ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095041

# AUDIO BARLOW TWINS: SELF-SUPERVISED AUDIO REPRESENTATION LEARNING

*Jonah Anton*[⋆] *Harry Coppock*[⋆] *Pancham Shukla*[⋆] *Björn W.Schuller*[⋆†]

[⋆]Imperial College London, UK
[†]EIHW, University of Augsburg, Germany

## ABSTRACT

The Barlow Twins self-supervised learning objective requires neither negative samples or asymmetric learning updates, achieving results on a par with the current state-of-the-art within Computer Vision. As such, we present ***Audio Barlow Twins***, a novel self-supervised audio representation learning approach, adapting Barlow Twins to the audio domain. We pre-train on the large-scale audio dataset AudioSet, and evaluate the quality of the learnt representations on 18 tasks from the HEAR 2021 Challenge, achieving results which outperform, or otherwise are on a par with, the current state-of-the-art for instance discrimination self-supervised learning approaches to audio representation learning. Code at `https://github.com/jonahanton/SSL_audio`.

## 1. INTRODUCTION

Inspired by recent successes in Computer Vision (CV) [1, 2, 3] and Natural Language Processing (NLP) [4, 5] in the generation of universal representations[1] through self-supervised learning (SSL) methodologies, much recent interest has been dedicated to using SSL to learn universal representations of audio data [6, 7, 8, 9]. Whilst generative approaches [8, 9] have produced state-of-the-art (SOTA) results for SSL methods in many audio tasks, the current SOTA SSL techniques in CV are dominated by instance discrimination (ID) approaches [10, 11, 12], which build a meaningful representation space through training an encoder network to embed similar instances near one another.

Barlow Twins [13] is one such ID approach, which encourages the empirical cross-correlation matrix between the embeddings of two views of a mini-batch of data samples towards the identity matrix. Through forcing the cross-correlation matrix to the identity, Barlow Twins embeds instances which encode similar semantic content near one another whilst minimising the redundancy between the individual components of the extracted embedding vectors, encouraging the latent representations to be maximally informative. Barlow Twins requires neither negative samples [1] nor asymmetric learning updates [2, 14], instead preventing representational collapse by design. As a result, Barlow Twins i) directly enforces invariances to the applied data augmentations without having to sample negative pairs, and ii) prevents representational collapse in an intuitive and explainable manner [15], unlike approaches such as BYOL [2] which are theoretically poorly understood (although some attempts have recently been made [16]). Within the audio domain, the sampling of negative pairs is also potentially problematic, since obtaining such a pair from two different audio signals within a mini-batch [6, 17] can lead to low-quality solutions since two signals may share common sounds, such as a chord sequence in music.

It seems reasonable, therefore, that Barlow Twins, when adapted to the audio domain, would produce robust and generalisable audio representations. To this end, we present ***Audio Barlow Twins*** (ABT), a novel self-supervised audio representation learning method which adapts Barlow Twins [13] to the audio domain. Figure 1 details ABT's high level architecture. ABT achieves results which outperform, or otherwise are on a par with, the current state-of-the-art for ID self-supervised learning approaches to audio representation learning.

## 2. BACKGROUND

**Instance Discrimination** Instance discrimination (ID) SSL approaches [1, 2, 18] are built on the core idea of similarity: instances which encode similar semantic content should be embedded near one another in representation space. These methods make use of a Siamese network, where each 'arm' of the network processes a different view of the data sample. The extracted feature representations of the two views are then pushed together. Solely enforcing representational similarity of positive pairs is vulnerable to mode collapse onto a constant vector for all inputs, a phenomenon known as representational collapse. Contrastive ID approaches, such as SimCLR [1], prevent representational collapse through the use of negative pairs, which are forced apart in representation space. Non-contrastive ID approaches such as BYOL [2] prevent representation collapse, instead, through introduction of asymmetry into the learning framework.

