



HAL
open science

Masked Latent Prediction and Classification for Self-Supervised Audio Representation Learning

Aurian Queennec, Pierre Chouteau, Geoffroy Peeters, Slim Essid

► **To cite this version:**

Aurian Queennec, Pierre Chouteau, Geoffroy Peeters, Slim Essid. Masked Latent Prediction and Classification for Self-Supervised Audio Representation Learning. 2024. hal-04921274

HAL Id: hal-04921274

<https://hal.science/hal-04921274v1>

Preprint submitted on 31 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Masked Latent Prediction and Classification for Self-Supervised Audio Representation Learning

Aurian Quelellenc, Pierre Chouteau, Geoffroy Peeters, Slim Essid
 LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
 {firstname}.{surname}@telecom-paris.fr

Abstract—Recently, self-supervised learning methods based on masked latent prediction have proven to encode input data into powerful representations. However, during training, the learned latent space can be further transformed to extract higher-level information that could be more suited for downstream *classification* tasks. Therefore, we propose a new method: **M**Asked **l**aten**T** **P**rediction **A**nd **C**lassification (MATPAC), which is trained with two pretext tasks solved jointly. As in previous work, the first pretext task is a masked latent prediction task, ensuring a robust input representation in the latent space. The second one is unsupervised classification, which utilises the latent representations of the first pretext task to match probability distributions between a teacher and a student. We validate the MATPAC method by comparing it to other state-of-the-art proposals and conducting ablations studies. MATPAC reaches state-of-the-art self-supervised learning results on reference audio classification datasets such as OpenMIC, GTZAN, ESC-50 and US8K and outperforms comparable supervised methods’ results for musical auto-tagging on Magna-tag-a-tune.

Index Terms—self-supervised, audio representation learning, audio spectrogram transformers

I. INTRODUCTION

Recently, Self Supervised Learning (SSL) methods have become a predominant representation learning approach across various application domains, especially speech and audio processing [1]. To alleviate the absence of labels in SSL, one needs to elaborate pretext tasks where the model learns valuable representations from the input data itself. A wide variety of pretext tasks have been developed [2]. In general, models of this kind may operate either from a single view of the input [3], [4] or multiple views obtained through hand-crafted data augmentations [5], [6]. While the latter variant may be sensitive to the bias that results from the choice of views [7], single-view methods learn representations only from the original data, which requires less prior knowledge. One may also categorize pretext tasks with respect to the problem they solve, *i.e.*, classification [8], reconstruction [3] or prediction [4].

Pretext tasks that consider unsupervised classification as an objective have proven quite successful. Models like HuBERT [9] or BEATs [10] start by assigning a cluster identity to each input, using either K-Means or a learned tokenizer, before learning a representation that can successfully predict the input-cluster association. In computer vision, DINO [8] tries

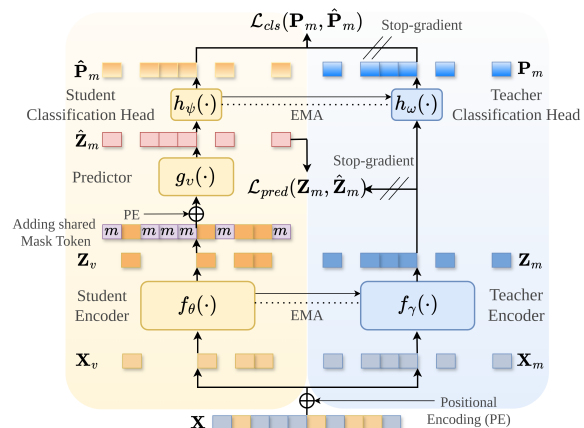


Fig. 1. Overview of proposed method, MATPAC, combining a prediction and classification pretext task.

to assign the same classes to multiple different views of the original input, using a teacher-student architecture.

