

SEAR: Semantically-grounded Audio Representations

Rajat Hebbar
University of Southern California
Los Angeles, CA, USA
rajatheb@usc.edu

Digbalay Bose
University of Southern California
Los Angeles, CA, USA
dbose@usc.edu

Shrikanth Narayanan
University of Southern California
Los Angeles, CA, USA
shri@ee.usc.edu

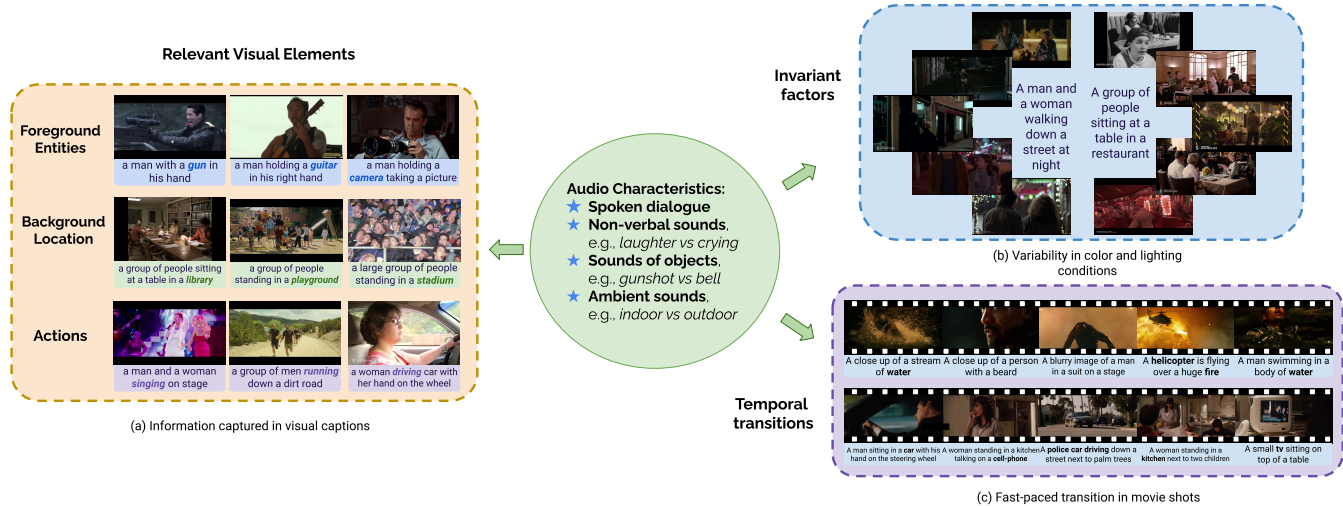


Figure 1: Relationship between audio and visual elements in movies and how (a) visual captions can help tease out *acoustically-relevant* factors while (b) being invariant to *acoustically-irrelevant* factors. (c) Captions are also a concise yet efficient way to aggregate information across fast-paced transition in visual frames as often seen in movies.

ABSTRACT

Audio supports visual story-telling in movies through the use of different sounds. These sounds are often tied to different visual elements, including foreground entities, the interactions between them as well as background context. Visual captions provide a condensed view of an image, providing a natural language description of entities and the relationships between them. In this work, we utilize visual captions to semantically ground audio representations in a self-supervised setup. We leverage state-of-the-art vision-language models to augment movie datasets with visual captions at scale to the order of 9.6M captions to learn audio representations from over 2500 hours of movie data. We evaluate the utility of the learned representations and show state-of-the-art performance on two movie understanding tasks, genre and speaking-style classification, outperforming video based methods and audio baselines. Finally, we show that the learned model can be transferred in a zero-shot manner through application in both movie understanding tasks and general action recognition.

CCS CONCEPTS

• Computing methodologies → Machine learning; • Applied computing → Media arts.

KEYWORDS

Self-supervised learning, Movie understanding, Cross-modal supervision, Audio Representation Learning

ACM Reference Format:

Rajat Hebbar, Digbalay Bose, and Shrikanth Narayanan. 2023. SEAR: Semantically-grounded Audio Representations. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3581783.3612592>

1 INTRODUCTION

Computational media understanding [52] has gained tremendous interest in the past few years with the rapid proliferation of multimedia content including the curation of large-scale multimedia databases from the web [3, 58]. Computational analyses of media not only can enable automatic content characterization e.g., genre, but offer insights into finer narrative elements like interactions between the characters and any underlying bias in their portrayals. Movies, in particular, are a rich source of audiovisual content – depicting a wide variety of scenes, making it conducive for self-supervised representation learning [28].