**Audio SSL (ID)** Many self-supervised learning methods have been proposed to learn generalisable audio repre-

[1]A representation is a lower-dimensional and compressed, but highly informative, distillation of an input.

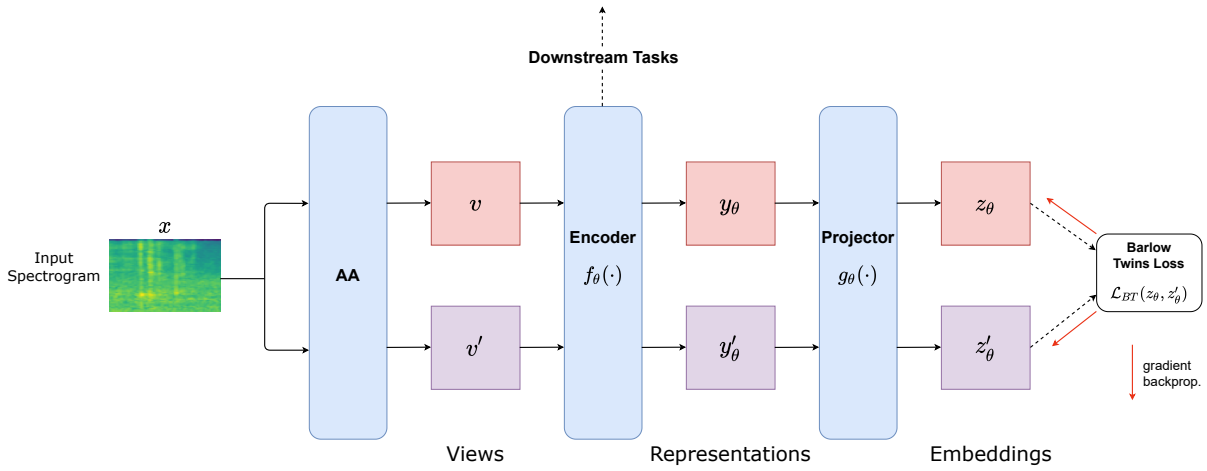


**Fig. 1**: The Audio Barlow Twins learning framework.

sentations[2]. [17, 6, 20] all adapt SimCLR [1] to the audio domain. [17] additionally propose an augmentation which they term *mix-back*, where the incoming spectrogram is mixed with another clip randomly drawn from the training dataset whilst ensuring that the incoming patch remains dominant. [7] present BYOL-A, adapting BYOL [2] to the audio domain with minimal modifications from the original learning framework. The key modification they make is their proposed data augmentation module, used to generate the two spectrogram views. BYOL-A also makes use of a lightweight convolutional encoder architecture, based on a network used in the solution of the NTT DCASE2020 Challenge Task 6 (Automated Audio Captioning) [21], which we use in our experiments and term the *AudioNTT* encoder.

## 3. METHOD

**Generation of views** ABT first produces two views, $v, v'$ of an input spectrogram $x$ by stochastic application of the audio augmentation (AA) module $v, v' \sim \text{AA}(x)$. The audio augmentation module consists of three different augmentation blocks: Mixup, Random Resize Crop (RRC), and Random Linear Fader (RLF) [22]. The spectrogram input is first normalised by the dataset mean and standard deviation.

**Extraction of embeddings** The two views are passed through the encoder to obtain the representations, $y_\theta = f_\theta(v), y'_\theta = f_\theta(v')$. The representations are then passed through the projector network to obtain the embeddings, $z_\theta = g_\theta(y_\theta), z'_\theta = g_\theta(y'_\theta)$.

