ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746490

CONFORMER-BASED SELF-SUPERVISED LEARNING FOR NON-SPEECH AUDIO TASKS

Sangeeta Srivastava[⋆§], *Yun Wang*[†], *Andros Tjandra*[†], *Anurag Kumar*[‡], *Chunxi Liu*[†]
Kritika Singh[†], *Yatharth Saraf*[†]

[⋆]The Ohio State University, USA [†]Meta AI, USA
[‡]Reality Labs Research, USA

srivastava.206@osu.edu, {yunwang,androstj,anuragkr,chunxiliu,skritika,ysaraf}@fb.com

ABSTRACT

Representation learning from unlabeled data has been of major interest in artificial intelligence research. While self-supervised speech representation learning has been popular in the speech research community, very few works have comprehensively analyzed audio representation learning for non-speech audio tasks. In this paper, we propose a self-supervised audio representation learning method and apply it to a variety of downstream non-speech audio tasks. We combine the well-known wav2vec 2.0 framework, which has shown success in self-supervised learning for speech tasks, with parameter-efficient conformer architectures. Our self-supervised pre-training can reduce the need for labeled data by two-thirds. On the AudioSet benchmark, we achieve a mean average precision (mAP) score of 0.415, which is a new state-of-the-art on this dataset through audio-only self-supervised learning. Our fine-tuned conformers also surpass or match the performance of previous systems pre-trained in a supervised way on several downstream tasks. We further discuss the important design considerations for both pre-training and fine-tuning.

Index Terms— Self-supervised learning, representation learning, conformer, sound events, wav2vec.

1. INTRODUCTION

In recent years, self-supervised learning (SSL) has proven to be an effective method for representation learning without the need for labeled supervision, not only in the language [1] and visual [2] domains but also in the audio domain. In the audio domain, SSL for speech tasks have shown extremely promising results [3], but SSL for non-speech audio tasks has been explored to a lesser extent. Examples of SSL for non-speech tasks include multimodal self-supervised learning [4] and audio captioning [5], but an extensive evaluation of SSL for non-speech audio-only tasks remains to be seen.

A number of self-supervised representation learning frameworks have been proposed, of which contrastive learning based methods have been found to be particularly effective in both image and speech domains. It aims to push semantically comparable samples closer together and dissimilar samples apart in the feature space. Wav2vec 2.0 [3] is one such contrastive learning framework which have been demonstrated to be highly efficient in speech-related tasks [6, 7].

Wav2vec 2.0 uses transformers to build contextualized representations of the speech sequence, but transformers can only simulate global dependencies. While convolutional networks (CNN) can capture local associations, obtaining global information from the model would necessitate the use of additional layers or parameters. Conformers [8], on the other hand, captures both local and global dependencies while requiring fewer parameters as compared to convolutional models. Prior works [9, 10] have shown that conformers can outperform transformers and CNN models on both speech and non-speech tasks.

In this work, we combine the conformer architecture with contrastive learning to learn representations for non-speech audio in a self-supervised manner. We replace the transformer in wav2vec 2.0 with a conformer in order to capture both global and local information in the audio signal. Unlike prior self-supervised works on speech representation learning (*e.g.* [3]) which directly learn from raw audio waveforms, we use logmel spectrograms as the fundamental representation of the audio signal. Besides their proven efficacy in different audio tasks [11], logmels are much more compact and lead to better memory and compute efficiency. The network is pre-trained on a large-scale unlabeled dataset of ∼67,000 hours to learn audio representations, and then fine-tuned on several audio tasks to show the generalization capabilities of the proposed SSL framework.

The contributions of this work are the following: **(1)** We propose a conformer-based SSL framework for general-purpose audio representation learning applicable to many audio tasks, which can reduce the need for labeled data by two-thirds; **(2)** On the AudioSet [12] benchmark, our fine-tuned representations achieve a new state-of-the-art (SOTA) mean average precision (mAP) of 0.415 for self-supervised learning with only audio, outperforming the previous best score of 0.329; **(3)** Our method can surpass or match the performance of previous systems pre-trained in a supervised way on 3 out of 4 classification tasks: acoustic scenes, music and human actions; **(4)** We show the effect of the design choices during pre-training and identify the main parameters to tune to avoid overfitting during fine-tuning.