This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0108-5/23/10.
<https://doi.org/10.1145/3581783.3612592>

Sound in movies is carefully designed to set the mood, convey emotion and enhance the visual imagery shown on the screen. Different elements of movie sound including background soundtrack, sound effects and ambient sounds are cohesively crafted to augment the spoken dialogue and visual imagery to depict a complex narrative structure [11]. In this work, we aim to leverage this inherent multimodal structure in movies to learn robust audio representations in a self-supervised manner.

Cross-modal supervision in multimedia using contrastive learning has been catalyzed by the availability of large-scale datasets [4]. Vision-language (VLN) models [7, 46, 47] have shown enormous potential in diverse sets of discriminative and generative learning tasks. These models are usually pretrained on weakly aligned data curated from the web, thus eliminating the need for manual annotation. During the pretraining phase, such models can exploit cross-modal correspondence to learn generalizable representations that can be used for a wide array of downstream tasks including classification, retrieval, question-answering, and generative modeling. Potentially, it is possible learn rich audio representations in a similar fashion using manually-annotated audio captions describing the sounds occurring within them [13, 15, 29]. However, such datasets do not offer the scale of data often required for pretraining.

In this work, we propose to use *visual captions* to learn rich audio representations, leveraging large-scale vision-language models to extract visual captions for movies at scale. As shown in Fig. 1, visual captions can serve as comprehensive, yet condensed natural language descriptions of raw image frames in a movie-scene. High-level audio concepts such as audio events and acoustic scenes are strongly tied to foreground visual entities such as objects and persons as well as the ambient visual context. We use these visual captions to learn a semantically-aligned acoustic representation using contrastive methods.

To this end, we outline three main contributions of our work:

- 1) We propose a method to learn audio representations in movies by leveraging cross-modal supervision. We do so by augmenting existing movie datasets with captions at scale to distill semantic information and explore novel ways of pooling contextual information across shots.
- 2) We show the first application of audio representation learning for movie understanding applications. Our method using only audio outperforms SotA video-based methods on two out of three movie understanding benchmarks.
- 3) Finally, we leverage the use of lexical embeddings for alignment by showing zero-shot capability of our method in both movie understanding tasks and generic action recognition datasets.

The rest of the paper is organized as follows: We provide background on related work in self-supervision and movie understanding in Sec. 2. We then discuss the motivation of our approach and the methodology adopted in learning semantically-grounded audio representations (SEAR) in Sec. 4. We show the utility of these representations on several movie understanding and general classification tasks, which we discuss in Sec. 4.3. We also explore zero-shot capabilities of the learned representations and show both quantitative and qualitative analysis in Sec 5. Finally, we conclude in Sec. 6 by briefly discussing the limitations and fairness considerations of our work and identify potential extension of this work.

2 RELATED WORK

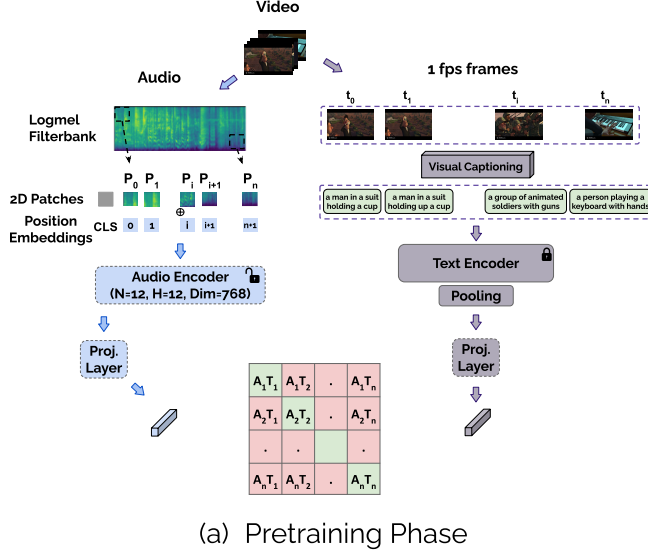
Self-Supervision for Audio Classification: Transformer architectures [18, 30] have shown tremendous success in audio classification tasks such as Audio Event Detection (AED). Recently, several self-supervised representation learning methods have been proposed and shown to generalize well for audio classification. A simple contrastive formulation involves using audio segments within an audio clip as positive samples and those across different audio clips as negative samples [50]. Another commonly used approach is to leverage augmented versions of an audio clip as positive samples, within a contrastive [54] or non-contrastive [43] framework. More recently, inspired by success in computer vision, masked reconstruction of audio features was proposed as a proxy task to learn rich representations [19]. Authors show that augmenting standard MSE-Loss with a patch-level contrastive loss boosts performance in downstream classification tasks.