More recently, promising results have been obtained using pretext tasks where the goal is to predict the latent representation of a masked part of the input from the visible part. This leverages the idea of Masked Language Modeling (MLM) [11]. When combined with a teacher-student paradigm, such as in Bootstrap Your Own Latent (BYOL) [5], this leads to M2D [12] and I-JEPA [4].

Contributions This work stems from the intuition that in order to better target *classification* downstream problems, an effective pretext task could be realised by modifying the goal of predicting the latent representation of a masked part of the input given a visible part (as done in I-JEPA and M2D), to instead predict the *cluster identity* of the masked part.

Thus, we propose a new SSL method where we consider two pretext tasks which are solved jointly:

- i) a *masked prediction task*, solved in the latent space, as in previous work [4], [12], [13];
- ii) an *unsupervised classification task* where, as illustrated in Fig.1, the predicted and target representations are separately projected by a student and teacher classification head into probability distributions to be matched through a classification loss.

We call this new method: **M**Asked **l**aten**T** **P**rediction **A**nd **C**lassification (MATPAC), which encodes the input into a robust predictive representation thanks to the first pretext task, while the classification pretext task makes it easier to extract

This work was supported by the Audible project, funded by French BPI, and was performed with GENCI-IDRIS resources (Grant 2024-AD011013929R1)

high-level information (*i.e.* cluster identity) further up from the latent representation, thus preparing the latent space for easy exploitation by classification downstream tasks.

We show through extensive experimentation that MATPAC sets new state-of-the-art (SOTA) performance for SSL on musical instrument recognition, musical genre recognition and event classification tasks.

II. RELATED WORKS

SSL methods based on masked prediction such as M2D [12] or I-JEPA [4] rely on pretext tasks that exploit a single view of the original input, using a teacher-student architecture to solve a prediction problem in the latent space. The student encodes only the visible patches, while the teacher encodes the masked patches as a target latent representation. Then, from the latent representation of the visible patches, a shallow predictor tries to match the target latent representations. While M2D was designed for audio signals and I-JEPA for images, they consider different masking strategies. Riou et al. [14] studied which of those masking strategies is the most suited for audio and found out that random masking is the most effective.

Besides, BEATs [10] for general audio, HuBERT [9] for speech, or MERT [15] for music, are models relying on unsupervised classification as a pretext task, but they either need iterative procedures or other models for initialization. Alternatively, other works in the field of computer vision use classification as an objective in their pretext task [8], [16], [17], considering multiple views of the input in a teacher-student architecture, thus avoiding the need for third-party models to provide the initial classification targets. DINO [8] passes different augmentations of an input image to the student and teacher networks, projects them into a probability distribution, and learns to match them through classification. In [17], the authors combine a DINO-like loss with the MLM iBOT [16] loss to obtain a teacher-student system relying only on a classification objective.

For learning general audio representations, SSAST [18] and MAE-AST [19] combine two pretext tasks: a masked input reconstruction and a classification objective. But as shown in [12], using a masked prediction pretext task in the latent space is much more effective than reconstructing the masked part of the input. To our knowledge, no general self-supervised representation models combine a teacher-student architecture with both classification and prediction pretext tasks that consider the unsupervised classification of the latent representations resulting from a teacher-student prediction with masking.

III. METHOD

Fig. 1 depicts our method’s entire pipeline. For the masked prediction pretext task, our approach is mainly inspired by M2D [12] and I-JEPA [4] as we use a teacher-student architecture with a form of MLM in the latent space. In addition, we use the predicted and target latent representation to solve an unsupervised classification pretext task.

Input of the model As input, we extract flattened patches, \mathbf{X} , from the log-scale Mel spectrogram of each audio sample. A two-dimensional learned positional encoding, \mathbf{p} is added to \mathbf{X} before randomly partitioning the sequence into \mathbf{X}_v , the visible patches, and \mathbf{X}_m , the masked patches. As opposed to I-JEPA [4], and like M2D [12], we chose a random masking strategy.

