# BYOL-S: Learning Self-supervised Speech Representations by Bootstrapping

**Gasser Elbanna*** GASSER.ELBANNA@EPFL.CH
*Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*
**Neil Scheidwasser-Clow** NEIL.SCHEIDWASSER-CLOW@EPFL.CH
*Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*
**Mikolaj Kegler** MIKOLAJ.KEGLER16@IMPERIAL.AC.UK
*Imperial College London, London, United Kingdom*
**Pierre Beckmann** PIERRE.BECKMANN@UNIL.CH
*Université de Lausanne, Lausanne, Switzerland*
**Karl El Hajal*** KARL.ELHAJAL@EPFL.CH
*Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*
**Milos Cernak** MILOS.CERNAK@IEEE.ORG
*Logitech Europe S.A., Lausanne, Switzerland*



## Abstract

Methods for extracting audio and speech features have been studied since pioneering work on spectrum analysis decades ago. Recent efforts are guided by the ambition to develop general-purpose audio representations. For example, deep neural networks can extract optimal embeddings if they are trained on large audio datasets. This work extends existing methods based on self-supervised learning by bootstrapping, proposes various encoder architectures, and explores the effects of using different pre-training datasets. Lastly, we present a novel training framework to come up with a *hybrid* audio representation, which combines handcrafted and data-driven learned audio features. All the proposed representations were evaluated within the HEAR NeurIPS 2021 challenge for auditory scene classification and timestamp detection tasks. Our results indicate that the hybrid model with a convolutional transformer as the encoder yields superior performance in most HEAR challenge tasks.

**Keywords:** audio embeddings, representation learning, self-supervised learning, hybrid representations

## 1. Introduction

Humans are able to learn, memorize, and distinguish various auditory patterns from limited data by projecting low-level audio inputs to high-level representations in the brain (Griffiths, 1999). Inspired by human capabilities, a substantial body of research has been dedicated over the past few decades to build models capable of extracting and representing auditory information. Historically, handcrafted feature sets, based on digital signal processing (DSP), have been employed to extract audio representations (Liu et al., 1998; Eyben et al., 2010). However, the recent success of deep learning in computer vision and natural language

* GE and KEH performed this work as interns at Logitech.

processing has propelled the development of data-driven frameworks, where deep neural networks (DNNs) are trained on large audio corpora to capture crucial features (Hershey et al., 2017; Van den Oord et al., 2018).

Two main methods currently coexist to build deep-learning-based audio representations: supervised (Hershey et al., 2017; Beckmann et al., 2021) and self-supervised learning (Baevski et al., 2020; Niizumi et al., 2021). While supervised methods have been at the center of most initial breakthroughs in vision, audio, and language understanding, they are inevitably limited by their reliance on well-defined labels for each training data input. Conversely, self-supervised learning aims at leveraging relations within input data to generate pseudo-labels, thus creating proxy supervised tasks (Liu et al., 2021; Murphy, 2022). Recently, several self-supervised models have been proposed as robust general-purpose audio representations (Van den Oord et al., 2018; Shor et al., 2020; Saeed et al., 2021; Niizumi et al., 2021; Shor et al., 2022). Most models are trained using contrastive learning setups, where an encoder network learns to produce a latent space representation by assessing the degree of similarity between input examples (Van den Oord et al., 2018; Shor et al., 2020; Saeed et al., 2021). In this framework, similar inputs should be mapped closer in the latent space, whereas unrelated examples should appear more distant. In the context of audio, *similarity* can be measured in terms of temporal proximity (Shor et al., 2020) or, more simply, whether two audio segments originate from the same source or not (Saeed et al., 2021).

However, Niizumi et al. (2021) argue that contrastive learning frameworks may potentially suffer from several limitations when it comes to audio representation learning. For example, similar rhythmic patterns can be found in different audio sources. Alternatively, short impulsive sounds, such as glass breaking, may be found only in a single example of an audio clip and thus appear as "dissimilar" to other inputs from the same clip. Consequently, Niizumi et al. (2021) proposed Bootstrap Your Own Latent for Audio (BYOL-A) to learn audio representations by comparing augmented views of a single audio segment. Inspired by the success of BYOL for self-supervised image representation (Grill et al., 2020), BYOL-A achieved competitive results in various tasks, including speaker identification, language identification, speech commands, and musical instrument classification.

That being said, handcrafted DSP-based feature sets remain widely used in various speech- and music-related applications. For instance, extensive feature sets such as openSMILE (Eyben et al., 2010) often constitute a strong baseline in paralinguistic tasks and challenges (Schuller et al., 2013, 2016, 2020). Unlike DNNs, such frameworks are completely transparent and interpretable. Moreover, they do not require any training, thus reducing both computational costs and risks of model overfitting (omitting any bias introduced by the handcrafted feature set designer). While the above-outlined properties make the DSP-based features remain relevant for many applications, most recent pre-trained DNNs significantly outperform handcrafted feature sets (Shor et al., 2020; Scheidwasser-Clow et al., 2022). Accordingly, in this paper, we propose new extensions of BYOL-A for speech representation learning. Whereas the original BYOL-A model was trained on the entire AudioSet (Gemmeke et al., 2017), a large audio dataset with more than 5800 hours of audio, here we retrained different models on a speech-specific subset from AudioSet creating BYOL for speech (BYOL-S). Originally, the BYOL-S model was developed for speech emotion recognition (SER) and outperformed BYOL-A and other pre-trained models in the con-

text of a speech emotion recognition adaptation benchmark (SERAB) (Scheidwasser-Clow et al., 2022). Here, we extend our previous work to assess BYOL-S, as a general-purpose audio representation and to thoroughly study the impact of different hyperparameters and training protocols on the model performance. In particular, we introduce different encoder architectures (Section 2.1) than the default one used in Niizumi et al. (2021). With the aim of leveraging the best of both DSP- and DNN-based approaches, we finally assessed the impact of incorporating DSP-based features to the BYOL training paradigm. This led to the development of a novel pre-training protocol for BYOL-S that combines learned and fixed DSP-based handcrafted features (Section 2.2). Such a *hybrid* approach can facilitate the pre-training of the model by grounding the complex DNN features with the considerably simpler DSP-based ones.

All models were evaluated in the context of the Holistic Evaluation of Audio Representations (HEAR) 2021 challenge[1], a challenge aimed at designing general-purpose audio embeddings. Launched at NeurIPS 2021, the HEAR challenge featured a new 16-task benchmark suite to compare models, including scene-based (i.e., audio classification) and timestamp-based tasks (i.e., sound event detection). Importantly, the benchmark comprises data from a variety of sources, e.g., human speech, environment sounds, and music.