2. SELF-SUPERVISED CONFORMER TRAINING

2.1. Upstream Data

The data for pre-training was selected from de-identified audio tracks of publicly shared Facebook user videos. We ran an in-house version of TALNet [13] on the audio tracks of 440 million user videos up to five minutes long, and obtained estimated probabilities of the 527 types of acoustic events in the AudioSet ontology. Then, for each event type, we selected 9,500 videos with the highest probabilities (some videos might be selected for multiple event types). The resultant dataset contains ∼3.9M videos totaling ∼67k hours. We expect that this large dataset captures enough relevant acoustic phenomena and is sufficiently diverse for different non-speech audio tasks.

2.2. Architecture

Frontend. We use 64-dimensional logmel spectrograms extracted with a window size of 64 ms and a hop size of 20 ms as the input X.

Encoder. As shown in Figure 1, the logmel spectrograms are fed into two types of encoders: feature and context encoders. In the feature encoder, the time stacking layer stacks every $R = 4$ consecutive

§ This work was done during an internship at Facebook.

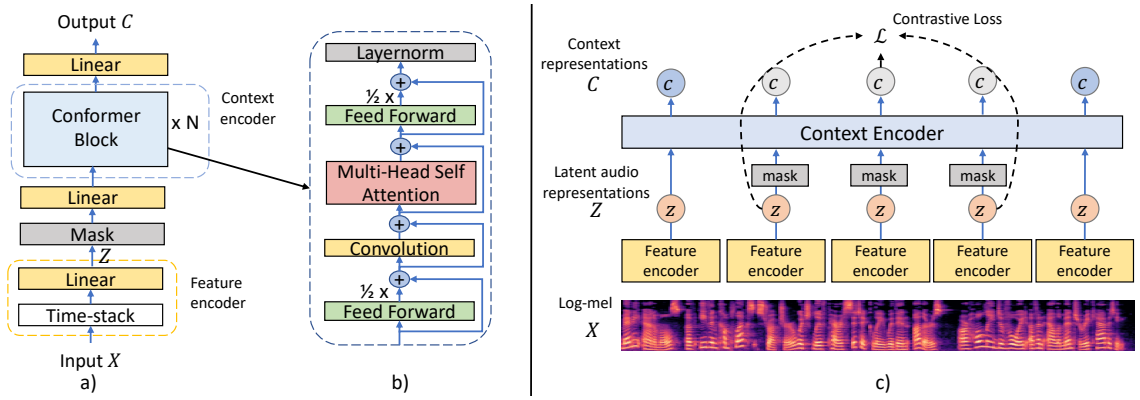

Fig. 1: *a) Logmel wav2vec 2.0 conformer architecture. b) A conformer block. c) An illustration of how wav2vec 2.0 generates context representation and solves the contrastive task.*

frames into a single frame; then a linear layer converts the stacked spectrogram into latent representation vectors $Z = [z_1, ..., z_{T/R}]$.

Before feeding the latent representation Z to the context encoder, we mask a proportion of it, similar to the original wav2vec 2.0 framework. The context encoder consists of a linear layer, multiple conformer blocks [8] and another linear layer at the end. We experiment with two context encoder architectures of differing sizes:

(i) **conformer small** (*cf_S*): 12 conformer blocks with 256-D encoder embeddings and 8 attention heads (18.4M parameters).

(ii) **conformer large** (*cf_L*): 12 conformer blocks with 768-D encoder embeddings and 12 attention heads (88.1M parameters).

Both encoders use a feed-forward network (FFN) dimension of 1024, dropout 0.1, kernel size 31 in the first conformer block and 15 for the rest. We further explore different design choices in Section 5.1.2. Additionally, following [14], we remove the original relative positional encoding, and reuse the existing convolution module for positional encoding by swapping the order of the convolution and multi-head self-attention modules, which speeds up both training and inference.