Masked prediction pretext task The student encoder f_θ , of parameters θ , projects \mathbf{X}_v such that $\mathbf{Z}_v = f_\theta(\mathbf{X}_v)$ is the latent representation of the visible patches. Similarly, the masked patches are projected with the teacher encoder f_γ , and $\mathbf{Z}_m = f_\gamma(\mathbf{X}_m)$. The teacher’s encoder parameters γ are updated using an Exponential Moving Average (EMA) of f_θ . The update rule is $\gamma \leftarrow \lambda\gamma + (1 - \lambda)\theta$, where λ is the decay rate.

The input of the predictor, g_v , is all the encoded visible patches to which we append the same shared learnable token \mathbf{m} at each masked position. A new learned positional embedding is added to give location information to the masked token. The predictor outputs a representation for visible and masked patch positions. However, we select only the prediction of the latent representation for masked patches to compare them with the targets. The prediction for the masked latent representation is $\hat{\mathbf{Z}}_m = \{g_v(\mathbf{m_concat}(\mathbf{Z}_v, \mathbf{m}, \mathcal{S}_m))^{(i)}, i \in \mathcal{S}_m\}$, where \mathcal{S}_m is the set of the masked indices. The prediction loss is then the square error of l_2 -normalized versions of \mathbf{Z}_m and $\hat{\mathbf{Z}}_m$, denoted by \mathbf{Z}'_m and $\hat{\mathbf{Z}}'_m$:

$$\mathcal{L}_{pred}(\hat{\mathbf{Z}}_m, \mathbf{Z}_m) = \sum_i (\hat{\mathbf{z}}_m^{(i)} - \mathbf{z}'_m^{(i)})^2; \quad (1)$$

where $\mathbf{Z}'_m = \{\mathbf{z}'_m^{(i)}, i \in [1, N]\}$, and N is the number of targets, *i.e.* the number of masked patches.

Classification pretext task Given a student h_ψ and teacher h_ω projection heads, we transform the target and predicted latent representations into probabilities distribution of dimension K . \mathbf{P}_m and $\hat{\mathbf{P}}_m$, respectively the target and predicted probability distributions, are obtained from \mathbf{Z}_m and $\hat{\mathbf{Z}}_m$ as follow:

$$\hat{\mathbf{P}}_m = \text{Softmax}((h_\psi(\hat{\mathbf{Z}}_m)/\tau_s); \quad (2)$$

$$\mathbf{P}_m = \text{Softmax}((h_\omega(\mathbf{Z}_m) - \mathbf{C})/\tau_t); \quad (3)$$

where τ_s and τ_t are temperature parameters used to sharpen the distribution, and \mathbf{C} is used to center the distribution. \mathbf{C} is an EMA of the mean of $h_\omega(\mathbf{Z}_m)$. Sharpening and centering were introduced by DINO’s authors [8] to avoid the collapse of their method. Our preliminary experiments showed that, even when the masked prediction pretext task is, we need the sharpening and centering operations to avoid the collapse of h_ω and h_θ into a trivial solution where one dimension dominates.

Finally, the classification pretext task matches the distributions with a cross-entropy loss depending on ψ of h_ψ :

$$\mathcal{L}_{cls}(\hat{\mathbf{P}}_m, \mathbf{P}_m) = \sum_i -\mathbf{p}_m^{(i)} \log(\hat{\mathbf{p}}_m^{(i)}). \quad (4)$$

The parameters ω of h_ω are updated with an EMA of h_ψ with a decay rate ζ different from λ (the decay rate of f_θ EMA).

Total loss The final training objective is simply the sum of \mathcal{L}_{cls} and \mathcal{L}_{reg} with a weight α :

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{cls} + \alpha\mathcal{L}_{pred}. \quad (5)$$

IV. EXPERIMENTS

In this section, we evaluate the MATPAC method, showing 1) the effectiveness of combining masked prediction of the latent representation with unsupervised classification, 2) the balancing between the two tasks, and 3) the impact of the design of the classification head.