## 2. Methods

All proposed models constitute extensions of BYOL-A (Niizumi et al., 2021), an adaptation of *bootstrap your own latent* (BYOL) (Grill et al., 2020) for general-purpose audio representation learning. More specifically, we extended our previous work (Scheidwasser-Clow et al., 2022) for speech representation by varying the encoder networks within the BYOL framework, as shown in Figure 1. In addition, we explored *hybrid* approaches by combining BYOL-like networks with hand-crafted features from openSMILE (Eyben et al., 2010), with the aim of assisting the self-supervised network with spectral and prosodic information during the training process.

While contrastive learning setups typically rely on comparing different audio segments to learn representations, BYOL-A models learn by comparing two augmented versions of a single audio input (Niizumi et al., 2021). During pre-training, the input audio is first preprocessed into a 64-band log-mel magnitude spectrogram (LMS) to produce a 2D input. All spectrograms were generated within a frequency range of 60 to 7800 Hz, a sampling rate of 16 kHz, a window length of 25 ms, and a hop size of 10 ms. By default, the BYOL-A framework uses random 96-frame-long segments from the input LMS for model training (Niizumi et al., 2021), approximately corresponding to 0.95 s of audio. We experimented with changing the training window size, as discussed in Section 3.2. However, the pre-trained model can be fed with audio samples of any duration during the inference stage. The extracted log-mel spectrograms are then of dimension $64 \times T$, where $T$, the number of spectrogram frames, depends on the sample length.

Subsequently, the LMS input is fed to an augmentation module, which first standardizes the spectrogram before applying two different data augmentations: mixup (Zhang et al., 2018), i.e., adding randomly mixed audio samples in the background of the input, and random resize cropping, which in the context of audio spectrograms can be equated with

1. https://neuralaudio.ai/

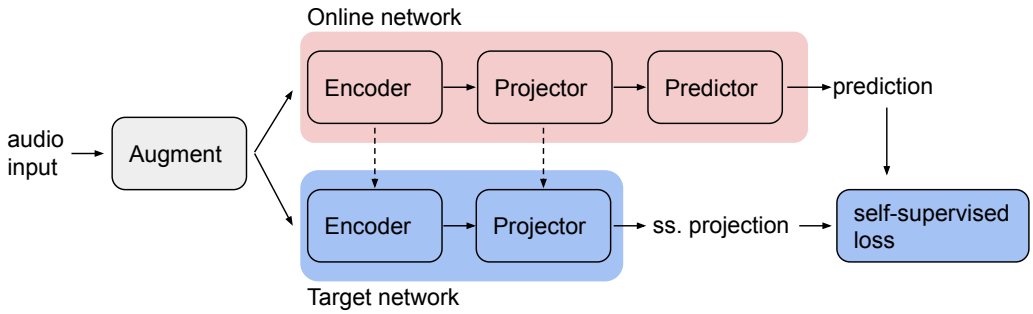


Figure 1: Schematic of the BYOL-A architecture, showing the main modules involved in the training paradigm. Adapted from (Grill et al., 2020; Niizumi et al., 2021). ss: self-supervised. The dotted lines indicate the fact that the target network parameters are updated as a moving average of the online parameters.

pitch-shifting and time-stretching. The random augmentation is applied twice to produce two randomly augmented views of the input spectrogram. These views thus share the same input source signal but are processed using two slightly different audio augmentations. At last, both views are re-standardized to account for any statistical drifts.

The models are trained to predict the representation of the first augmented version from the representation of the second. To that end, the two augmented views are fed respectively to an *online* and a *target* network (Figure 1), which have a different set of weights. The rationale behind encompassing two different networks is to build a self-supervised learning paradigm that makes the target network act as a pseudo label generator that is, then, compared to the online network output. Both networks comprise an *encoder* and a *projection* head: the encoder extracts a representation of the augmented input, whereas the *projection* head helps to map the representation to a lower dimensional latent space. Additionally, the online network includes a predictor head to avoid collapsing solutions with constant representations (Grill et al., 2020). Finally, a simple mean-squared error (MSE) is used as a loss function to minimize the difference between the online predictor and the target projector outputs (ss. projection). While the online parameters were updated using Adam optimization (Kingma and Ba, 2015) with a learning rate of 0.0003, target parameters were updated as an exponential moving average of the online parameters, which empirically showed to improve training stability and yield more robust embeddings (Grill et al., 2020).

The default encoder used in BYOL-A is a simple convolution neural network (CNN) adapted from a solution to the Automated Audio Captioning task of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Challenge (Niizumi et al., 2021). More specifically, the network comprises three convolution blocks (1 block = Conv2d-BatchNorm-ReLU-MaxPool2d) with 64 3x3 filters each, followed by two linear layers projecting the output of the final block onto a 2048-dimensional embedding.

In addition to the default encoder, we explore different encoder architectures in Section 2.1 and evaluate their performance across the HEAR challenge tasks.

### 2.1. Encoder Architectures

Three different encoding networks were benchmarked against the HEAR tasks to gain further insight into the robustness of certain audio representations. These networks, presented in Table 1, are: ResNetish-34 (Hershey et al., 2017), a convolutional LSTM inspired by (Passricha and Aggarwal, 2020) and the convolutional vision transformer (CvT) (Wu et al., 2021). While both ResNetish and CLSTM networks rely on CNNs for feature extraction, CLSTMs also incorporate recurrent neural networks (RNNs) to leverage the temporal properties of audio data. Inspired by the success of transformers in computer vision, Wu et al. (2021) showed with CvT that including convolutional token embeddings and projections within transformer layers yielded competitive results for image classification.

Table 1: Encoder architectures studied within the BYOL-S framework.

| **Encoder** | **Parameters (M)** | **Embedding size** |
|---|---|---|
| Default (CNN) (Niizumi et al., 2021) | 5.3 | 2048 |
| ResNetish-34 (Hershey et al., 2017) | 21.3 | 2048 |
| CLSTM (Passricha and Aggarwal, 2020) | 18.6 | 1024 |
| CvT (Wu et al., 2021) | 5.0 | 2048 |

#### 2.1.1. ResNetish Model

An audio version of a 34-layer residual network (He et al., 2016) was implemented using the same modifications as Hershey et al. (2017) for large-scale audio classification. In accordance with Shor et al. (2020), we refer to this encoder as ResNetish-34 (Table 1). Although ResNetish-50 was the most robust for AudioSet classification compared to other common CNNs (Hershey et al., 2017), we implemented a lighter version, ResNetish-34, to avoid overfitting the pre-training datasets. The implementation was derived from an adaptation of ResNetish in `PyTorch`[2] with a final embedding size of 2048.

#### 2.1.2. Convolution with LSTM

The ability of RNNs such as LSTM-based networks to capture temporal dependencies in sequential data led us to implement a bidirectional LSTM (BiLSTM) (Graves and Schmidhuber, 2005) network with CNN features (CLSTM) to explore the benefits of RNNs. This model is inspired by previous work by Passricha and Aggarwal (2020) on automatic speech recognition (ASR). Their model comprised two 256-filter convolution layers (with 9x9 and 4x3 (frequency x time) kernels), followed by two BiLSTM layers and three feedforward layers to generate higher-order representations. While the CNN architecture was similar to the original implementation, we only used one 512-layer BiLSTM layer and one 1024-dimensional fully connected layer to prevent overfitting during pre-training.