**Barlow Twins objective** The Barlow Twins objective, $\mathcal{L}_{BT}$, is calculated on the embeddings, $\mathcal{L}_{BT}(z_\theta, z'_\theta)$. $\mathcal{L}_{BT}$, since it uses batch statistics in its calculation of the embeddings' cross-correlation matrix $C$, cannot in practice be calculated on an input-by-input basis, but instead must be calculated over a batch of embeddings $Z_\theta, Z'_\theta$, with $Z_\theta = [z^1_\theta, ..., z^B_\theta] \in \mathbb{R}^{B \times d}$, and likewise for $Z'_\theta$. Formally

$$\mathcal{L}_{BT} = \alpha \sum_i (1 - C_{ii})^2 + \lambda \sum_{i \neq j} C_{ij}^2, \tag{1}$$

where the first term enforces representational invariance to the applied audio augmentations, and the second term minimises the redundancy between the individual components of the embedding vectors. The positive constants $\alpha$ and $\lambda$ control the trade-off between the importance of these two terms, and by default $\alpha$ is set to 1 and $\lambda$ to 0.005 (as in the original Barlow Twins publication [13]). The cross-correlation matrix $C$ is computed between the embeddings within the batch $B$,

$$C_{ij} = \sum_{b=1}^{B} \hat{Z}^b_{\theta,i} \hat{Z'}^b_{\theta,j}, \tag{2}$$

where $\hat{Z}_\theta$ is the normalised embedding $Z_\theta$ along the batch dimension, and $\hat{Z}^b_{\theta,i}$ corresponds to the $i^{th}$ component of the $b^{th}$ batch element of $\hat{Z}_\theta$.

## 4. EXPERIMENTS

We pre-train on the large-scale audio dataset AudioSet [23] for 100 epochs with a batch size of 128, which corresponds to $\sim$ 1.3M training iterations. We successfully download $1,629,756$ clips (corresponding to $\sim 4,500$ hours of audio) from AudioSet's unbalanced train subset, which are used for ABT pre-training.

**Audio preprocessing** All audio samples are converted with a sampling frequency of 16 kHz to (log-scaled) mel-spectrograms using a 64 ms sliding window with a 10 ms step size, extracting $F = 64$ mel frequency bins in the range $60 - 7,800$ Hz. By default, during pre-training, we randomly crop $T = 96$ time frames (all clips with shorter duration are

[2]A full and in-depth analysis on the current SOTA audio self-supervised learning methods can be found in the survey produced by [19].

padded with zeros), corresponding to 950 ms of audio. This produces a mel-spectrogram of size $F \times T = 64 \times 96$.

**Architecture** We consider two encoders, the AudioNTT convolutional encoder [21], and the ViT$_C$ encoder [24]. We consider the ViT$_C$-*B*(ase)[3] model, using a patch size of $16 \times 8$. A learnable `[CLS]` token is prepended to the sequence of patches, and its output representation, $O_{[\texttt{CLS}]}$, is taken as representative of the clip as a whole. Fixed sinusoidal positional encodings are added to each patch.

**Downstream Tasks** We use 18 tasks from the HEAR 2021 Challenge [25] for evaluation. HEAR includes two types of tasks, i) scene-based tasks, corresponding to classification of an entire audio clip, and ii) timestamp-based tasks, corresponding to sound event detection and or transcription over time. For each task, the representations are extracted from the frozen pre-trained model and then evaluated using the `hear-eval`[4] toolkit, which trains a shallow Multilayer Perceptron (MLP) classifier on the extracted representations.

## 5. RESULTS

We compare the performance of the ABT pre-trained models on the 18 HEAR tasks with two baseline models, CREPE [27], wav2vec2.0 [28], and to BYOL-A*[5] [7]. The results for the scene-based and timestamp-based tasks are detailed in Tables 1,2 and 3.