Cross-modal supervision: Contrastive learning using cross-modal supervision has been well-studied in recent literature, primarily in vision-language applications [7, 46, 47]. Extension of such ideas to audio domain has been attempted in the past, by either using the pretrained visual CLIP encoder to distill knowledge into an audio encoder [59], or by training a joint 3-way text-image-audio alignment [21].

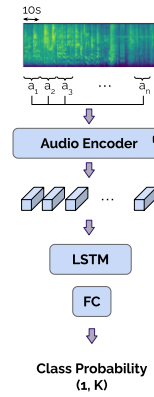
More recently, there has been prevailing work in learning text supervised audio representations for applications of audio retrieval and automatic audio captioning. CLAP [14] applied the contrastive framework on audio captioning datasets and showed the zero-shot capabilities on several audio classification tasks. In a similar vein, MusCALL [35] leveraged a music dataset, tagged with textual descriptions, to learn music representations in a contrastive manner. While these are effective methods, they rely on manually tagged descriptions of audio clips, which are challenging to acquire.

Movie Understanding: Several resources have been curated to address content understanding in long-form multimedia such as movies, including for person/action recognition [17, 20, 31, 37, 53, 56], scene understanding [6, 44, 48, 49], speech and audio understanding [5, 8, 23–25] and general movie understanding [3, 27, 51, 58].

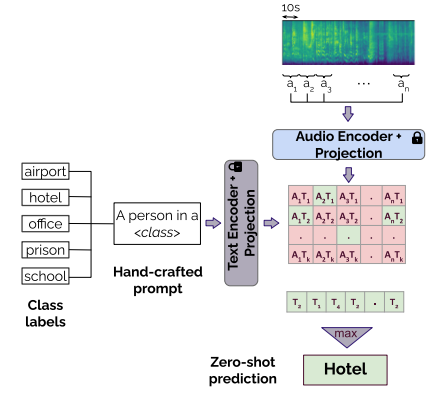
There has been a few efforts recently to learn rich representations using self-supervision toward better movie understanding. Proximity-based contrastive loss and masked reconstruction loss have been explored to learn video-representations for movie classification tasks [58]. Contrastive learning of audiovisual representations in movies using modified negative sampling has also been shown to be useful for action recognition tasks [28]. However, this was trained on an in-house movie dataset more than 5x the size of the largest publicly available movie database. Language descriptions have been used to ground visual representations in movies in two cases. MAD [51] curated professional audio descriptions from 3rd party services which were then automatically transcribed and used to learn grounded visual representations. CMD [3] explored contrastive alignment of movie scenes to textual descriptions for video retrieval application. However, a shortcoming of this method is that only a single scene description was used for the entire movie scene (~2 min), hence the model can only capture high-level concepts and not entity-specific attributes.



(b) Feature extraction



(c) Zero-shot transfer



Downstream Classification

Figure 2: Schematic overview of SEAR. Audio Encoder is learned during pretraining phase and kept frozen for downstream movie understanding tasks

3 METHOD

In this section, we motivate the use of visual captions to distill cross-modal information to learn audio representations in Sec. 3.1. We then describe our proposed method in detail in Sec. 3.2.

3.1 Motivation

Multimedia content such as movies showcase a diverse variety of audiovisual scenes. These two modalities are typically complementary to each other, in order to build a cohesive narrative structure to tell compelling stories in media. The audio profile in a visual clip is strongly linked to foreground entities, the actions they perform as well as the background context (See Fig. 1a). For example, human sounds are typically tied to their actions, including emotive expressions such as crying and laughter and physical actions like running and fighting. Similarly, sounds of things are linked to physical objects shown on screen such as a gun firing or the sound of a door slamming shut. Finally, different visual scenes are typically characterized by their corresponding ambient soundscapes, for example, a library filled with a lot of people is typically associated with a quiet audio profile, whereas a stadium filled with a crowd of people can be assumed to sound quite noisy.