2. https://github.com/daisukelab/sound-clf-pytorch

#### 2.1.3. Convolution Vision Transformer

Inspired by transformer architectures for vision (Dosovitskiy et al., 2021), Wu et al. (2021) proposed the Convolutional vision Transformer (CvT), which leverages the advantages of both CNNs (i.e., detecting fine-grained local patterns) and Transformers (i.e., learning long-range global context). The network architecture comprises three stages. Each stage comprises a convolutional token embedding layer and a convolutional transformer block, the latter of which includes a convolutional projection layer followed by multi-head self-attention. We adapted the implementation from[3] and built a lightweight version of CvT with one transformer block per stage with embedding sizes of 64, 256, and 512, respectively. Using temporal aggregation, the final layer outputs a vector of dimensions 2048. Such an encoder yielded competitive results in SER tasks (Scheidwasser-Clow et al., 2022).

### 2.2. Hybrid Representations

Despite current progress in deep learning-based audio and speech representation, hand-crafted (DSP-based) feature sets remain competitive in various paralinguistic challenges (Schuller et al., 2013, 2020). This motivated us to study the benefits of combining DSP-based features and data-driven features by adding a third *supervision* module to the BYOL-S framework. Here, the module simply consists of features extracted using the ComParE 2016 acoustic feature set (Schuller et al., 2013) from openSMILE (OS) (Eyben et al., 2010)[4]. This extensive feature set comprises 6373 static features, including acoustic functionals, low-level descriptors (LLDs) and LLD derivatives.

We denote the resulting model as *Hybrid BYOL-S*, since the online network learns to strike a balance between self-supervised, learned features and supervised features from a fixed feature set, as illustrated in Figure 2. More specifically, the model is trained to optimize a sum of two loss functions: the *self-supervised* BYOL loss between the online and target outputs ($\mathcal{L}_{ss}$), and a second *supervised* loss ($\mathcal{L}_{sup}$), computed as the mean squared error between the outputs of the online and DSP-based networks. To get a more exhaustive view of the performance of the hybrid model, we finally explored different weights $\alpha$ and $\beta$ for $\mathcal{L}_{sup}$ and $\mathcal{L}_{ss}$, respectively, leading to the following hybrid loss:

$$\mathcal{L}_{hybrid} = \alpha\mathcal{L}_{sup} + \beta\mathcal{L}_{ss} \tag{1}$$

Note that the output dimension of the online and target networks' projectors was changed from 256 to 6373 to accommodate the size of the openSMILE features. However, the embedding size of the encoders (2048) remained unchanged.

### 2.3. Implementation Details

For all experiments, we used the same pre-training hyperparameters used in Niizumi et al. (2021)[5], in which models were trained for 100 epochs using Adam and a batch size of 256 with a learning rate of 0.0003. Pre-trained models and code for embedding computation (explained in Section 2.4) are available at[6].

3. https://github.com/lucidrains/vit-pytorch
4. https://audeering.github.io/opensmile-python
5. https://github.com/nttcslab/byol-a
6. https://github.com/GasserElbanna/serab-byols

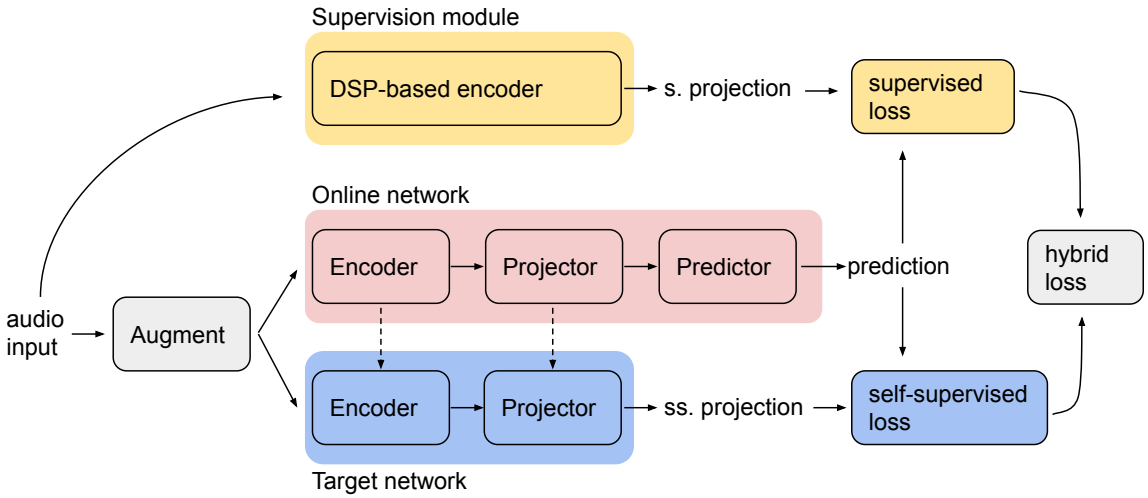


Figure 2: The hybrid BYOL-S framework leverages both self-supervised features and hand-crafted features from openSMILE. *s*: supervised, *ss*: self-supervised. The dotted arrows indicate that the target parameters are updated as a moving average of the online parameters.

### 2.4. Embedding Generation and Downstream Evaluation

All 16 tasks from the HEAR challenge were used to evaluate the robustness of our pre-trained models, which include both scene-based and timestamp-based tasks. Scene-based tasks (Table 2) consist of multi-class or multi-label audio classification and can be further divided into three categories: speech, environmental sounds, and music. Although containing human and musical sounds, FSD50K was included in the "environmental sounds" category. On the other hand, timestamp-based tasks (Table 3) consist of sound event detection or transcription and include two datasets: MAESTRO and Task 2 from DCASE 2016.

All audio clips were sampled at 48 kHz[7]. For each dataset, all recordings were resampled to 16 kHz using `torchaudio`[8] to align with our model implementation. Subsequently, embeddings were generated from pre-trained models differently depending on the task type. For scene-based tasks, the entire audio samples were used to produce the embedding, while timestamp-based tasks required the sample to be chunked into fixed-size segments before generating window representations aligned with their corresponding timestamps. Finally, the embeddings were evaluated using the `hear-eval`[9] toolkit. This toolkit trains a shallow fully-connected predictor on a train set of a downstream task, which is optimized on a validation set before evaluation on an unseen test set. Despite slight discrepancies between the official challenge results and the results presented herein due to different resampling methods, the produced scores remained stable and consistent with Turian et al. (2022). We observe however a higher discrepancy for timestamp-related tasks, especially MAESTRO, which we ascribe to higher sensitivity to resampling methods.

