-defined mask ratio to mask out image patches and adds the reconstruction supervision to the masked regions. In this study, we investigate the learning of general audio representations using MIM applied to the audio spectrogram, which we call masked audio modeling (MAM) [4, 5, 6]. MAM splits the audio spectrogram into patches along the time and frequency axes, enabling the model to learn both temporal and frequency structures.

Currently, MAM has been applied to general audio represenation learning, like SS-AST [7] and AudioMAE [4], showing good performance on audio and speech classification tasks. This indicates that MAM learns robust auditory representations for both speech and non-speech signals. However, it is generally believed that the reconstruction loss mainly accounts for signal-level quality at time-frequency domain but is still limited in learning high-level audio semantic knowledge [8, 9]. Besides, audio exhibits different characteristics from images. It is not easy to find an auditory-perception-aware patching strategy that fits various sound events with diverse scales. Masking with an extremely high ratio (80% mask ratio on AudioMAE [4]) would

unavoidably harm sound events with a short time/frequency span.

In addition to MAM, supervised pretraining is also widely studied for audio semantic modeling, which leverages either out-of domain data (e.g. ImageNet [10]) or in-domain audio data (e.g. AudioSet [11]) for pretraining. As for the out-of-domain supervised pretraining, HTS-AT [12] employs the Swin Transformer-based architecture [13] as the backbone, which obtains superior audio classification performance. For in-domain supervised pretraining, inspired by the language-vision pretraining method CLIP [14], CLAP [15, 16] utilizes a contrastive language-audio pretraining task to learn the text-enhanced audio representations with audio and text pairs. Compared with MAM, these methods learn different levels of semantics, which could complement each other.

In this paper, we aim to enhance the semantics for MAM by leveraging external cross-model information. To this end, we present MAM-CLAP, a simple yet effective framework that incorporates multimodal information into MAM. We take AudioMAE [4] as the benchmark for MAM, which learns to efficiently encode the visible patches into latent representations that carry essential information for reconstructing masked patches. Our motivation is that multimodal data can provide additional richer semantic knowledge. In this case, we apply the CLAP model to provide semantic guidance for MAM as it shows great robustness on various audio downstream tasks [16, 17, 18]. Moreover, unlike most MAM methods applying the reconstruction supervision on masked patches, we find that supervisions on both masked and visible patches with the CLAP target can achieve remarkable performance, which demonstrates that the visible patches can also effectively extract rich semantic information from CLAP, performing like the feature distillation.

Furthermore, most existing MAM methods [4, 5, 19] only learn feature interactions among patches. No global features, i.e., features that can represent the entire spectrogram, are explicitly learned during pretraining. To complement this, we extend MAM to a supervised setting (SupMAM) by adding a branch for supervised classification in parallel with the reconstruction target, thereby enabling MAM to effectively learn global features from labels. To make it sample-efficient, only a subset of visible patch features are used for the classification-based pretraining branch, where an average pooling is employed to get the global audio representation followed by a MLP for classification.

In a nutshell, our contributions are threefold:

- We provide a novel perspective that the CLAP-targeted feature supervision on the spectrogram patches is a good choice for audio representation learning. We develop a simple yet effective framework MAM-CLAP to improve the semantic
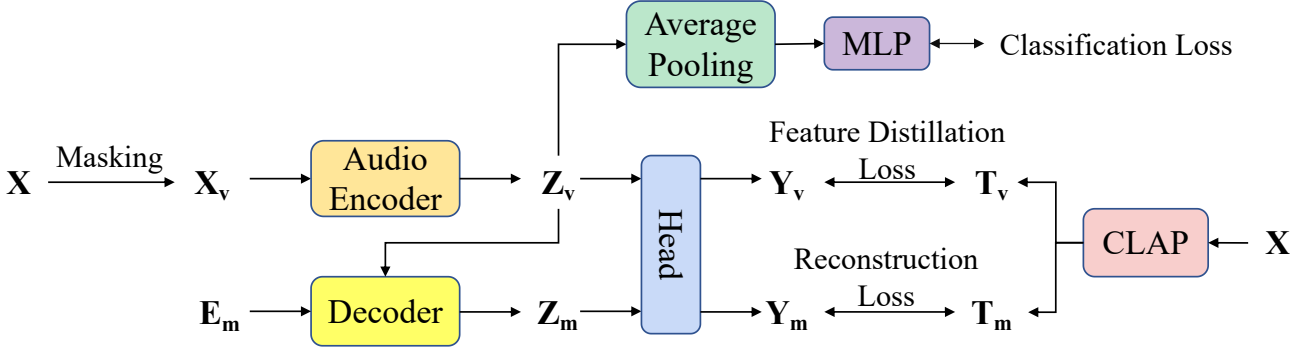