ABT, with the AudioNTT encoder, generally performs on a par with, or outperforms, BYOL-A*, which uses the same AudioNTT encoder architecture, on the scene-based tasks, and consistently outperforms BYOL-A* on the timestamp-based tasks. We see further consistent improvements over CREPE and wav2vec2.0, except on the type of tasks on which these models have been specialised (music for CREPE, speech for wav2vec2.0). These results demonstrate the robustness of ABT in the generation of general-purpose audio representations. Interestingly, ABT pre-training appears damaging to performance on several of the music tasks, often leading to performance degradation from the random baselines. We find this to be particularly evident for the Mridingam Stroke and Tonic, NSynth (5h and 50h), and MAESTRO tasks. This extends to other ID methods, with BYOL-A*, performing similarly poorly. The aforementioned tasks all require a sound's pitch to be correctly discerned. However, invariance to pitch perturbations is enforced through RRC, and as such, it is intuitive that a model will consequently struggle to classify pitch. That said, we find that RRC does considerably improve the quality of the learnt representations, as detailed in Section 5.1. We therefore observe an issue with transferring ID methods from CV to audio, since such methods rely on applying data augmentations to generate two views, and any given data augmentation may benefit one type of audio task but harm another. This provides support for generative self-supervised methods for learning universal audio representations [8, 9], since they don't require the use of any data augmentations.

### 5.1. Ablation study: the effect of strong data augmentations

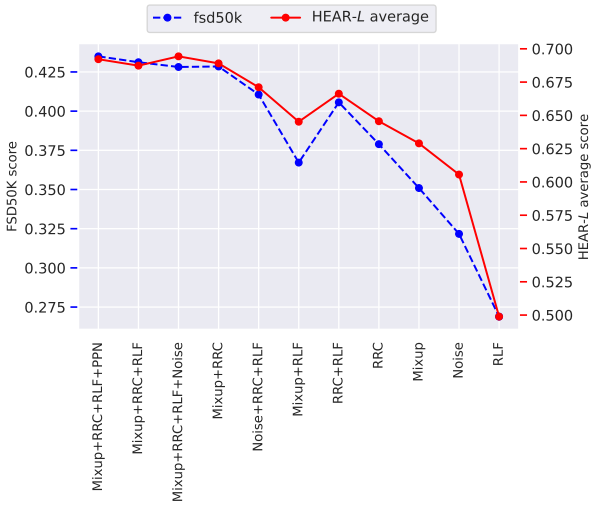


**Fig. 2**: We compare the effect of pre-training with different combinations of the components of the Audio Barlow Twins audio augmentation (AA) module. Results are shown both evaluated on FSD50K (blue) and the average score on the 5 HEAR-*L* tasks (red).

We consider using different combinations of the components of the audio augmentation module, which by default consists of Mixup, Random Resize Crop (RRC), and Random Linear Fader (RLF). We further consider two different variations, Pre-Post-Norm (PPN) and *Noise*. PPN refers to removal of the normalisation block and replacing it with the Pre- and Post-Normalisation blocks proposed by [7] in BYOL-A. Noise refers to interpolation with random noise[6]. ABT pre-training is performed for each ablation study for 100 epochs with the AudioNTT encoder on the FSD50K development subset, and linear evaluation is performed on five HEAR tasks, which we term HEAR-*L*[7]. From Figure 2 we see that strong audio augmentations are essential for the learning of high-quality representations, since when all the augmentations are removed from the baseline except RLF (remove RRC and Mixup), ABT performance drops significantly, by 19 points from 69% to 50% average on the HEAR-*L* tasks. We also obvserve that

[3]The ViT$_C$-*B* corresponds to the ViT$_C$-18GF model proposed in the original publication [24].

[4]`https://github.com/hearbenchmark/hear-eval-kit`

[5]BYOL-A* is a reimplementation of BYOL-A [7] by [26], which we use since [7] did not evaluate on the HEAR tasks.

[6]We sample the noise from a Gaussian distribution $\mathcal{N}(0, \lambda)$, where $\lambda \sim U(0, \alpha)$, $\alpha = 0.2$.

[7]HEAR-*L* consists of five HEAR tasks covering all three of the scene-based task subcategories: CREMA-D (speech), LibriCount (speech), FSD50K (environmental sound), ESC-50 (environmental sound), and GTZAN Genre (music).