7. https://doi.org/10.5281/zenodo.5887964
8. https://pytorch.org/audio/stable/transforms.html#torchaudio.transforms.Resample
9. https://github.com/neuralaudio/hear-eval-kit

Table 2: HEAR scene-based tasks and datasets. clf.: classification. Adapted from Turian et al. (2022).

| Dataset | Task | # Classes | Total duration (h) |
|---|---|---|---|
| | Speech | | |
| CREMA-D (Cao et al., 2014) | Emotion recognition | 6 | 5.3 |
| LibriCount (Stöter et al., 2018) | Speaker count estimation | 11 | 7.9 |
| Speech Commands 5h/full (Warden, 2018) | Keyword spotting | 12 | 6.4/27.9 |
| Vocal Imitations (Kim et al., 2018a) | Vocal imitation clf. | 302 | 17.5 |
| VoxLingua (Valk and Alumäe, 2021) | Language identification | 10 | 5.0 |
| | Environment sounds | | |
| Beehive states (Nolasco et al., 2019) | Beehive identification | 2 | 96 |
| ESC-50 (Piczak, 2015) | Sound classification | 50 | 2.8 |
| FSD50K (Fonseca et al., 2022) | Sound classification | 200 | 108.3 |
| Gunshot triangulation (Cooper and Shaw, 2020) | Gunfire location | 4 | 0.04 |
| | Music | | |
| Beijing Opera (Tian et al., 2014) | Instrument classification | 4 | 0.31 |
| GTZAN (Genre) (Tzanetakis and Cook, 2002) | Genre classification | 10 | 8.3 |
| GTZAN (Speech/Music)[10] | Speech vs music clf. | 2 | 1.07 |
| Mridangam (Anantapadmanabhan et al., 2013) | Stroke & tonic clf. | 10 & 6 | 1.6 |
| NSynth 5h/50h (Engel et al., 2017) | Pitch & chroma clf. | 88 & 12 | 5.6/54.5 |

Table 3: HEAR timestamp-based tasks and datasets.

| Dataset | Task | Total duration (h) |
|---|---|---|
| MAESTRO (Hawthorne et al., 2019) | Music transcription | 6.2 |
| DCASE 2016 Task 2 (Mesaros et al., 2018) | Office sound detection | 2.4 |

## 3. Results

Tables 5, 6 and 7 present the results for scene-based tasks (Table 2) pertaining to speech, environmental sounds, and music, respectively. Test accuracy (in percentage) was used as the evaluation metric for each task except FSD50K (mAP). The best scores are shown in bold. For comparison, each set of experiments within each table is evaluated against the scores obtained by two HEAR baseline models:

- `wav2vec2` (Baevski et al., 2020), a self-supervised framework pretrained on 100K hours of speech from VoxPopuli (Wang et al., 2021). The model comprises a 1D convolutional feature encoder followed by a positional transformer for context representation.
- `CREPE` (Kim et al., 2018b), a 1D CNN for pitch estimation pre-trained on 16 hours of synthesized music.

10. http://marsyas.info/downloads/datasets.html#music-speech

### 3.1. Contribution of pre-training datasets: BYOL-A, BYOL-S and BYOL-S++

First, we considered re-training BYOL-A from scratch with different pre-training datasets. Whereas BYOL-A was pre-trained on AudioSet (Gemmeke et al., 2017), our early submission (BYOL-S) was only pre-trained on speech samples of AudioSet. Moreover, we introduce BYOL-S++, trained on LibriSpeech (Panayotov et al., 2015) in addition to the speech subset of AudioSet. With approximately 960 hours of data (Table 4), enriching the pre-training corpus with LibriSpeech samples further diversifies the format of speech samples from which the model could learn. Whereas AudioSet mostly features spontaneous speech surrounded by environmental sounds or music, LibriSpeech consists of read English speech derived from audiobooks recorded in a studio environment. All results pertaining to these experiments are shown in the **Pre-training dataset** section of Tables 5, 6 and 7.

Table 4: Datasets used for self-supervised training

| Model Name | Dataset | Duration (h) |
|---|---|---|
| BYOL-A | AudioSet | 5800 |
| BYOL-S | AudioSet (Speech subset) | 2190 |
| BYOL-S++ | AudioSet (Speech subset) + LibriSpeech | 3150 |

### 3.2. Contribution of the pre-training window size

All models were originally pre-trained with random 96-frame segments of audio spectrograms, corresponding to 0.95 s of audio. To study the effect of this window length, we compared four versions of BYOL-S, pre-trained with a window size of 0.5, 0.95, 1.425 and 2 s, respectively. As mentioned in Section 2, all models could still be fed with samples of any duration for downstream evaluation. The **BYOL-S window size (s)** section of Tables 5, 6 and 7 shows the results obtained by training BYOL-S with the window sizes mentioned above. To reduce training time, the batch size was reduced from 256 to 128 for these experiments.

### 3.3. Comparison of encoder architectures: ResNetish-34, CLSTM, CvT

The **BYOL-S encoder** section of Tables 5, 6 and 7 shows the results obtained by replacing the default encoder in BYOL-S with the DNNs described in Section 2.1. The other pre-training parameters were unchanged.

### 3.4. Hybrid versions of BYOL-S and BYOL-S/CvT

Here, we evaluated the performance of hybrid models (Figure 2), obtained by combining handcrafted and learnable feature sets during model pre-training. Following Section 3.3, two models were evaluated against the HEAR benchmark: Hybrid BYOL-S (using the "default" BYOL-S encoder) and Hybrid BYOL-S/CvT (i.e., with CvT encoding).

Moreover, to assess the impact of self-supervised and supervised losses on overall performance, we tested different values for loss weights $\alpha$ and $\beta$ (Eq. 1) when pre-training the hybrid BYOL-S/CvT. All results pertaining to these experiments are shown in the **Hybrid models** and **Hybrid BYOL-S/CvT** sections of Tables 5, 6, and 7.

Lastly, we validated the relevance of the hybrid training protocol against embeddings produced using only openSMILE features and the concatenation of the latter with BYOL-S/CvT features. Here, we modified the embedding generation scripts in the `hear-eval` toolkit by applying per-fold standardization to account for large values produced by openSMILE (mainly due to the computation of LLD functionals and deltas), which hindered convergence during downstream training. All results pertaining to these experiments are shown in the **Hybrid** section of Tables 5, 6 and 7.