Figure 1: *The overview of our SupMAM-CLAP.*

modeling of MAM-based audio pre-training.

- We introduce a multi-objective learning method that extends AudioMAE to a fully-supervised setting (SupMAM) by adding a supervised classification branch. SupMAM only uses a subset of the visible patches to do classification rather than all patches used in standard supervised pre-training.

- By combining our MAM-CLAP with SupMAM, our SupMAM-CLAP can achieve state-of-the-art results across various audio and speech classification benchmarks.

2. SupMAM-CLAP

2.1. MAM-CLAP

The overview of our SupMAM-CLAP is illustrated in Figure 1. Let X denote an input spectrogram. We take AudioMAE [4] as the backbone. Our MAM-CLAP first embeds X into N patches, which are then randomly masked by a specific proportion γ. These N patches are naturally split into two non-overlapped sets, i.e., visible patches X_v and masked patches X_m, where $N = |v| + |m|$. The mask ratio is thus denoted as $\gamma = |m|/N$. The AudioMAE encoder $e(\cdot)$ maps the visible patches X_v to the latent representations Z_v. The decoder $g(\cdot)$ predicts the latent representations Z_m for the masked patches from mask tokens E_m, conditioned on the visible latent representation Z_v. After that, the predictions of visible patches Y_v and masked patches Y_m are obtained via a head $h(\cdot)$. In this work, we only use a fully-connected layer followed by a layer normalization in $h(\cdot)$. For the target supervision, we directly input the spectrogram X into the CLAP model $f(\cdot)$ to generate the target supervision T. T is then split into T_v and T_m corresponding to the positions of X_v and X_m. The optimization is applied on Y_v to approach T_v, and we also add the supervision on Y_m using T_m.

2.2. SupMAM

Unlike standard supervised pretraining methods [23, 20, 26] that use all patch features, SupMAM only uses a subset of the visible patches to do classification, which makes SupMAM more sample-efficient. This is mainly based on the intuition that humans can recognize sound events with partial information (a subset of patches). Also, from the perspective of data augmentation, random masking can generate different training samples for each iteration [27, 5, 28], serving as a strong regularization.

In this case, a global pooling first condenses the visible local patch features Z_v into the global representation G_v of the spectrogram, which is then used to predict the sound event la-

bels. The classification branch is complementary to the feature reconstruction branch as it can bring global feature learning into the framework. For the classification head, we use a two-layer MLP, with a batch normalization and a ReLU activation injected in-between to project the global representation into the logits of the predicted probabilities p_v. During fine-tuning, we only use the AudioMAE-based transformer encoder for downstream tasks.

2.3. Loss Functions

Most previous MAM methods [4, 6, 19] only apply the reconstruction supervision on the predictions of masked patches. With CLAP as the target, we supervise both visible and masked patches. The L2 loss functions are used as follows:

$$\mathcal{L}_{target} = \frac{1}{l_v}\|Y_v, T_v\|_2 + \frac{1}{l_m}\|Y_m, T_m\|_2, \quad (1)$$

where l_v and l_m denote the number of visible and masked tokens.

For SupMAM, we use the cross-entropy (CE) loss $\mathcal{L}_{cls}$ to supervise the training process with the classification label y:

$$\mathcal{L}_{cls} = CE(p_v/\tau, y), \quad (2)$$

where τ is a temperature parameter that controls the concentration level of the distribution, which is widely used in supervised and self-supervised feature learning.