A. Experimental Setup

Pre-training dataset and input processing We use AudioSet [20] as a pre-training dataset. Our version of AudioSet has 2,012,215 samples over 6s length. Each audio sample is processed to a log-scale Mel spectrogram with a sampling rate of 16,000Hz using a window size of 25 ms, a hop size of 10 ms, and 80 Mel bins spaced between 50 and 8,000 Hz. The log-scale Mel spectrograms are standardized with the dataset statistics. To process the log-scale Mel spectrogram into the input of the encoders, we used a patch size of 16×16 and a masking ratio of 0.7, as it has shown to work well across different tasks [12], [14].

Encoders and predictor details We based our code on M2D’s PyTorch implementation of the teacher-student encoder and predictor parts, making the minimum number of changes while keeping the same parameters.¹ To keep a fair comparison with SOTA methods like M2D and ATST-Clip [21], and as it seems a good design choice [14], we train our method with audio segments of 6 s. We randomly crop 6-s audio segments from the audio samples above this duration in the pre-training dataset. Similarly to M2D, we train for 300 epochs, with a batch size of 2048, a warm-up of 20 epochs, the same base learning rate and optimizer, and the EMA decay rate λ follows the same update policy.

Classification head details We use the same architecture and parameters as in DINO for the teacher-student classification heads.² It is divided into two parts. First, there are three fully connected layers with a bottleneck output of dimension 256. Then, a weight-normalized fully connected layer projects the l_2 -normalized output of the first layers into the probability distribution space of K dimensions [8]. By default, we set the hidden dimension to 2048 and $K = 2048$. We linearly increase the temperature parameter τ_t in Eq. (2), from 0.04 to 0.07 for $n_{\tau, epoch}$ epochs. This encourages the target’s probability distribution to be peakier at the beginning of the training, forcing some “classes” to appear. We keep $\tau_s = 0.1$ for the whole training. Finally, the EMA decay rate ζ of the classification heads update is linearly interpolated from 0.998 to 1 at the end of the training and $\alpha = 0.5$.

Linear probing evaluation We train a linear classifier consisting of a single fully connected layer that maps the latent representation of the model we evaluate to the number

TABLE I
EVALUATION DOWNSTREAM TASKS. TVT STANDS FOR TRAIN/VALIDATION/TEST

Dataset	#Samples	#Classes	Type	Split	Metric
OpenMIC	20,000	20	ML	TVT	mAP
NSynth	305,979	11	MCSL	TVT	Acc.
GTZAN	930	10	MCSL	TVT	Acc.
MTT	25,863	50	ML	TVT	mAP
FSD50K	51,197	200	ML	TVT	mAP
ESC-50	2,000	50	MCSL	5-Fold	Acc.
US8K	8,732	10	MCSL	10-Fold	Acc.

of classes of the task. We use a batch size of 128 for all tasks and a learning rate of $1e^{-4}$ with an ADAM optimizer. We kept the model with the best validation loss, and each downstream experiment was repeated five times. We report the average scores with their Confidence Interval (CI) at 95%.

For MATPAC, if the input audio from the downstream task exceeds 6s, we split the audio every 6s without overlapping and perform average pooling over time to have one embedding per clip. Since our method encodes each time-frequency patch (extracted from the input log-scaled Mel spectrogram) with a specific latent representation, we concatenate the patches over frequencies and average them over time, as done in M2D.

Downstream Datasets To evaluate our method, we selected a variety of classification downstream tasks that cover music and environmental sound domains. For instrument classification tasks, we have chosen the NSynth dataset [24], a synthetic dataset of musical notes. We also consider OpenMIC [25] for instrument classification as its samples are extracted from real music recordings. For genre classification tasks, we use GTZAN [26] with the corrected labels following Sturm’s work [27]. For music autot-agging, we use the Magna-Tag-A-Tune (MTT) dataset. While GTZAN and NSynth are widely used in evaluation frameworks [6], [28], [29], OpenMIC and MTT are less common. For environmental sound analysis tasks, we consider FSD50K [30], a sound event event recognition dataset whose classes are from the AudioSet ontology. In addition, we consider the widely-used ESC-50 [31], and UrbanSound8K (US8K) [32] which is an urban sound classification dataset. Table I provides more details about each of the datasets, for instance, the type of classification, which is either Multi-Label (ML) or Multi-Class Single-Label (MCSL).