Table 5: Top-1 accuracy of all proposed models on speech-related tasks. + denotes concatenation.

| | CREMA-D | LibriCount | Speech Commands (5h) | Speech Commands (all) | Vocal Imitations | VoxLingua | Average |
|---|---|---|---|---|---|---|---|
| **HEAR baselines:** | | | | | | | |
| CREPE | 36.2 | 49.9 | 16.8 | 19.6 | 5.1 | 15.1 | 23.8 |
| wav2vec2 | 65.7 | 67.6 | 79.7 | 88.5 | 7.2 | **49.7** | 59.7 |
| **Pre-training dataset:** | | | | | | | |
| BYOL-A | 62.3 | 78.8 | 89.6 | 92.4 | 13.7 | 39.0 | 62.6 |
| BYOL-S | **66.4** | 78.5 | 92.6 | 94.3 | 15.1 | 41.2 | 64.7 |
| BYOL-S++ | **66.4** | **80.0** | **93.2** | **95.0** | **15.4** | 47.8 | **66.3** |
| **BYOL-S window size (s):** | | | | | | | |
| 0.5 | **66.5** | **86.0** | 91.2 | 93.4 | 14.2 | 43.0 | 64.8 |
| 0.95 | 65.5 | 83.0 | 91.9 | **94.5** | 14.8 | 44.4 | 65.2 |
| 1.425 | 65.7 | 81.0 | 90.0 | 93.1 | **16.0** | 43.1 | 64.7 |
| 2 | 65.6 | 81.8 | **92.2** | 93.4 | 14.3 | 47.3 | **65.8** |
| **BYOL-S encoder:** | | | | | | | |
| Default | **66.4** | 78.6 | **92.6** | **94.3** | 15.1 | 41.2 | 64.7 |
| Resnetish-34 | 63.5 | 77.0 | 83.2 | 88.9 | 13.3 | 35.8 | 60.3 |
| CLSTM | 64.0 | 78.1 | 91.3 | 92.5 | 11.9 | 24.3 | 60.4 |
| CvT | **66.4** | **84.8** | 92.0 | 93.1 | **16.2** | 37.0 | **64.9** |
| **Hybrid models:** | | | | | | | |
| openSMILE (OS) only | 59.4 | 66.7 | 70.1 | 80.2 | 13.1 | 25.1 | 52.4 |
| BYOL-S/CvT + OS | 65.4 | 82.7 | 77.6 | 86.9 | **17.4** | 29.5 | 59.9 |
| Hybrid BYOL-S/Default | 66.1 | 81.6 | 91.6 | 93.8 | 14.2 | 43.8 | 65.2 |
| Hybrid BYOL-S/CvT | **67.2** | **83.5** | **92.6** | **95.8** | 16.3 | 42.2 | **66.3** |
| **Hybrid BYOL-S/CvT:** $\alpha$**:**$\beta$ **ratio** | | | | | | | |
| 1:4 | 62.9 | **84.4** | 86.2 | 90.7 | 12.0 | 27.1 | 60.6 |
| 1:2 | 64.4 | 83.5 | 89.6 | 92.7 | 13.6 | 35.3 | 63.2 |
| 2:3 | 66.0 | 84.1 | 91.9 | 93.8 | 14.6 | 38.7 | 64.9 |
| 1:1 | **67.2** | 83.5 | 92.6 | **95.8** | **16.3** | 42.2 | **66.3** |
| 3:2 | 66.7 | 82.7 | 92.5 | 94.8 | 15.2 | 34.0 | 64.3 |
| 2:1 | 65.8 | 81.2 | 92.7 | 94.7 | 15.8 | 39.8 | 65.0 |
| 4:1 | 66.5 | 80.6 | **93.1** | 94.6 | 16.2 | 38.0 | 64.8 |

Table 6: Performance on environmental sound-related datasets. Top-1 accuracy was used for each dataset but FSD50K (mAP). Due to exhausting GPU memory problems for several models (shown with superscript *), the reported average does not include *Beehive states*.

| | Beehive states | ESC | FSD50K | Gunshot triangulation | Average |
|---|---|---|---|---|---|
| **HEAR baselines:** | | | | | |
| CREPE | 50.4 | 29.4 | 15.9 | 91.7 | 45.7 |
| wav2vec2 | -* | 59.2 | 34.6 | 77.1 | 57.0 |
| **Pre-training dataset:** | | | | | |
| BYOL-A | 48.8 | 78.9 | 48.9 | 87.5 | 71.8 |
| BYOL-S | 52.8 | **81.9** | **49.9** | **96.4** | **76.1** |
| BYOL-S++ | **55.0** | 80.0 | 49.4 | 92.9 | 74.1 |
| **BYOL-S window size (s):** | | | | | |
| 0.5 | **59.7** | **83.3** | 49.5 | 89.3 | 74.0 |
| 0.95 | 57.5 | 81.4 | **50.0** | **95.2** | **75.5** |
| 1.425 | 56.5 | 80.3 | 49.9 | 89.3 | 73.2 |
| 2 | 53.5 | 81.3 | 49.6 | **95.2** | 75.4 |
| **BYOL-S encoder:** | | | | | |
| Default | **52.8** | **81.9** | **49.9** | **96.4** | **76.1** |
| Resnetish-34 | **52.8** | 71.5 | 43.3 | 86.3 | 67.0 |
| CLSTM | 51.4 | 73.9 | 39.4 | 86.9 | 66.7 |
| CvT | -* | 79.9 | 48.3 | 89.3 | 72.5 |
| **Hybrid models:** | | | | | |
| openSMILE (OS) only | -* | 68.2 | 34.5 | **98.8** | 67.2 |
| BYOL-S/CvT + OS | -* | 76.9 | 44.2 | **98.8** | 73.3 |
| Hybrid BYOL-S/Default | **53.0** | 82.4 | 48.9 | 95.2 | 75.5 |
| Hybrid BYOL-S/CvT | -* | **83.8** | **52.0** | 96.4 | **76.8** |
| **Hybrid BYOL-S/CvT:** $\alpha$**:**$\beta$ **ratio** | | | | | |
| 1:4 | -* | 72.5 | 42.7 | 88.7 | 68.0 |
| 1:2 | -* | 78.4 | 46.1 | 90.5 | 71.7 |
| 2:3 | -* | 79.6 | 47.8 | 96.4 | 74.6 |
| 1:1 | -* | **83.8** | **50.2** | 96.4 | **76.8** |
| 3:2 | -* | 82.3 | 47.9 | **97.6** | 75.9 |
| 2:1 | -* | 83.1 | 48.3 | 95.8 | 75.7 |
| 4:1 | -* | 82.4 | 48.7 | 94.0 | 75.0 |

Table 7: Top-1 accuracy of all proposed models on music-related tasks. S/M: Speech vs. Music task.