Finally, our SupMAM-CLAP is optimized with both the CLAP-target loss and the classification loss, which simultaneously learns fine-grained local and global features. We use a weighted sum of these two loss terms as our total loss as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{target} + \lambda_{cls}\mathcal{L}_{cls}, \quad (3)$$

where λ_{cls} are weights to balance the objective. Note that when only using our SupMAM framework, $\mathcal{L}_{target}$ is replaced by the mean square error (MSE), i.e., the reconstruction loss of AudioMAE between the reconstructed and the input masked patches.

3. Experiments

We perform extensive evaluations on five tasks, including audio classification on AudioSet (AS-2M, AS-20K) [11] and Environmental Sound Classification (ESC-50) [29], and speech classification on Speech Commands (SPC-1 and SPC-2) [30].

Table 1: *Comparing with the SOTA models on audio and speech classification tasks. IN, AS, and LS denote the ImageNet, AudioSet, and LibriSpeech datasets, respectively. TA denotes the text-audio pairs for CLAP pretraining. † represents results from our implementation.*

Method	Data	AS-2M	AS-20K	ESC-50	KS1	KS2
No Pretraining						
PANN [20]	-	43.1	27.8	83.3	-	61.8
ERANN [21]	-	45.0	-	89.2	-	-
Out-of-domain Supervised Pretraining						
PSLA [22]	IN	44.0	31.9	-	-	96.3
AST [23]	IN	45.9	34.7	88.7	95.5	98.1
PaSST [24]	IN+AS	47.1	-	-	-	-
HTS-AT [12]	IN	47.1	-	-	-	98.0
In-domain Supervised Pretraining						
PANN [25]	AS	-	-	94.7	-	-
ERANN [21]	AS	-	-	96.1	-	-
PaSST [24]	IN+AS	-	-	96.8	-	-
HTS-AT [12]	IN	-	-	97.0	-	-
CLAP [15]	TA	-	-	96.7	-	-
Self-Supervised Pretraining						
SS-AST [7]	AS+LS	-	31.0	88.8	96.0	98.0
MSM-MAE [19]	AS	-	-	85.6	-	87.3
MaskSpec [6]	AS	47.1	32.3	89.6	-	97.7
MAE-AST [5]	AS+LS	-	30.6	90.0	95.8	97.9
AudioMAE [4]	AS	47.3	37.1	94.1	96.9	98.3
AudioMAE†	AS	47.1	36.8	93.8	96.7	98.1
Ours						
MAM-CLAP	AS+TA	**47.6**	**37.8**	**96.8**	**97.2**	**98.1**
SupMAM	AS	**47.7**	**37.6**	**96.2**	**97.3**	**98.2**
SupMAM-CLAP	AS+TA	**48.1**	**38.2**	**97.2**	**97.6**	**98.3**

3.1. Datasets

AudioSet (AS-2M, AS-20K) [11] contains about 2 million 10-second YouTube clips with 527 sound events. The full training set has 2 subsets: a class-wise balanced (22,176 clips) and an unbalanced (2,042,985 clips) set. The eval set has 20,383 clips. We collected and processed around 1.9M unbalanced training, 21K balanced training, and 19K evaluation clips due to the frequent change in YouTube videos available (e.g., videos being removed or taken down). For the AS-2M experiments, we use all unbalanced and balanced training audio clips for pretraining and finetuning. For the AS-20K experiments, we employ AS-2M for pretraining and the 20K balanced set for fine-tuning. We evaluate our models on the 19K eval set with the mean average precision (mAP) evaluation metric.

Environmental Sound Classification (ESC-50) [29] is an audio classification dataset including 2,000 5-second environmental sound recordings annotated with 50 classes. Each recording is only tagged with one class. We follow the 5-fold cross-validation evaluation setting as the previous work [4] and report the classification accuracy as the evaluation metric.

Speech Commands (SPC-2, SPC-1) [30] are two keyword spotting tasks. There are 35 speech commands in SPC-2. The training/validation/testing set contain 84,843/9,981/11,005 1-second recordings, respectively. In SPC-1, there are 10 classes of keywords, 1 silence class, and 1 unknown class that includes all the other 20 common speech commands. We report the testing accuracy based on the data and split provided by the SUPERB benchmark [31].