Contrastive Loss. We learn audio representations by solving a contrastive task: identify the true latent representation z_t for a masked time step t within a set of $K+1$ candidates $\tilde{Z} = \{\tilde{z}_1, \ldots, \tilde{z}_{K+1}\}$, which includes z_t and K distractors. Similar to the original wav2vec 2.0, the distractors are uniformly sampled from other masked time steps of the same audio snippet. Given the context network output c_t centered over a masked time step t, the contrastive loss is defined as:

$$L = -\log \frac{\exp(\text{sim}(c_t, z_t))}{\sum_{\tilde{z} \in \tilde{Z}} \exp(\text{sim}(c_t, \tilde{z}))} \quad (1)$$

where $\text{sim}(\mathbf{x}, \mathbf{y})$ is the cosine similarity.

2.3. Hyper-parameters

For all pre-training experiments, we use the Adam [15] optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$. We warm up the learning rate for the first 10k updates linearly to a peak of 3×10^{-4}, and then decay it linearly to zero until 300k updates. We regularize the model using weight decay with a decay factor of 0.01. For the contrastive loss we use $K = 100$ distractors and mask 30% of the latent representations Z.

3. DOWNSTREAM TASKS

We evaluate the generalizability of the learnt representations on sound event detection (SED) and other non-speech audio tasks. Within SED, we evaluate on the AudioSet dataset separately in Sec. 5, because it is much larger than the data used in all the other downstream tasks, and has been used in several priors works as well.

3.1. Sound Event Classification

AudioSet [12]: AudioSet contains $\sim$ 2 million 10-second audio clips from YouTube videos, labeled using an ontology of 527 sound classes. Each clip may have multiple labels. The *Balanced* training set of AudioSet is a subset with at least 59 clips per class. The balanced training, full training, and evaluation sets contain 22k, 2M, and 20k samples, respectively. For AudioSet, we use the larger model (*cf_L*).

ESC-50 [16]: ESC-50 consists of 2,000 5-seconds audio clips belonging to 50 environmental sound categories. It contains 40 samples for each category, and comes divided into 5 folds for cross validation.

FSDKaggle2019 [17]: FSDKaggle2019 is a multi-label dataset consisting of 29,266 audio files annotated with 80 labels form the AudioSet ontology. It further consists of a *curated* set and a *noisy* set: the former is annotated by humans but the labels may be incomplete; the latter contains many wrong labels.

3.2. Other Audio Tasks: Scenes, Music, and Human Actions

Acoustic scene classification: We use the dataset from Task 1a of the 2019 DCASE challenge [18]. It consists of 9,185 and 4,185 segments in the training and the test sets respectively, belonging to 10 acoustic scene classes such as "airport", "park", and "metro station".

Music tagging: This is a multi-label classification problem where each recording can have one genre label and multiple instrument labels. For this task, we use the MagnaTagATune [19] dataset which consists of 25,863 music clips, each clip lasting 29 seconds. We follow the most common 12:1:3 split of training, validation, and evaluation data, and only use the 50 most popular tags out of the 188.

Human action classification: This task involves recognizing human actions such as "dancing" and "playing guitar" in video recordings. The Kinetics700 dataset [20] is a collection of 650k 10-second video clips covering 700 human action classes. This problem has primarily been addressed using images in a uni- [21] or multimodal [22] approach; so far there is only one audio-only solution [11].

4. FINE-TUNING

After the self-supervised pre-training, the conformer model can generate an embedding vector (the context representation) for each frame of the input audio. The downstream tasks, however, require predictions over the entire input audio. For AudioSet, we first add a linear classification layer with the sigmoid activation to predict the

Model	Balanced (20k)	60k	200k	600k	Full (∼2M)
	1%	3%	10%	30%	100%
From scratch	0.138	-	-	-	0.366
Pre-train + fine-tune	**0.276**	0.305	0.337	0.356	**0.415**

Table 1: *Performance (measured in mAP) of the cf_L model on the test set of AudioSet, with and without pre-training, and with different amounts of fine-tuning data.*