This multimodal correspondence has been used to learn audio representations that are aligned to corresponding visual information using contrastive formulations [21, 40, 59, 61]. However, the inherent variability of images frames depicting similar semantic context (See Fig. 1b) can make it challenging to correlate with accompanying audio characteristics. Modern visual captioning models are capable of capturing these visual entities reliably, acting as an efficient "filter" to extract acoustically relevant attributes of the image. As shown in Fig. 1a, visual captions capture information about objects, location and actions which are linked to the sounds that are heard. Furthermore, as shown in Fig. 1b, captions are also

effective at teasing out this information while being invariable to changes in visual attributes such as texture, color, shape and lighting conditions, which typically have limited acoustic relevance.

Movie scenes typically contain faced-paced shots (typically 2-4s), with certain sound-effects and ambient audio spanning across several shots. This makes it harder to learn coarse audio representations due to high variation in images across shots. Captions, on the other hand, are more convenient since they can be aggregated in the embedding space while still retaining semantic meaning.

A past effort in using captions to learn audio-visual representations leveraged image-captioning datasets to mine additional data in-the-wild by using similarity in the image embedding space [41]. In our work, we address two major limitations of such an approach: (1) Movies depict highly-variable scenes that may not occur in natural-world domains, (2) Slight variations in image space can exhibit vastly diverse acoustic characteristics, f.e., "a police car driving on the road" vs "a police car driving on the road with it's lights on". In the first case, the audio profile might comprise generic vehicular sound, while the latter would exhibit siren-like sounds. Image-space similarities as used in [41] likely will not capture such subtle variations, while captions can easily distinguish them.

Finally, an additional benefit of using language in a contrastive-formulation is the potential to perform zero-shot transfer of learned representations to unseen tasks and domains [46], which we explore for our method.

3.2 SEAR: Semantically-grounded audio representations

We present SEAR, a novel approach to learn audio representations using cross-modal supervision with visual captions. The schematic in Fig. 2a shows the overview of our proposed model. Given a

segment of a movie clip of duration T seconds, we obtain audio and textual representations in the following manner:

Audio Encoding: We use a single audio encoder to model the entire clip. This is done in order to be able to model both short duration sound-effects as well as longer duration sounds such as of ambience and background music. Given an audio segment of T second duration, we obtain token representation of 2D patches from spectrogram representations which are then fed into a transformer encoder. A single learnable token is appended to learn contextual audio representations that are projected onto fixed dimension - which we call the audio representation a .

Visual Captioning: We extract visual captions using off-the-shelf vision-language models (VLN) for image frames corresponding to the clip in the following manner:

$$t_{vc}[k] = VLN(v[k * f_s]) \quad (1)$$

where v - image frames for a movie clip, $k = [1, 2, \dots, T/f_s]$ correspond to frame indices, and f_s - caption sampling rate

Since movie-shots are typically short duration ($< 3s$), we chose the caption sampling rate f_s to be the same as the sampling rate of the clip - hence we extract visual captions for every second of the video (1fps). In this manner, we obtain T contextual captions for a given movie clip. Note that since we use off-the-shelf VLNs, this method is easily scalable and can potentially be improved with better visual captioning systems that provide more descriptive captions.

Textual Encoding: Following this, we use pretrained language models (LM) to obtain representations for each of the T contextual captions:

$$t_{emb}[k] = LM(t_{vc}[k]) \quad (2)$$

where $k = [1, 2, \dots, T]$, $t_{emb} \in \mathbb{R}^{(Tx D)}$ are the contextual caption embeddings and D is the dimension of the LM representations.

Finally, we can pool the multiple contextual caption embeddings to obtain a textual representation t for the movie clip. We explore three different *context-pooling* mechanisms to do this:

1. *Average Pooling:* Here we simply pool the multiple caption representations to obtain a single representation for the clip.

$$t = \frac{1}{T} \sum_{k=1}^T t_{emb}[k] \quad (3)$$

2. *Attention Pooling:* We use scaled dot-product attention mechanism to pool multiple caption representations. This can enable the model to attend to more *acoustically-relevant* captions within a clip.

$$t = \frac{1}{T} \sum softmax\left(\frac{QK^T}{\sqrt{D}}\right)V, \quad (4)$$

$$Q = t_{emb}W^Q; K = t_{emb}W^K; V = t_{emb}W^V$$

where, W^Q , W^K and $W^V \in \mathbb{R}^{(D \times D)}$ are learned during training.