B. Comparison with SOTA methods

We compare MATPAC with other SOTA methods on linear evaluation. Results shown in Table II prove that our method is better than previous SSL methods on all tasks except NSYNTH and FSD50K, where respectively M2D and ATST models get higher scores. Performance discrepancy between M2D and MATPAC without the classification task are mostly due to the differences in the list of AudioSet excerpts used in our experiment (owing to no longer active YouTube videos). Moreover, while on most tasks, the supervised learning methods PaSST [22], HTS-AT [23] and BEATs iter3+ still outperform all self-supervised methods, MATPAC achieves score of 41.1% on MTT, outperforming both supervised and self-supervised learning methods considered here³. MATPAC

¹<https://github.com/ntccslab/m2d>

²<https://github.com/facebookresearch/dino>

³Note that SOTA scores on MTT are achieved by [33]

TABLE II

LINEAR EVALUATION COMPARISON WITH CI AT 95%. † SCORES AND CI ARE FROM THE ORIGINAL PAPER. SUPERVISED LEARNING (SL) METHODS ARE GREYED OUT. BOLD SCORES ARE THE BEST SSL RESULTS. UNDERLINED ONES ARE THE BEST SCORES OVERALL. POSSIBLE TRAINING DATA ARE AUDIOSET (AS) AND LIBRISPEECH (LS).

Models	Training data	OpenMIC mAP	NSynth Acc(%)	GTZAN Acc(%)	MTT mAP	FSD50K mAP	ESC-50 Acc(%)	US8K Acc(%)	Avg.
MATPAC - $n_{\tau,epoch} = 10$	AS	85.3±0.0	74.3±0.2	85.3±0.4	41.1 ±0.0	55.2±0.1	93.5 ±0.1	89.4 ±0.1	74.9
MATPAC - $n_{\tau,epoch} = 20$	AS	85.4 ±0.0	74.6±0.5	85.9 ±0.3	41.0±0.0	54.8±0.1	93.2±0.1	88.0±0.0	74.7
MATPAC w/o Classification	AS	85.1±0.0	73.7±1.1	84.2±0.4	40.8±0.0	54.4±0.2	92.5±0.1	87.5±0.1	74.0
M2D - ratio = 0.7 [12]	AS	84.8±0.1	76.9 †±1.3	83.9†±1.4	40.6±0.1	52.8±0.8	89.8†±0.3	87.1†±0.3	73.7
ATST-Clip [21]	AS	84.2±0.0	76.2†	79.9±0.2	39.1±0.0	58.5 †	91.4±0.1	85.8†	73.6
ATST-Frame [21]	AS	83.1±0.1	75.9†	80.7±0.7	39.1±0.1	55.1†	87.5±0.2	85.8†	72.5
BEATs iter3 [10]	AS	82.7±0.1	72.4±0.7	80.0±0.3	38.3±0.0	46.7±0.1	87.5±0.1	85.4±0.1	70.4
MAE-AST [19]	AS+LS	79.4±0.1	71.2±0.4	64.1±1.2	36.9±0.2	41.1±0.1	84.4±0.2	81.3±0.2	65.5
PaSST [22]	AS	86.5±0.1	72.9±0.5	87.4±2.0	40.4±0.1	<u>61.3</u> ±0.2	<u>97.0</u> ±0.1	89.1±0.1	76.4
HTS-AT [23]	AS	86.4±0.0	68.6±0.3	85.9±0.4	40.1±0.0	59.4±0.0	95.9±0.1	85.3±0.0	74.5
BEATs iter3+ [10]	AS	<u>86.7</u> ±0.0	71.6±0.5	86.0±0.4	40.3±0.0	60.6±0.1	96.1±0.1	<u>89.5</u> ±0.1	75.8

also has comparable performances with supervised methods for instrument classification on GTZAN and urban sound classification on US8K.