| | Beijing Opera | GTZAN | | Mridangam | | NSynth (5h) | | NSynth (50h) | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Genre | S/M | Stroke | Tonic | Pitch | Chroma | Pitch | Chroma | |
| **HEAR baselines:** | | | | | | | | | | |
| CREPE | 93.2 | 64.5 | 86.7 | 88.7 | 82.3 | **87.2** | **93.4** | **89.5** | **95.2** | **86.7** |
| wav2vec2 | 89.4 | 78.0 | 92.3 | 94.7 | 82.8 | 40.0 | 44.6 | 66.7 | 71.9 | 73.4 |
| **Pre-training dataset:** | | | | | | | | | | |
| BYOL-A | 91.9 | 83.5 | **96.9** | 97.0 | 90.0 | 29.0 | 36.0 | 64.2 | 67.5 | 72.3 |
| BYOL-S | 91.1 | 83.8 | 92.3 | 97.3 | **92.9** | 42.0 | 44.0 | 70.0 | 73.4 | 76.3 |
| BYOL-S++ | **95.3** | **83.9** | 93.8 | **97.4** | 91.7 | 40.0 | 41.8 | 69.6 | 73.2 | 76.3 |
| **BYOL-S window size (s):** | | | | | | | | | | |
| 0.5 | **94.9** | 83.5 | 96.9 | **97.3** | **93.0** | 38.2 | 40.6 | 71.2 | 74.5 | 76.7 |
| 0.95 | 93.2 | 82.5 | 96.2 | **97.3** | 92.5 | 38.2 | 39.0 | 69.0 | 72.1 | 75.6 |
| 1.425 | 94.5 | **83.6** | **98.5** | 97.0 | 92.5 | 35.2 | 37.6 | 67.7 | 71.6 | 75.4 |
| 2 | 94.1 | 83.4 | **98.5** | 97.1 | 91.9 | 35.8 | 37.4 | 66.2 | 69.2 | 74.8 |
| **BYOL-S encoder:** | | | | | | | | | | |
| Default | 91.1 | 83.8 | 92.3 | 97.3 | 92.9 | 42.0 | 44.0 | 70.0 | 73.4 | 76.3 |
| Resnetish-34 | 92.4 | 77.4 | 96.9 | 96.1 | 88.7 | 26.0 | 22.4 | 44.9 | 47.4 | 65.8 |
| CLSTM | **97.0** | 75.9 | 96.2 | 96.7 | 89.7 | 28.6 | 32.0 | 57.6 | 65.0 | 71.0 |
| CvT | 96.2 | **84.2** | **97.7** | **97.5** | **94.0** | 53.0 | 55.2 | 76.1 | 85.0 | 82.1 |
| **Hybrid models:** | | | | | | | | | | |
| openSMILE (OS) only | 89.4 | 78.2 | **97.6** | 95.6 | 88.0 | 47.0 | 51.0 | 73.9 | 78.3 | 77.7 |
| BYOL-S/CvT + OS | 92.8 | 83.4 | 96.9 | 96.8 | 93.1 | 53.4 | 56.6 | 80.9 | 85.6 | 82.2 |
| Hybrid BYOL-S/Default | 92.8 | **85.9** | 96.9 | 96.1 | 88.7 | 35.8 | 38.0 | 71.1 | 75.0 | 75.6 |
| Hybrid BYOL-S/CvT | **94.5** | 85.8 | 96.2 | **97.7** | **94.9** | 56.8 | 59.2 | 79.5 | 83.0 | 83.1 |
| **Hybrid BYOL-S/CvT:** $\alpha$**:**$\beta$ **ratio** | | | | | | | | | | |
| 1:4 | 94.9 | 82.2 | 95.3 | 97.2 | 96.0 | 66.4 | 68.6 | 80.8 | 84.6 | 85.1 |
| 1:2 | **96.2** | 84.8 | 94.6 | 97.6 | **96.7** | 65.4 | 69.6 | 81.6 | 85.8 | 85.8 |
| 2:3 | 93.2 | 86.1 | 93.8 | 97.5 | 96.6 | 67.0 | 69.8 | 81.1 | 85.4 | 85.6 |
| 1:1 | 94.5 | 85.8 | **96.2** | 97.7 | 94.9 | 56.8 | 59.2 | 79.5 | 83.0 | 83.1 |
| 3:2 | 95.8 | 85.6 | 94.6 | 97.7 | **96.7** | 64.0 | 67.0 | 81.9 | 85.9 | 85.5 |
| 2:1 | 94.9 | 86.4 | 94.6 | **97.8** | 96.6 | 62.8 | 66.4 | 80.9 | 85.0 | 85.0 |
| 4:1 | 94.9 | **86.8** | 92.3 | **97.8** | 96.4 | 63.2 | 65.6 | 79.9 | 84.3 | 84.6 |

### 3.5. Timestamp Embeddings

In this section, we evaluate our models on tasks that depend on generating embeddings from segments, instead of the entire audio recording, to detect sound events—onset and offset—or music transcription. We tested this format on the two timestamp-based tasks from the HEAR challenge: DCASE and MAESTRO (Table 3). For the best model (hybrid BYOL-S/CvT), two hyperparameters were tuned: the window size and the hop size. The evaluation results are presented in Tables 8 and 9.

Table 8: Results on timestamp-based tasks. All models used input windows of 1 s with a 50 ms hop size. ↓ indicates lower is better. Due to exhausting GPU memory problems for openSMILE features with MAESTRO (shown with superscript *), no reported average for these features.

| | DCASE | | MAESTRO | | Average |
|---|---|---|---|---|---|
| | Onset FMS | Error rate ↓ | Onset FMS | Onset w/ Offset FMS | Onset FMS |
| **HEAR baselines:** | | | | | |
| CREPE | 0.552 | 0.420 | **0.3910** | **0.15** | 0.472 |
| wav2vec2 | 0.670 | 0.320 | 0.0328 | 0.009 | 0.351 |
| **Models:** | | | | | |
| BYOL-A | 0.499 | 0.503 | 0.0028 | 0.00029 | 0.251 |
| BYOL-S | 0.650 | 0.356 | 0.0043 | 0.00048 | 0.327 |
| BYOL-S++ | 0.512 | 0.583 | 0.0457 | 0.01055 | 0.279 |
| Hybrid BYOL-S | 0.526 | 0.504 | 0.0067 | 0.00090 | 0.266 |
| BYOL-S/CvT | **0.891** | **0.152** | 0.0817 | 0.01488 | **0.486** |
| openSMILE (OS) only | 0.857 | 0.194 | -* | -* | n/a |
| BYOL-S/CvT + OS | 0.865 | 0.18 | -* | -* | n/a |
| Hybrid BYOL-S/CvT | 0.889 | 0.153 | 0.0746 | 0.01242 | 0.482 |

Table 9: Results on timestamp-based tasks using the hybrid version of BYOL-S/CvT. ↓ indicates lower is better.

| | DCASE | | MAESTRO | | Average |
|---|---|---|---|---|---|
| | Onset FMS | Error rate ↓ | Onset FMS | Onset w/ offset FMS | Onset FMS |
| **HEAR baselines:** | | | | | |
| CREPE | 0.552 | 0.420 | **0.3910** | **0.15** | 0.472 |
| wav2vec2 | 0.670 | 0.320 | 0.0328 | 0.009 | 0.351 |
| **Hybrid BYOL-S/CvT: Window/Hop size** | | | | | |
| 1s/50ms | **0.889** | 0.153 | 0.0746 | 0.01242 | 0.482 |
| 0.5s/50ms | 0.880 | **0.144** | 0.1739 | 0.04406 | **0.527** |