We can now use a standard contrastive loss objective - *infoNCE* [57] as shown below to learn audio representations that are *semantically aligned* to visual caption embeddings,

$$L_{NCE} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{e^{(a_i^T t_i)/\tau}}{\sum_{j=1}^N e^{(a_i^T t_j)/\tau}} \right) \quad (5)$$

where a_i and t_i are the audio and text representations for a clip, τ is the temperature parameter, and N is the batch size.

3. *MIL-NCE:* Finally, we also explore a third way to pool representations from multiple captions. Here, we incorporate the pooling within the objective using a multiple instance learning formulation [38] in the following manner:

$$L_{MIL} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\sum_{k=1}^T e^{(a_i^T t_{ik})/\tau}}{\sum_{j=1}^N \sum_{k=1}^T e^{(a_i^T t_{jk})/\tau}} \right) \quad (6)$$

where T is the number of text captions for a single clip, and t_{ik} refers to the k^{th} caption representation for the i^{th} sample.

As we can see from the objective function, the MIL-NCE formulation posits multiple positive pairs for each sample in a batch. Since sequential captions can capture information about transitions and hence different sound-profiles, we use this formulation to account for multiple *acoustically-relevant* captions within a movie clip.

4 EXPERIMENTS

We employ SEAR for several movie understanding tasks as well as general classification tasks and probe zero-shot capabilities. In this section, we will describe the datasets used for pretraining and downstream classification in Sec 4.1. We will then discuss implementation details for pretraining and evaluation strategies for downstream classification in Sec 4.2. We will then discuss the results for different movie and general classification tasks in Sec 4.3.

4.1 Datasets

4.1.1 *Training:* To train SEAR, we use the following two movie data sources:

Movie Clips: We use the Condensed Movie [3] dataset, which is a set of movie scenes from the Fandango MovieClips YouTube channel. The dataset consists of roughly 33K clips from over 3K movies. Accounting for attrition, i.e., YouTube videos being taken down or made private, we were able to use a total of 31K movie scenes for development.

Full-length Movies: We also use a set of 684 full-length (Hollywood) movies from the years 2014-19. We use the most-popular movies by box-office ratings in order to source movies across a variety of diverse genres.

4.1.2 *Downstream Classification:* We probe the utility of the pretrained audio encoder via several downstream classification tasks. We organize our evaluation along two dimensions: I. in-domain (movies) and out-of-domain tasks, and II. predominantly audio, audiovisual and predominantly visual tasks. We hope to uncover the utility of the pretrained model for both movie and non-movie applications via the experiments along dimension I. Furthermore, although we only use the audio encoder for downstream application, we use the classification of tasks along dimension II to study the performance of our model in tasks that are typically characterized

Table 1: Movie training data statistics

Attribute	Values
# Movies	684 (full) + 3605 (clips)
# Hours	2512
# Captions	9,620,380
# Unique	1,449,402
# words/caption	9.95 ± 1.64
# nouns/caption	3.24 ± 0.8
# verbs/caption	0.97 ± 0.49

by fine-grained audio features (e.g., speaking-style classification, audio event detection), in comparison to tasks that are typically audiovisual [AV] in nature (genre classification, acoustic scene classification), and tasks that are generally model using visual features [V] (visual scene classification, action recognition).

LVU: Long-form Video Understanding [58] benchmark comprises several video classification tasks for movies. We show results for two content understanding tasks, i.e., scene and speaking-style classification and one metadata understanding task, i.e., genre-classification. We chose these subset of tasks since audio plays a direct role in each of these tasks. In fact, to the best of our knowledge, this is the first effort toward movie understanding using audio as a primary signal. Along II, we categorize speaking-style as the primarily audio task, genre classification as the audiovisual task and scene classification as the primarily visual task. Due to attrition in YouTube videos, a small fraction of videos are not available since the original release. Details about dataset size, number of classes and evaluation protocol are shared in the Supplementary section.

General Classification: In order to perform comprehensive evaluation of the pretrained audio encoder, we chose different audio classification tasks - including fine-grained audio event detection (AED), which is a predominantly audio task, acoustic scene classification which is an audiovisual task, and action recognition which is a predominantly visual task. We show results on the following four datasets (details included in Supplementary) :

ESC-50 [45] is an environmental sound classification dataset used to benchmark AED models. It comprises 50 different fine-grained sound classes, varying based on source; including human and animal sounds, indoor and outdoor sounds and natural sounds.