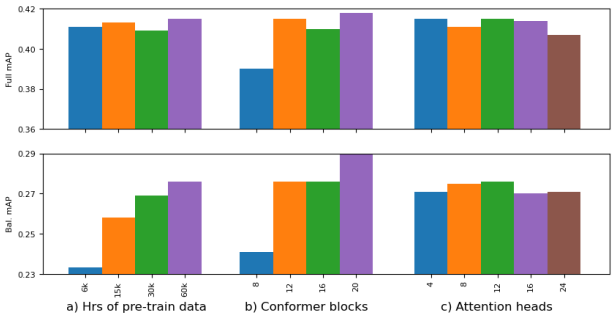

Fig. 2: *Effect of the amount of pre-training data, and the number of conformer blocks and attention heads (Full & Balanced AudioSet).*

frame-level probabilities of each event type, then add a linear softmax pooling layer [13] to pool them into global probabilities. For all other downstream tasks, we first average-pool the embeddings across all the frames, then stack a linear classification layer to make predictions.

During fine-tuning, we allow the weights of the conformer to update. We find this is critical to make the SSL schema work. This is different from some existing works on transfer learning, where a shallow classifier is stacked on top of an encoder with frozen weights.

4.1. Data Processing

To deal with the highly skewed label distribution in Full AudioSet, we apply data balancing to ensure that the training sees different sound classes at approximately equal frequencies. We also apply data augmentations in the following order to all downstream tasks:

Temporal jittering: Prior to extracting logmel features, we shift the waveform by a random amount between -200 and 200 samples. The maximum shift (200 samples) is equal to half the frame length.

SpecAugment [23]: For each 10 s video, we mask out the logmel features of one random temporal interval up to 2 s. We do not mask out any frequency bins because it does not work well with Mixup.

Mixup [24]: For a minibatch of n videos, with logmel features denoted by $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we shuffle the instances to get $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_n\}$, and then mix up the two batches according to

$$\mathbf{x}''_i = \alpha_i \mathbf{x}_i + (1 - \alpha_i)\mathbf{x}'_i, \quad i = 1, \ldots, n \tag{2}$$

The mixing coefficient α_i is sampled from the beta distribution $B(0.5, 0.5)$, and it is enforced that $\alpha_i \geq 1 - \alpha_i$. The labels of $\mathbf{x}''_i$ inherits those of $\mathbf{x}_i$; we do not mix them with the labels of $\mathbf{x}'_i$.

4.2. Consistency Loss

In addition to the binary cross-entropy (BCE) loss, we employ a consistency loss as a regularizer for training. For each minibatch, we apply the data augmentation twice with different random numbers, and compute the symmetric KL divergence between the model's predicted event probabilities on the two versions of data. This KL divergence is multiplied by 2 and added to the BCE loss.

Pre-training	Model	Input	AudioSet pre-train?	Fine-tune?	Test mAP
Audio-only Self-supervised	C^3 [25]	l	✓		0.206
	Triplet [26]	l	✓		0.259
	CPC [27]	w	✓		0.277
	Multi-Format [28]	w + l	✓		0.329
	Separation-based [29]	l	✓		0.326
	Ours (*cf_L*)	**l**	✗	✓	**0.411**
Multimodal Self-supervised	C^3 [25]	l + v	✓		0.285
	MMV [30]	l + v + t	✓		0.309
	VATT [4]	l + v + t	✓	✓	0.394
	Multimodal COLA [31]	w + l + v	✓		0.424
Supervised (ImageNet)	PSLA (w/o ensemble) [32]	w + l		✓	0.444
	AST (w/o ensemble) [33]	l		✓	0.459
No pre-training (From scratch)	TALNet [13]	l			0.383
	WEANet-SUSTAIN [34]	l			0.398
	PANNs [35]	w + l			0.431
	ERANN [36]	l			0.450

Table 2: ***Performance comparison with prior works on Full AudioSet.*** *Input forms include waveforms (w), logmel (l), video (v), and text (t). We label systems that are pre-trained on AudioSet, or update the encoder weights during fine-tuning. Higher mAP is better.*

4.3. Hyperparameter Tuning

We find that the fine-tuning process is prone to overfitting, but it can be avoided by carefully tuning the learning rate, the learning rate schedule, and the batch size. We utilize a three-stage learning rate scheduler [23] for all downstream datasets. The three stages are: warmup, hold, and exponential decay, and they typically last for 30%, 30%, and 40% of all the updates. While high-resource datasets perform best with larger batches, smaller datasets prefer smaller batches. For example, we use a batch size of 640 and 64 for Full AudioSet and MagnaTagATune, respectively. We also regularize the model more for low-resource datasets by adding dropout to the output layer of the pre-trained conformer. We sweep different combinations of the parameters to find the one that optimizes validation performance.