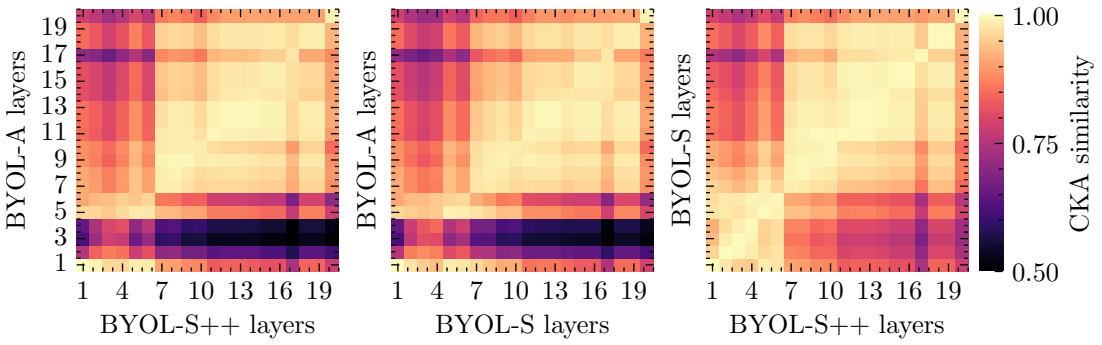


Figure 3: CKA-based model similarity analysis.

## 4. Discussion

### 4.1. Contribution of pre-training datasets: BYOL-A, BYOL-S and BYOL-S++

Table 5 shows that BYOL-S++ outperformed BYOL-A and BYOL-S in all speech-related tasks. This result is consistent with the fact that BYOL-S++ was trained on a larger and more diverse speech corpus, with both spontaneous and anechoic scripted speech. In addition, BYOL-S surpassed the other models in all environmental sound tasks as well as in most music-related tasks (Tables 6, 7). The latter observation comes as less intuitive: one could expect BYOL-A, a general-purpose audio representation, to perform best on these tasks, whereas BYOL-S is speech-specific. However, it is worth mentioning that the speech subset used for BYOL-S pre-training included music, instrumental sounds and noise in the background which might contribute to its ability to perform better on non-speech tasks.

On the other hand, BYOL-A performed significantly better than the other models in differentiating music from speech (GTZAN; Table 7), which could be due to being trained on all AudioSet ontologies and thus acquiring better discrimination of signal types.

Unsurprisingly, the HEAR baseline CREPE (Kim et al., 2018b) consistently outperformed the proposed approaches in the NSynth pitch discrimination task. Indeed, CREPE is a specialized pitch representation, rather than a general-purpose audio model. The performance gap might be explained by different pitch properties in speech and instrumental music. Since the proposed model is predominantly a speech feature extractor, it does not fully capture pitch representation in musical tasks. However, the performance gap seems to decrease for larger downstream dataset sizes (i.e., NSynth (50h); Table 7).

To gain further insight into model performance, we used the centered kernel alignment (CKA) method (Cortes et al., 2012; Kornblith et al., 2019) to assess the similarity between layers of the BYOL-A, BYOL-S, and BYOL-S++ representations (Figure 3). Following Raghu et al. (2021) and Subramanian (2021)[11], we first fed an unseen dataset (CREMA-D) to two different models (e.g., BYOL-A and BYOL-S++; Figure 3, left) to generate the activation maps for each layer. Using the CKA algorithm, we were then able to compute pairwise similarity scores between each layer using the Gram matrix of their activation maps, resulting in a "similarity matrix" between the two models.

11. https://github.com/AntixK/PyTorch-Model-Compare

Accordingly, Figure 3 shows that only the last convolution layers of BYOL-A shared a high CKA similarity with those of BYOL-S and BYOL-S++, constituting reasonable evidence that BYOL-A learned different features from the speech-specific frameworks. Conversely, high similarity was observed across all layers between BYOL-S and BYOL-S++. This observation is consistent with Tables 5-7, where the BYOL-S and BYOL-S++ achieved similar results on the HEAR benchmark. Thus, the lower similarity between the first layers of BYOL-A and BYOL-S-derived models could mean that these first layers have a detrimental influence on model performance. More generally, these results confirm that pre-training dataset selection is, unsurprisingly, critical to produce robust audio representations, especially as all models presented here shared the speech portion of AudioSet for pre-training.

### 4.2. Contribution of the pre-training window size

Table 5 depicts a trend towards higher performance in speech-related scene-based tasks for larger window sizes during pre-training. This influence of the training window size is not unreasonable since a typical utterance can span across several seconds. Conversely, for environmental sounds and music (comprising transient musical notes), Tables 6 and 7 reflect a reverse tendency, i.e., the smaller the window size the better in most tasks. This might be due to the fact that brief environmental sounds (r.g., gunfire sounds from the Gunshot triangulation dataset or glass breaking and dog barking in the ESC dataset) do not require a large context window to be identified, making the models that were pre-trained on a small context window a better fit. As window size optimization appears to be highly dependent on the dataset, developing training protocols to effectively capture context across multiple timescales should constitute for a crucial step to produce universal audio representations.

### 4.3. Comparison of encoder architectures: Resnetish-34, CLSTM, CvT

Among the alternative encoders to the default CNN-based encoder network in BYOL-S (Section 2.1), only BYOL-S/CvT outperformed the "vanilla" BYOL-S. For instance, BYOL-S/CvT was the best representation for music-related tasks (Table 7), and achieved similar performance to BYOL-S in speech-related tasks (Table 5). On the other hand, the default BYOL-S outperformed all other encoders in environmental sound-related tasks (Table 6). To explain such discrepancies, one could hypothesize that using the other encoders, CLSTM and ResNetish34, caused overfitting problems during pre-training due to their higher model capacity. This hypothesis is especially motivated by the fact that both frameworks had a remarkably lower pre-training error compared to BYOL-S and BYOL-S/CvT. Thus, one could argue that simple encoder models might be preferable for generating robust audio representations when using the BYOL-A paradigm. That being said, larger transformer-based models and trained on large-scale datasets tend to yield superior audio representations (Shor et al., 2022). However, due to their size, the application of such models is costly and necessitates a longer inference time. Here, we focused on comparatively smaller models ($< 30$M parameters), which can be trained and applied using manageable computing resources.

### 4.4. Hybrid versions of BYOL-S and BYOL-S/CvT

In this study, we proposed *hybrid* models of BYOL-S as an attempt to take advantage of DSP-based features and eventually increase feature interpretability for data-driven features.