The code and pre-trained weights are publicly available.⁴

C. Ablation experiments

TABLE III
IMPACT OF α , THE WEIGHTING BETWEEN THE CLASSIFICATION AND THE MASKED PREDICTION PRETEXT TASK

α	0.25	0.5	0.75	1
Music	71.5	71.5	71.4	71.0
Environment	78.5	79.3	78.5	78.1
All	74.5	74.9	74.4	74.0

Importance of solving jointly the masked prediction and classification pretext tasks Firstly, in Table II, we observe that, for both $n_{\tau,epoch}$ values, MATPAC has better scores to those without the classification pretext task. The two configurations achieved an average score of 74.9 and 74.7, significantly above the average score of 74.0 obtained when training without the classification pretext task. For $n_{\tau,epoch} = 20$ epochs, MATPAC achieves the best overall score on OpenMIC, NSYNTH and GTZAN, while it has better scores on environmental tasks for $n_{\tau,epoch} = 10$. Without the masked prediction pretext task, however, the model collapses as the classification head is not able to ensure that \mathbf{Z}_m is close to $\hat{\mathbf{Z}}_m$. Therefore \mathbf{P}_m and $\hat{\mathbf{P}}_m$ do not match either. Those results emphasise the utility of considering both pretext tasks: masked prediction and unsupervised classification.

Influence of α in the loss In Table III, we show that $\alpha = 0.5$ yields the best average score of 74.9. For lower values, the classification pretext task has more importance in the loss and scores start to drop. On the other hand, when $\alpha = 1$, only the masked prediction pretext task is taken into account in the loss Eq. (5), and the average score of 74.0 shows that it is less effective than using both pretext tasks.

Classification head architecture Concerning the classification head, we explore which output dimensionality K is

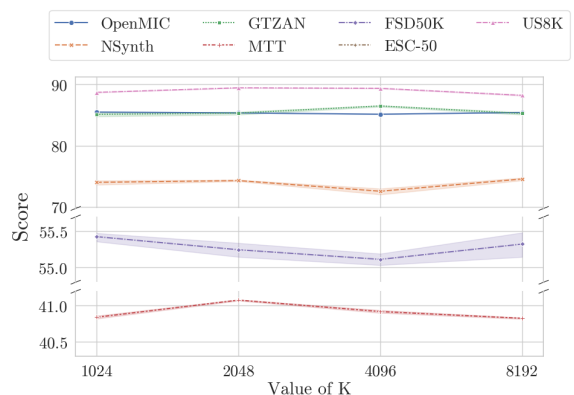


Fig. 2. Influence of the dimensionality K in the classification head

best. In Fig. 2, we report the scores on each downstream task for $K \in [1024, 2048, 4096, 8192]$. We observed that scores vary very little as a function of K , but overall, for $K = 2048$ we obtain the best results. We can conclude that the classification pretext task only plays the role of regularizer of the latent space, making the representation effective in downstream tasks.

V. CONCLUSIONS

We proposed a new Self Supervised Learning method, **MA**sKed **l**atent **P**rediction **A**nd **C**lassification (MATPAC), which combines two pretext tasks to learn a robust representation for classification downstream tasks. The first pretext task performs masked prediction in the latent space, which had already proved effective in previous work. The second is an unsupervised classification task that projects the first task’s predicted and target latent representation into probability distributions matched through a classification loss. Through our evaluation of MATPAC on various downstream tasks and our ablation studies, we prove the effectiveness of solving two pretext tasks jointly. Notably, MATPAC achieves state-of-the-art results compared to other SSL methods on OpenMIC, GTZAN, ESC-50 and US8K while outperforming comparable fully supervised competitors on NSynth and Magna-tag-a-tune.