TUT2017 [36] is an acoustic scene classification dataset consisting of recordings collected from distinct locations (both indoor and outdoor).

Moments in Time [39] is a large-scale action recognition dataset, developed to understand short-duration (~3s) actions and events in videos in the wild.

Kinetics-Sounds [1] is a subset of the larger Kinetics-400 dataset, consisting of 36 action classes that have characteristic audio profiles. Due to attrition, we show results on 29 classes that we were able to download.

4.2 Implementation Details

4.2.1 Pretraining: To generate visual captions, we use a BLIP-based model [32], made available for open-source use through the LAVIS

Table 2: Movie Classification Results on LVU Benchmark. Top-1 Classification Accuracy averaged over 5 repeated runs for genre and 10 runs for speaking-style/scene for audio baselines and proposed method. Self-supervised audio baselines include hubert, wav2vec2, SSAST-(A/M). AST is the fully-supervised audio baseline

Method	Modality	Speaking-Style	Genre	Scene
VideoBERT [55]	V	37.9±0.9	51.9±0.6	54.9±1.0
Object Transformer [58]	V	39.4±1.2	54.6±0.6	56.9±1.4
Movie2Scenes [9]	V	42.2	55.9	68.2
hubert [26]	A	37.98±2.18	47.20±0.83	31.20±0.65
wav2vec2 [2]	A	38.19±3.39	43.07±1.48	30.6±2.36
SSAST-A [19]	A	36.06±2.73	53.11±2.58	28.92±1.32
SSAST-M	A	38.83±1.26	53.86±1.31	30.84±1.72
AST	A	40.9±2.12	53.24±2.62	40.84±3.90
Random	-	22.89	25.41	18.36
SEAR-ZS	A	29.79	47.6	49.4
SEAR	A	47.66±1.39	56.76±1.5	48.67±1.45

library¹. We extract visual captions at 1 fps for video snippets of T=10 seconds each, resulting in a total of 10 captions per clip during training. Table 1 shows additional details about caption statistics.

Text Encoder: We use three different text encoders for our experiments. We evaluate two language models, BERT [12] and RoBERTa [33], since these have been shown to be powerful for learning audio representations using acoustic captions [60]. We also evaluate CLIP [46] which was originally trained for learning image and language representations with web-scraped text-image pairs. In our early experiments, we find that keeping the text encoder frozen results in better learned audio representations, so we entirely freeze each of these encoders and only learn a single text-projection layer to map to the audio encoder dimensions.

Audio Encoder: We use the SSAST architecture [19] inspired from vision transformers, which uses mel-spectrograms as inputs and a patch-based 2D-CNN to learn token embeddings as input to the transformer.

We use AdamW [34] optimizer with an initial learning rate of 1e-4. We perform linear warmup for 12K training iterations with a batch size of 180 and subsequently use an exponential-decay scheduler with gamma parameter of 0.8. Similar to previous contrastive methods [46], we initialize the temperature parameter in Eq. 5 to a value of 0.07 and learn its value during training.

4.2.2 Downstream Classification Tasks: For the movie understanding tasks, classification is performed at scene level, which are around 2 min long. We use the trained audio encoder as a feature extractor to extract segment-level features. A single-layer LSTM model with a hidden dimension of 512 is trained for each of the classification tasks (See Fig. 2b). Genre classification performance is reported as the average of 5 repeated runs, and speaking style and scene classification results are reported over 10 repeated runs. We use top-1 classification accuracy as the evaluation metric [58].

For the general-domain classification tasks, we perform full-finetuning of the pretrained audio encoder in order to account for domain mismatch. We show comparison with training from randomly-initialized weights (scratch) to highlight the benefit of pretraining.