**Table 1**: Results on HEAR speech and environmental sound scene-based tasks. Top two performing models for each task are shown underlined and highlighted. % $\uparrow_{\text{RAND}}$ refers to the fractional increase from the average score obtained by the random baseline for that model.

| | Speech | | | | | | | Environmental Sound | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | CREMA-D | LbC | SPC-5h | SPC-F | VocIm | VoxL | **Avg** (% $\uparrow_{\text{RAND}}$) | ESC-50 | FSD50K | Gunshot | **Avg** (% $\uparrow_{\text{RAND}}$) |
| [HEAR] CREPE | 0.383 | 0.499 | 0.180 | 0.211 | 0.051 | 0.142 | 0.244 | 0.301 | 0.159 | 0.863 | 0.441 |
| [HEAR] wav2vec2.0 | 0.656 | 0.692 | 0.838 | 0.879 | 0.080 | 0.493 | 0.606 | 0.561 | 0.342 | 0.848 | 0.584 |
| [HEAR] BYOL-A* | 0.623 | 0.788 | 0.896 | 0.924 | 0.137 | 0.390 | 0.626 | 0.789 | 0.489 | 0.875 | 0.7180 |
| [ABT] AudioNTT | 0.594 | 0.745 | 0.882 | 0.910 | 0.111 | 0.324 | 0.594 (17%) | 0.786 | 0.474 | 0.905 | 0.721 (24%) |
| [ABT] ViT$_C$-*B* (16 × 8) | 0.581 | 0.812 | 0.724 | 0.771 | 0.087 | 0.312 | 0.548 (140%) | 0.705 | 0.446 | 0.845 | 0.666 (49%) |

**Table 2**: Results on HEAR music scene-based tasks.

| | Music | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | Beijing | GTZAN-Genre | GTZAN-M/S | Mrd-Stroke | Mrd-Tonic | NSynth 5h | NSynth 50h | **Avg** (% $\uparrow_{\text{RAND}}$) |
| [HEAR] CREPE | 0.928 | 0.645 | 0.929 | 0.898 | 0.824 | 0.870 | 0.900 | 0.856 |
| [HEAR] wav2vec2.0 | 0.907 | 0.780 | 0.946 | 0.943 | 0.828 | 0.402 | 0.653 | 0.780 |
| [HEAR] BYOL-A* | 0.919 | 0.835 | 0.969 | 0.970 | 0.900 | 0.290 | 0.642 | 0.789 |
| [ABT] AudioNTT | 0.966 | 0.818 | 0.962 | 0.970 | 0.932 | 0.476 | 0.740 | 0.838 (−1%) |
| [ABT] ViT$_C$-*B* (16 × 8) | 0.869 | 0.765 | 0.992 | 0.952 | 0.897 | 0.280 | 0.632 | 0.769 (15%) |

**Table 3**: Results on HEAR timestamp-based tasks. Error rate (↓) indicates that a lower error rate is better. Table format adapted from [26].

| | DCASE | | MAESTRO | | **Avg** (% $\uparrow_{\text{RAND}}$) |
|---|---|---|---|---|---|
| **Model** | Onset FMS | Error rate (↓) | Onset FMS | Onset w/ Offset FMS | Onset FMS |
| [HEAR] CREPE | 0.552 | 0.420 | 0.3910 | 0.15 | 0.472 |
| [HEAR] wav2vec2.0 | 0.670 | 0.320 | 0.0328 | 0.009 | 0.351 |
| [HEAR] BYOL-A* | 0.499 | 0.503 | 0.0028 | 0.00029 | 0.251 |
| [ABT] AudioNTT | 0.761 | 0.274 | 0.04801 | 0.00672 | 0.405 (27%) |
| [ABT] ViT$_C$-*B* (16 × 8) | 0.722 | 0.275 | 0.0263 | 0.00429 | 0.374 (−10%) |

RRC is the most effective audio augmentation, attaining the highest performance of any of the audio augmentations when applied alone, achieving a HEAR-*L* average score of 65% (compared with 63% for Mixup, 60% for Noise, and 50% for RLF). RLF is by far the least effective augmentation.