⁴<https://github.com/aurianworld/matpac>

REFERENCES

- [1] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *Patterns*, vol. 3, no. 12, p. 100616, 2022.
- [2] R. Balestrieri, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, "A cookbook of self-supervised learning," *arXiv:2304.12210*, 2023.
- [3] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *NeurIPS*, New Orleans, December 2022.
- [4] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. G. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. CVPR, Vancouver, BC, Canada, June 17-24, 2023*, Vancouver, June 2023, pp. 15 619–15 629.
- [5] J. Grill, F. Strub, F. Althé, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - A new approach to self-supervised learning," in *Proc. NeurIPS*, Virtual, December 2020.
- [6] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for audio: Exploring pre-trained general-purpose audio representations," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 137–151, 2023.
- [7] M. Assran, R. Balestrieri, Q. Duval, F. Bordes, I. Misra, P. Bojanowski, P. Vincent, M. G. Rabbat, and N. Ballas, "The hidden uniform cluster prior in self-supervised learning," in *Proc. ICLR*, Rwanda, May 2023.
- [8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, Montreal, October 2021, pp. 9630–9640.
- [9] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
- [10] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, vol. 202, Honolulu, July 2023, pp. 5178–5193.
- [11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, June 2019, pp. 4171–4186.
- [12] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *Proc. ICASSP*, June 2023, pp. 1–5.
- [13] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. ICML*, vol. 162, Baltimore, July 2022, pp. 1298–1312.
- [14] A. Riou, S. Lattner, G. Hadjeres, and G. Peeters, "Investigating design choices in joint-embedding predictive architectures for general audio representation learning," in *Proc. ICASSP*, Seoul, April 2024, pp. 680–684.
- [15] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. B. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu, "MERT: acoustic music understanding model with large-scale self-supervised training," in *Proc. ICLR*, Vienna, May 2024.
- [16] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. L. Yuille, and T. Kong, "Ibot: Image BERT pre-training with online tokenizer," in *Proc. ICLR*, Virtual, April 2022.
- [17] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *Trans. Machine Learning Research*, January 2024.
- [18] Y. Gong, C. Lai, Y. Chung, and J. R. Glass, "SSAST: self-supervised audio spectrogram transformer," in *Proc. AAI*, Virtual, March 2022, pp. 10 699–10 709.
- [19] A. Baade, P. Peng, and D. Harwath, "MAE-AST: masked autoencoding audio spectrogram transformer," in *Proc. Interspeech*, Incheon, September 2022, pp. 2438–2442.
- [20] J. F. Gemmeke, D. P. W. 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 *ICASSP*, New Orleans, March 2017, pp. 776–780.
- [21] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1336–1351, 2024.
- [22] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, Incheon, September 2022, pp. 2753–2757.
- [23] 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 *Proc. ICASSP*, Singapore, May 2022, pp. 646–650.
- [24] J. H. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70, Sydney, August 2017, pp. 1068–1077.
- [25] E. Humphrey, S. Durand, and B. McFee, "Openmic-2018: An open data-set for multiple instrument recognition," in *Proc. ISMIR*, Paris, September 2018, pp. 438–444.
- [26] G. Tzanetakis and P. R. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.
- [27] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," vol. abs/1306.1461, 2013. [Online]. Available: <http://arxiv.org/abs/1306.1461>
- [28] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*, December 2022, pp. 125–145.
- [29] L. Wang, P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira, and A. van den Oord, "Towards learning universal audio representations," in *Proc. ICASSP*, Singapore, April 2022, pp. 4593–4597.
- [30] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829–852, 2022.
- [31] K. J. Piczak, "ESC: dataset for environmental sound classification," in *Proc. MM*, Brisbane, October 2015, pp. 1015–1018.
- [32] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. MM*, November 2014, pp. 1041–1044.
- [33] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proc. ISMIR*, Virtual, November 2021, pp. 88–96.