¹<https://github.com/salesforce/LAVIS>

Table 3: Results on general classification tasks, *zeroshot, **audio-only results reported. First three rows show self-supervised baselines, FS SotA: best fully-supervised performance. Scratch: SEAR encoder finetuned from randomly initialized weights. All metrics shown are top-1 accuracy

Method	ESC-50	TUT2017	Moments in Time	Kinetics-Sounds
SSAST [19]	88.80	-	-	-
MS-CLAP [14]	96.70	74.63	-	-
LAION-CLAP* [60]	91.00	-	-	-
MBT** [42]	-	-	8.20	52.60
FS SotA (Audio)	98.10 [10]	84.30 [22]	8.20	52.60
Scratch	54.00	64.14	8.16	45.66
SEAR	86.95	82.30	14.34	66.65

Table 4: Effect of different context-pooling methods

Context-pooling	Speaking-Style	Genre	Scene
Average	47.66±1.39	55.12±1.53	47.35±3.41
Attention	43.72±2.33	56.14±1.82	47.35±1.33
MIL-NCE	42.45±2.08	56.76±1.5	46.14±2.95

Baseline Models: We show comparison with previous video based methods for the movie understanding tasks, Due to the lack of previous audio methods, we also perform classification experiments with three competitive self-supervised audio representation models, namely, Hubert [26], wav2vec2.0 [2] and SSAST [19]. The first two are masked-reconstruction based methods primarily developed for speech applications including speech recognition and emotion recognition, while the third is a self-supervised method for speech and audio classification, i.e., including audio event detection and spoken language understanding. For fair comparison, and to better understand the merits of movie data for pretraining, we train a version of SSAST on our movie training data. We refer to the original model pretrained on Audioset + LibriSpeech as SSAST-A, and the version trained on movies as SSAST-M. Finally, we also show competitive performance with a fully-supervised (FS) model, AST [18], trained for AED on Audioset [16].

For the general classification tasks, we show previous methods based on self-supervised pretraining and the fully-supervised state-of-the-art for each task.

4.2.3 Zero-shot Task Transfer: Following training of the contrastive method outlined in 3, we can perform zeroshot transfer of the learned representations using the aligned text-representations. The framework for zeroshot transfer to scene classification is shown in Fig. 2c. Contextual prompts can be crafted for each of the classes, following which we use simple dot-product between the audio representation of a given sample and each of the class-embeddings from the text encoder to ascertain the predicted class in a zeroshot manner. Due to the discrepancy in temporal resolution of movie clips (1-3 min instead of 10s used in pretraining), we can then aggregate zeroshot predictions over multiple non-overlapping audio segments to obtain a class prediction for a given sample. For general classification tasks, we can directly perform zeroshot transfer without the need for aggregation. We discuss zeroshot performance for one action recognition dataset and also show some qualitative analysis in Sec. 5.

Table 5: Ablation for different language models used during pretraining

Language Model	Speaking-Style	Genre	Scene
CLIP-Text [46]	47.66±1.39	55.12±1.53	47.35±3.41
BERT-base [12]	43.4±3.37	55.94±1.36	45.06±1.96
RoBERTa-base [33]	45.27±2.22	54.81±1.03	48.67±1.45

4.3 Results

4.3.1 Movie Understanding. We present the results for movie understanding on LVU in Table 2. We slightly outperform external video baselines for genre-classification (56.76 vs 55.9) and significantly outperform video methods for speaking-style classification (47.66 vs 42.2). Since a movie scene is primarily a visual depiction, it is not surprising that the audio based methods significantly underperform compared to the video methods (48.67 vs 68).

However, we see that SEAR clearly outperforms other audio-only (SSL and FS) baselines on all three tasks. The boost in performance is more significant for scene classification, which seems to suggest that our representation is indeed capturing visually-relevant information.

Ablations: We also perform ablation experiments to analyze the effect of different *context-pooling* methods and language models used during pretraining. From Table 4, we see that the parametric approach using attention performs worse than average pooling on speaking-style classification and comparable on the other two tasks. Similarly, the multiple-instance loss formulation MIL-NCE to pool contextual embeddings performs slightly worse for speaking-style classification and slightly better for genre classification. We claim that the comparable nature of these approaches is due to the ability of language models to robustly aggregate multiple contextual embeddings while retaining semantically-relevant information from each of them. This allows for simple aggregation over captions to be sufficient to capture contextual information.

We show results for ablation on different text encoders in Table 5. As can be seen, there is no clear consensus on the best encoder across tasks. This is different from previous work with acoustic captions [60], which showed significantly worse results with clip encoder. A reason for this discrepancy could be due to the fact that clip [46] was trained with paired language-image pairs, which are already tuned to visual caption style of language. However, we do observe that model trained with RoBERTa embeddings performs best for scene classification.