5. RESULTS

5.1. Evaluation on AudioSet

5.1.1. *Effect of Pre-training*

First, we evaluate the merit of pre-training in our proposed SSL schema in Table 1. On both the Balanced and the Full training sets, we compare the performance of the *cf_L* model trained from scratch vs pre-trained then fine-tuned. The model trained from scratch performs **50%** and **12%** worse than the pre-trained and fine-tuned version on Balanced and Full AudioSet, respectively. This emphasizes the necessity of pre-training, especially for smaller downstream datasets.

The merit of pre-training can also be quantified by the reduced need for labeled data. By varying the amount of fine-tuning data from 20k to 1.9M audio clips, we find that pre-training and fine-tuning with 600k clips approximately matches the performance of training from scratch with 1.9M clips. That is, pre-training with ∼67k hours of unlabeled data reduces the need for labeled data by **two-thirds**.

5.1.2. *Ablation Studies*

We perform detailed ablation studies to investigate the effect of the amount of pre-training data and the design choices regarding the model structure, on both the Full and and Balanced training sets.

Amount of pre-training data. As shown in Fig. 2a, the performance on Full AudioSet stays relatively unchanged as we vary the amount of pre-training data from 6k to 60k hours. In other words, we can reduce the need for labeled data by two thirds even with only

Task	Dataset		# Training Samples	Metric	SS Conformer		Prior Works		
					cf_S	cf_L	SS + shallow	S + shallow	S + fine-tune
Sound Events	ESC50		1,600	Accuracy	80.7	88.0	86.3 [28]†	94.1 [11]	94.7 [35]
	FSDKaggle	curated	4,970	lwlrap	50.3	61.1	✗	72.8 [11]	✗
		noisy	19,185		33.2	40.1	✗	51.0 [11]	✗
Acoustic Scenes	DCASE2019 Task 1a		9,185	Accuracy	72.4	76.1	✗	68.0 [11]	✗
Music Tagging	MagnaTagATune		15,247	MAUC	90.2	91.2	89.3 [37]	91.5 [11]	✗
Human Actions	Kinetics700		504,443	Top-1 Accuracy	20.4	23.5	✗	18.0 [11]	✗

Table 3: *Performance of both versions of our conformer on downstream tasks, compared with prior works using either self-supervised (SS) or supervised (S) pre-training. Most prior works use shallow classifiers; only [35] uses fine-tuning as we do. All metrics are the higher, the better.* †*Performance using logmel representations only. Combining waveform and logmel features produces a higher accuracy of 90.5%.*

6k hours of unlabeled data. On Balanced AudioSet, however, the performance increases significantly and monotonically with every bit of extra pre-training data. This demonstrates that a sufficient amount of unlabeled data can make up for the scarcity of labeled data.

Number of conformer blocks. We vary the depth of the model from 8 conformer blocks (58.9M parameters) to 20 conformer blocks (146.6M parameters), keeping all other hyperparameters identical to the large model. Fig. 2b shows that the largest performance gain occurs when going from 8 to 12 blocks. Beyond that, the improvement is minimal (if any) for Full AudioSet, but still moderate for Balanced AudioSet from 16 to 20 blocks.

Number of attention heads. Attention heads allow varied levels of focus on different parts of the sequence, producing better predictions than a single weighted average. We vary the number of attention heads from 4 to 24 in our large model, using the same number of heads in all layers. As shown in Fig. 2c, there is a slight improvement from 4 to 12 heads, but the performance drops once the number of attention heads exceeds 12 for both Full and Balanced AudioSet.

In general, the pre-training data size and design choices have a larger impact when labeled data for the downstream task is limited.