## 6. CONCLUSION

In this paper, we presented ***Audio Barlow Twins*** (ABT), a novel self-supervised audio representation learning method which adapts Barlow Twins [13] to the audio domain. ABT pre-training on AudioSet [23] for 100 epochs with the AudioNTT encoder [21] results in model performance which is on a par with, and in several cases better than, BYOL-A [7]. We found commonly introduced augmentations to be harmful to ABT in certain settings. Future works should consider the effect on different downstream tasks of different augmentations that act directly on raw waveforms, within the ABT learning framework. Applying the augmentations directly on raw waveforms, instead of spectrograms, allows for a) a better control of the strength of these augmentations (as it is possible to *listen* directly to their effect), and b) a greater number of augmentations to be considered (e.g. pitch shift, time masking, time shift, time stretch, fade in/out, compression, etc.). We were unable to apply data augmentations during training directly on raw waveforms in this work due to an I/O bottleneck.

# References


[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. pages 1, 2

[2] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *CoRR*, vol. abs/2006.07733, 2020. pages 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," *CoRR*, vol. abs/2104.14294, 2021. pages 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. pages 1

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. pages 1

[6] Aaqib Saeed, David Grangier, and Neil Zeghidour, "Contrastive learning of general-purpose audio representations," *CoRR*, vol. abs/2010.10915, 2020. pages 1, 2

[7] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," 2021. pages 1, 2, 3, 4

[8] Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, and James R. Glass, "SSAST: self-supervised audio spectrogram transformer," *CoRR*, vol. abs/2110.09784, 2021. pages 1, 3

[9] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer, "Masked autoencoders that listen," 2022. pages 1, 3

[10] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong, "ibot: Image BERT pre-training with online tokenizer," *CoRR*, vol. abs/2111.07832, 2021. pages 1

[11] Xinlei Chen, Saining Xie, and Kaiming He, "An empirical study of training self-supervised vision transformers," *CoRR*, vol. abs/2104.02057, 2021. pages 1

[12] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas, "Masked siamese networks for label-efficient learning," 2022. pages 1

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *CoRR*, vol. abs/2103.03230, 2021. pages 1, 2, 4

[14] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," *CoRR*, vol. abs/2011.10566, 2020. pages 1

[15] Yao-Hung Hubert Tsai, Shaojie Bai, Louis-Philippe Morency, and Ruslan Salakhutdinov, "A note on connecting barlow twins with negative-sample-free contrastive learning," 2021. pages 1

[16] Yuandong Tian, Xinlei Chen, and Surya Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," *CoRR*, vol. abs/2102.06810, 2021. pages 1

[17] Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E. O'Connor, and Xavier Serra, "Unsupervised contrastive learning of sound event representations," 2020. pages 1, 2

[18] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *CoRR*, vol. abs/2006.09882, 2020. pages 1

[19] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabeleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W. Schuller, "Audio self-supervised learning: A survey," 2022. pages 2

[20] Haider Al-Tahan and Yalda Mohsenzadeh, "Clar: Contrastive learning of auditory representations," 2020. pages 2

[21] Yuma Koizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino, "The ntt dcase2020 challenge task 6 system: Automated audio captioning with keywords and sentence length estimation," 2020. pages 2, 3, 4

[22] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino, "Byol for audio: Exploring pre-trained general-purpose audio representations," 2022. pages 2

[23] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780. pages 2, 4

[24] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick, "Early convolutions help transformers see better," *CoRR*, vol. abs/2106.14881, 2021. pages 3

[25] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk, "Hear: Holistic evaluation of audio representations," 2022. pages 3

[26] Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak, "Byol-s: Learning self-supervised speech representations by bootstrapping," 2022. pages 3, 4

[27] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "Crepe: A convolutional representation for pitch estimation," 2018. pages 3

[28] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. pages 3