Table 2: *Influences of the mask ratio in our MAM-CLAP.*

Mask Ratio	AS-20K	ESC-50
10%	37.4	96.6
20%	**37.8**	**96.8**
30%	37.7	96.8
40%	37.6	96.5

3.2. Training Details

We employ a standard 12-layer ViT-B by default as the transformer encoder, where we keep the model size similar to AudioMAE for a fair comparison. For the decoder, we only use a single transformer block with shifted local attention for our MAM-CLAP and SupMAM-CLAP. In addition, we follow the training pipeline of AudioMAE [4] to train our models. We resample each raw waveform to 16 khz, and extract the 128-dimensional mel-filter bank features using a 25ms Hanning window that shifts every 10ms as the acoustic feature. The resulting spectrogram for a 10-second clip in AudioSet has a dimension of $1 \times 1024 \times 128$. Each acoustic feature is divided into 16×16 patches, which are then flattened into a patch sequence as the model input. We use AudioSet-2M for pretraining (including the supervised branch) and randomly iterate over all audio clips. We distribute the training load over 4 V100 GPUs with a batch size of 128 and a learning rate of 0.0002.

For our MAM-CLAP, we adopt the CLAP feature to supervise the masked and unmasked patches. For our SupMAM, we leverage the patch reconstruction branch and the supervised

Table 3: *Influences of the mask ratio in our SupMAM.*

Mask Ratio	AS-20K	ESC-50
20%	37.4	95.9
30%	37.5	96.1
40%	**37.6**	**96.2**
50%	37.5	96.0
60%	37.3	95.8

Table 4: *Influences of the pretraining objectives in our Sup-MAM.*

Pretraining Objectives	AS-20K	ESC-50
80% mask ratio		
rec	36.8	93.8
cls	33.1	91.3
rec+cls	37.2	95.4
40% mask ratio		
rec	36.6	93.3
cls	33.9	91.9
rec+cls	**37.6**	**96.2**

Table 5: *Influences of the classification loss ratio in our Sup-MAM.*

Classification Loss Ratio	AS-20K	ESC-50
0.05	37.2	95.7
0.02	37.5	96.0
0.01	**37.6**	**96.2**
0.005	37.3	95.8

classification branch simultaneously. For our SupMAM-CLAP, we use the CLAP feature to supervise the masked and unmasked patches while employing the supervised classification branch. λ_{cls} is set as 0.01 for SupMAM and 0.0001 for SupMAM-CLAP. Different from AudioMAE with a 80% mask ratio during pretraining, we use a masking ratio of 20% with unstructured random masking for our MAM-CLAP and SupMAM-CLAP, and adopt a 40% masking ratio for our SupMAM to achieve better performance. During pretraining, we do not apply any augmentation methods. During the finetune stage, we employ the SpecAugment [27] with 0.2 in time masking and 0.2 in frequency masking for our methods.

3.3. Experimental Results

Table 1 compares our methods to prior state-of-the-art results. With the CLAP target, our MAM-CLAP achieves the best performance across all tasks compared to other models with out-of domain supervised pretraining, and self-supervised pretraining methods, just slightly worse than HTS-AT with in-domain supervised pre-training on ESC-50. On AudioSet-2M, it improves the performance to 47.6 mAP. On AudioSet-20K and ESC-50, MAM-CLAP outperforms AudioMAE by a large margin. For the speech tasks (KS1, KS2), MAM-CLAP also achieves competitive performance compared with prior SOTA methods. What's more, our SupMAM also demonstrates superior performance on multiple downstream tasks. By combining our MAM-CLAP with SupMAM, we can achieve new SOTA results on all tasks compared to all previous SOTA models, which strongly demonstrates the effectiveness of our methods.

3.4. Ablation Study

In this part, we discuss the influence of the mask ratio in our MAM-CLAP, and the mask ratio, the pretraining targets, the classification loss ratio in our SupMAM. Here, we test the performance on the AS-20K and ESC-50 datasets.

Influences of the mask ratio in our MAM-CLAP. The mask ratio is a hyper-parameter that needs hand-design in MAM

methods. Given that we replace the reconstruction target with the CLAP feature and apply the supervision on both masked and visible patches in MAM-CLAP, we suppose that it may not be appropriate to adopt a high mask ratio as that in AudioMAE. Here, we show the results of different mask ratios, ranging from 10% to 40% with our MAM-CLAP in Table 2. It can be seen that unlike AudioMAE that achieves the best score with 80% mask ratio, our MAM-CLAP with 20% mask ratio achieves the best performance.