5.1.3. Comparison with Prior Works

Table 2 lists the Full AudioSet test performance of our system and some prior works. Depending on the type of data used for pre-training, we divide the works into four categories: self-supervised pre-training with audio only, self-supervised pre-training with multimodal data, supervised pre-training, and no pre-training. Not all comparisons are apple to apple: some works pre-train on AudioSet, and others (including ours) fine-tune the weights of the entire network, both of which may give them an advantage. Still, we report our performance with ∼6k hours of pre-training data without any augmentation, in order to match previous works that pre-train on AudioSet.

We give a quick overview of the previous works using audio-only self-supervised pre-training. [27] uses contrastive predictive coding (CPC) to single out the representation of a future step at a given distance, but also adds a nonlinear learnable similarity metric and adversarial perturbation. All the others [25, 26, 28, 29] learn encoders such that a pair of audio clips sampled within a certain temporal proximity have similar representations; the difference lies in how the pairs are presented to the encoder, and the loss function. [25] and [26] present both clips as spectrograms, and use the BCE loss and triplet loss, respectively. [28] presents one clip as a waveform and the other as a spectrogram, and uses a one-vs-many contrastive loss. [29] separates a mixed signal into two channels and forms a pair using one of the channels and the original signal, and combines the BCE and the one-vs-many contrastive losses.

The current state-of-the-art performance on AudioSet using audio-only SSL is held by [28], with an mAP of 0.329. Our method outperforms it by a huge **25%** relative, and achieves an mAP of 0.411. However, it should be pointed out that we update encoder weights during fine-tuning, while all the existing works in this category, including [28], use a shallow classifier upon frozen encoder weights.

On the multimodal self-supervised learning front, most methods use audio and visual signals to learn relationships between them [25, 30, 4, 31]. In addition to audio and video, some works [30, 4] also make use of textual information, but perform worse than our framework with only audio. Our audio-only learning method is competitive even to the best multimodal SSL work on AudioSet [31], inferior by only 3% relative.

Supervised pre-training on ImageNet [32, 33] outperforms our proposed SSL framework, but it requires a large amount of labeled data for pre-training. However, we notice some systems trained from scratch [35, 36] achieve exceedingly high performance. This may indicate that CRNNs may be better suited for SED than conformers.

5.2. Evaluation on Other Downstream Tasks

Table 3 summarizes the results for all the other downstream tasks. We compare the performance of both our conformer models (*cf_S* and *cf_L*) with prior works of transfer learning using either self-supervised or supervised pre-training. Most of these prior works use shallow classifiers; only [34] fine-tunes the entire network as we do. To our knowledge, no self-supervised prior work exists for FSDKaggle2019, DCASE2019, or Kinetics700.

For ESC50 and MagnaTagATune, we outperform baselines of self-supervised pre-training with our *cf_L* architecture. For music tagging, we nearly match the baseline system [11] pre-trained in a supervised fashion; for the classification of acoustic scenes and human actions, *cf_L* (*cf_S*) outperforms supervised pre-training baselines by 11.9% (6.5%) and 30.5% (13.3%), respectively. However, if we compare against supervised pre-training baselines for SED tasks, our best *cf_L* architecture performs 7.6% (ESC50) and 19.1% (FSDKaggle curated) worse. This significant disparity can be attributed to the fact that the baseline systems were pre-trained on Full AudioSet, which has a substantial overlap in the label space with both ESC50 and FSDKaggle, providing the baseline systems with an edge.

6. CONCLUSION

In this paper, we have proposed and evaluated a conformer-based self-supervised framework for learning representations for non-speech audio. The self-supervised pre-training can reduce the need for labeled data by two-thirds. On the widely known AudioSet, our framework produces an mAP of 0.415, outperforming the audio-only self-supervised learning SOTA of 0.329; it also achieves performance comparable to supervised pre-trained baselines on several other downstream non-speech tasks. Our ablation studies shows the significance of critical design choices such as the model depth.