In fact, the hybrid version of BYOL-S/CvT outperformed the other proposed models on most tasks of the HEAR benchmark (Tables 5-7). In particular, the benefit of this hybrid method was especially apparent when utilizing CvT as the encoder. Hence, adding DSP-based fixed features could be viewed as a good auxiliary task to support model pre-training. Consequently, at the end of the pre-training, the model should be able to strike the optimal trade-off between the pure-DSP and a fully data-driven, learned representation. The addition of fixed, non-trainable features could also improve training stability (i.e., preventing model collapse), which is known to be one of the issues associated with BYOL-style methods (usually mitigated by updating the target weights as a moving average of those of the online network) (Grill et al., 2020). While the supervision module (Figure 2) only consisted of fixed DSP-based features from openSMILE in this work, using other learnable, deep learning-based, feature extractors in the supervision module could constitute promising future work, where the two trainable systems could be even trained or fine-tuned in an iterative fashion. To explore the impact of the supervised and self-supervised components on hybrid model performance, we varied the weights $\alpha$ and $\beta$ given to their respective loss functions. Overall, the results of this tuning procedure seem to be dataset-dependent. In speech and environmental sound tasks (Tables 5 and 6), performance was maximal when $\alpha = \beta = 1$ and gradually decreased as the ratios become more unbalanced. In music-related tasks (Table 7), using $\alpha = \beta = 1$ yielded notably worse results than for other ratios.

To complement our experiments, we assessed the validity of the hybrid training protocol by comparing the hybrid version of BYOL-S/CvT with embeddings produced using only openSMILE features or the concatenation of the latter with BYOL-S/CvT embeddings. As shown in Tables 5-7, we found that the hybrid model remained the best representation, thus suggesting that the information learned during the hybrid training paradigm is more robust than a simple concatenation between features having different distributions.

### 4.5. Timestamp Embeddings

Table 8 showed that the CvT encoder and the hybrid models constituted on average our best performing models for timestamp-based tasks. It is noted that the performance of openSMILE features only, on the DCASE task, was improved by concatenating the latter features with BYOL-S/CvT, however, Hybrid BYOL-S/CvT still outperforms the simple concatenation between both features. While yielding marginal changes for the DCASE task (Table 9), decreasing the window size for the hybrid BYOL-S/CvT considerably improved the detection of note onsets in MAESTRO, in addition to note onset with offset frames. This could be because music notes tend to be short transient events, making small window lengths a better choice for music transcription tasks. That being said, our current implementation for timestamp embeddings remained a relatively simple extension of that for scene embeddings. Hence, it is likely that a specialized encoder, or an entire different pre-training protocol, could yield substantially better results. That is why HEAR's implementation of CREPE, specifically designed as pitch representation, as well as other CREPE-derived models (Turian et al., 2022), consistently topped the leaderboard on musical tasks.

## 5. Conclusion

In this paper, we present our submission for the 2021 NeurIPS HEAR challenge, a benchmark to evaluate audio representations on 16 downstream tasks. Our submission was based on BYOL-S, a re-trained version of BYOL-A on a speech subset of AudioSet. Although the model was originally designed as a speech-specific representation, we showed that it can also produce suitable performance in other tasks involving environmental sounds and music. Additionally, we carried out several experiments to delve deeper into the model components that contribute to generating audio representations. The experiments included trying different pre-training datasets, encoder architectures, pre-training audio windows, a hybrid pre-training protocol, and finally, the selection of hyperparameters to optimize timestamp embeddings. While the choice of pre-training audio size was generally subject-dependent (e.g., shorter and longer windows for environmental sounds and speech, respectively), opting for CvT-based encoding and a hybrid training protocol using openSMILE features yielded more robust results. Consequently, we observed that the hybrid version of BYOL-S/CvT, i.e., with CvT encoding and hybrid pre-training, outperformed our original HEAR challenge submission (BYOL-S), constituting our best audio representation with respect to the HEAR benchmark. Thus, combining original self-supervised pre-training paradigm with DSP-based handcrafted features used in the loss function of the hybrid model helped produce a more robust audio representation. This finding further validates the benefit of using ensemble embeddings obtained from several models for general-purpose audio representation (Turian et al., 2022). In particular, involving a considerably simple set of fixed features during training can substantially improve DNN-based audio representation learning.

## References

Akshay Anantapadmanabhan, Ashwin Bellur, and Hema A Murthy. Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185, 2013.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020.

Pierre Beckmann, Mikolaj Kegler, and Milos Cernak. Word-level embeddings for cross-task transfer learning in speech processing. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 446–450, 2021.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

Seth Cooper and Steven Shaw. Gunshots recorded in an open field using iPod touch devices. http://datadryad.org/stash/dataset/doi:10.5061/dryad.wm37pvmkc, 2020. Accessed: 2022-03-09.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning (ICML)*, pages 1068–1077, 2017.

Florian Eyben, Martin Wöllmer, and Björn Schuller. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, 2010.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Jort F Gemmeke, Daniel PW 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)*, pages 776–780, 2017.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

TD Griffiths. Human complex sound analysis. *Clinical Science*, 96(3):231–234, 1999.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21271–21284, 2020.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations (ICLR)*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan. Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, volume 1, pages 148–152, 2018a.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A Convolutional Representation for Pitch Estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018b.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, pages 3519–3529, 2019.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Zhu Liu, Yao Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI signal processing systems for signal, image and video technology*, 20(1):61–79, 1998.

A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):379–393, 2018.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT Press, 2022.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for Audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

Inês Nolasco, Alessandro Terenzi, Stefania Cecchi, Simone Orcioni, Helen L Bear, and Emmanouil Benetos. Audio-based identification of beehive states. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8256–8260, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

Vishal Passricha and Rajesh Kumar Aggarwal. A hybrid of deep cnn and bidirectional lstm for automatic speech recognition. *Journal of Intelligent Systems*, 29(1):1261–1274, 2020.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proc. ACM Multimedia (MM)*, pages 1015–1018, 2015.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879, 2021.

Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. SERAB: A multi-lingual benchmark for speech emotion recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7697–7701, 2022.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013*, pages 148–152, 2013.

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, sincerity & native language. In *Proc. Interspeech 2016*, pages 2001–2005, 2016.

Björn Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In *Proc. Interspeech 2020*, pages 2042–2046, 2020.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. Towards Learning a Universal Non-Semantic Representation of Speech. In *Proc. Interspeech 2020*, pages 140–144, 2020.

Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. Universal paralinguistic speech representations using self-supervised conformers. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3169–3173, 2022.

Fabian-Robert Stöter, Soumitro Chakrabarty, Emanuël Habets, and Bernd Edler. Libricount a dataset for speaker count estimation (v1.0.0) [Data set]. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Anand Subramanian. torch_cka. https://github.com/AntixK/PyTorch-Model-Compare, 2021. Accessed: 2022-02-14.

Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. Beijing opera percussion instrument dataset. https://doi.org/10.5281/zenodo.1285212, 2014. Accessed: 2022-03-09.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. HEAR 2021: Holistic Evaluation of Audio Representations. *arXiv preprint arXiv:2203.03022*, 2022.