Influences of the mask ratio in our SupMAM. In this part, we extend AudioMAE to a supervised setting. As shown in Table 3, too large mask ratio will decrease the performance under supervised setting. The reason may be that when there are too many patches masked out, some short or inconspicuous sound events are easily erased up, resulting in incorrect classification and adversely affecting the model training. A 40% mask ratio can bring the best performance while retaining the advantages of AudioMAE and the supervised classification branch.

Influences of the pretraining objectives in our Sup-MAM. Table 4 studies the pretraining objectives with 80% and 40% mask ratios. The method degrades into AudioMAE with only the reconstruction (rec) objective. If only the classification (cls) objective is used, the method degrades into the supervised pretraining method with 80% or 40% input patches masked out. We discover that the self-supervised pretraining is compatible with the supervised objective and neither the reconstruction nor the classification objective can perform well when used in isolation. This is because only with both objectives can 100% patches be exploited: (1) reconstruction operates on the masked patches to learn local features, and (2) classification operates on the visible patches to provide the global feature learning ability.

Influences of the classification loss ratio in our Sup-MAM. As shown in Table 5, to achieve a balance between the two pretraining goals for SupMAM, we first set the ratio of the reconstruction loss to 1, and then tune the classification ratio. We find that a 0.01 classification loss ratio is optimal. A ratio that is too high would cause SupMAM to degrade into supervised pretraining (with a large proportion of input patches masked out), which is detrimental to the network pretraining.

4. Conclusions

In this paper, we propose to enhance the semantic modeling of MAM by distilling cross-modality knowledge from CLAP representations for both masked and unmasked regions (MAM-CLAP) and leveraging a multi-objective learning strategy with a supervised classification branch (SupMAM), thereby providing more semantic knowledge for MAM and enabling it to effectively learn global features from labels. Experiments show that our MAM-CLAP significantly improves the performance on multiple downstream tasks. Moreover, by combining our MAM-CLAP with SupMAM, we can achieve new state-of-the-art on various audio and speech classification benchmarks.

5. References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[2] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[3] X. Zhang, J. Chen, J. Yuan, Q. Chen, J. Wang, X. Wang, S. Han, X. Chen, J. Pi, K. Yao *et al.*, "Cae v2: Context autoencoder with clip target," *arXiv preprint arXiv:2211.09799*, 2022.

[4] P.-Y. Huang, H. Xu, J. B. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *Advances in Neural Information Processing Systems*.

[5] A. Baade, P. Peng, and D. Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," *arXiv preprint arXiv:2203.16691*, 2022.

[6] D. Chong, H. Wang, P. Zhou, and Q. Zeng, "Masked spectrogram prediction for self-supervised audio pre-training," *arXiv preprint arXiv:2204.12768*, 2022.

[7] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[9] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *International Conference on Learning Representations*.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[11] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[12] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022.

[16] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] Y. Xin, D. Yang, F. Cui, Y. Wang, and Y. Zou, "Improving weakly supervised sound event detection with causal intervention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," *arXiv preprint arXiv:2204.12260*, 2022.

[20] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[21] S. Verbitskiy, V. Berikov, and V. Vyshegorodtsev, "Eranns: Efficient residual audio neural networks for audio pattern recognition," *Pattern Recognition Letters*, vol. 161, pp. 38–44, 2022.

[22] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.

[23] Yuan Gong and Yu-An Chung and James Glass, "Ast: Audio spectrogram transformer," *Proc. Interspeech 2021*, pp. 571–575, 2021.

[24] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[25] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[26] Y. Xin, D. Yang, and Y. Zou, "Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification," *Proc. Interspeech 2022*, pp. 1546–1550, 2022.

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech*, 2019.

[28] F. Liang, Y. Li, and D. Marculescu, "Supmae: Supervised masked autoencoders are efficient vision learners," *arXiv preprint arXiv:2205.14540*, 2022.

[29] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[30] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[31] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.