7. REFERENCES

[1] J. Devlin, M.-W. Chang, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[2] T. Chen, S. Kornblith, et al., "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[3] A. Baevski, Y. Zhou, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[4] H. Akbari, L. Yuan, et al., "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," *arXiv preprint arXiv:2104.11178*, 2021.

[5] X. Liu, Q. Huang, X. Mei, T. Ko, H. L. Tang, M. D. Plumbley, and W. Wang, "CL4AC: A contrastive loss for audio captioning," *arXiv preprint arXiv:2107.09990*, 2021.

[6] A. Tjandra, D. G. Choudhury, et al., "Improved language identification through cross-lingual self-supervised learning," *arXiv preprint arXiv:2107.04082*, 2021.

[7] Z. Fan, M. Li, et al., "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[8] A. Gulati, J. Qin, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*. ISCA, 2020, pp. 5036–5040.

[9] Y. Zhang, J. Qin, et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[10] K. Miyazaki, T. Komatsu, et al., "Convolution-augmented transformer for semisupervised sound event detection," in *DCASE*, 2020, pp. 100–104.

[11] A. Kumar, Y. Wang, et al., "Do sound event representations generalize to other audio tasks? A case study in audio transfer learning," in *Interspeech*. ISCA, 2021, pp. 1214–1218.

[12] J. F. Gemmeke, D. P. W. Ellis, et al., "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

[13] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP*. IEEE, 2019, pp. 31–35.

[14] B. Li, A. Gulati, et al., "A better and faster end-to-end model for streaming ASR," in *ICASSP*. IEEE, 2021, pp. 5634–5638.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[16] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM Int. Conf. on Multimedia*, 2015, pp. 1015–1018.

[17] E. Fonseca, M. Plakal, et al., "Audio tagging with noisy labels and minimal supervision," *arXiv preprint arXiv:1906.02975*, 2019.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[19] E. Law, K. West, et al., "Evaluation of algorithms using games: The case of music tagging," in *ISMIR*. Citeseer, 2009, pp. 387–392.

[20] W. Kay, J. Carreira, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[21] C. Feichtenhofer, H. Fan, et al., "A large-scale study on unsupervised spatiotemporal representation learning," in *CVPR*, 2021, pp. 3299–3309.

[22] A. Radford, J. W. Kim, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[23] D. S. Park, W. Chan, et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*. ISCA, 2019.

[24] H. Zhang, M. Cisse, et al., "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[25] A. Jansen, D. P. W. Ellis, et al., "Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision," in *ICASSP*. IEEE, 2020, pp. 121–125.

[26] A. Jansen, M. Plakal, et al., "Unsupervised learning of semantic audio representations," in *ICASSP*. IEEE, 2018, pp. 126–130.

[27] L. Wang, K. Kawakami, and A. van den Oord, "Contrastive predictive coding of audio with an adversary," in *Interspeech*. ISCA, 2020, pp. 826–830.

[28] L. Wang and A. v. d. Oord, "Multi-format contrastive learning of audio representations," *arXiv preprint arXiv:2103.06508*, 2021.

[29] E. Fonseca, A. Jansen, et al., "Self-supervised learning from automatically separated sound scenes," *arXiv preprint arXiv:2105.02132*, 2021.

[30] J.-B. Alayrac, A. Recasens, et al., "Self-supervised multimodal versatile networks," *NeurIPS*, vol. 2, no. 6, pp. 7, 2020.

[31] L. Wang, P. Luc, et al., "Multimodal self-supervised learning of general audio representations," *arXiv preprint arXiv:2104.12807*, 2021.

[32] Y. Gong, Y.-A. Chung, and J. Glass, "PSLA: Improving audio event classification with pretraining, sampling, labeling, and aggregation," *arXiv preprint arXiv:2102.01243*, 2021.

[33] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[34] A. Kumar and V. Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *ICML*. PMLR, 2020, pp. 5447–5457.

[35] Q. Kong, Y. Cao, et al., "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[36] S. Verbitskiy and V. Vyshegorodtsev, "ERANNs: Efficient residual audio neural networks for audio pattern recognition," *arXiv preprint arXiv:2106.01621*, 2021.

[37] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," *arXiv preprint arXiv:2103.09410*, 2021.