4.3.2 General Classification: Table 3 shows the results on general classification tasks. We see that our method is subpar to competitive audio-based methods for AED on ESC-50. This is understandable since our pretraining does not contain any information about fine-grained sounds. On the other hand, for acoustic scene classification we are able to perform comparably to SotA (82.3 vs 84.3). This could be because of the fact that our pretraining learns some knowledge about scene information similar to what we observed in the movie understanding task. Finally, for both of the action-recognition datasets, we are able to significantly outperform the previous SotA [42]. We can observe here the benefit of our method

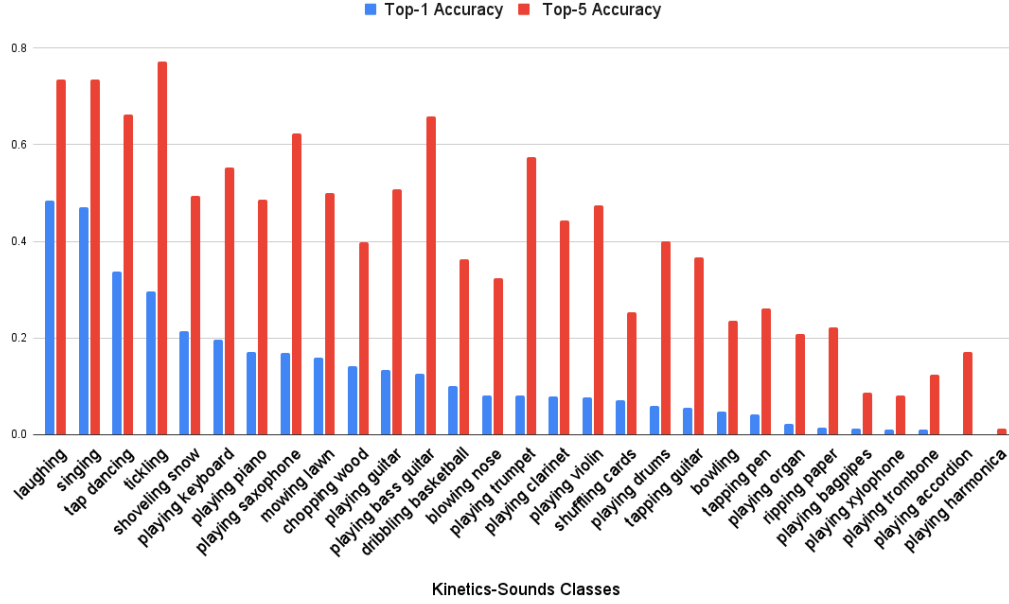


Figure 3: Zero-shot transfer to action recognition dataset

Table 6: Effect of different prompt choices for zero-shot scene classification

Prompt	Accuracy
<scene>	25.30
a <scene> scene	34.94
a visual ambience of <scene>	38.55
a scene of a <scene>	39.76
a person standing in a <scene>	49.4

for predominantly visual tasks such as action recognition due to the presence of actions in visual captions.

5 ZERO-SHOT TASK TRANSFER

In this section, we discuss in detail the zero-shot application of our method through both quantitative (Sec. 5.1) and qualitative (Sec. 5.2) analyses in order to probe the semantic information encoded by the learned audio representations.

5.1 Quantitative Results

5.1.1 Movie Understanding: We apply zero-shot transfer using our learned audio and text encoders for each of the movie understanding tasks. We use the following hand-crafted prompts for each task:

scene: "A person in a <class>"
speaking-style: "A person <class>"
genre: "People in a <class> movie"

Results for the zero-shot approach on LVU are shown in Table 2. We can notice that zero-shot performance is better than random

baseline for all of the tasks, which indicates that the model is capturing relevant information for each of the tasks during training. We can notice that performance for scene classification is comparable and even slightly better than the feature-extraction-based approach, and significantly outperforms all audio baselines without the need for any training. This can be attributed to the fact that visual captions capture scene information frequently and quite reliably, and hence the representations encode information about background context.

Similarly, poor performance for speaking-style classification (29.79%) can be due to the fact that the nature of labels seen (e.g., discuss vs explain) are not typically encoded in visual captions and require explicit supervision in order to be able to distinguish between them. Finally, we observe that zero-shot performance for genre classification is also quite competitive (47.6%), outperforming two of the four audio baselines. Since genre is an abstracted measure of different audio-visual elements, we do expect visual captions to capture information that is correlated with different genre types resulting in competitive zero-shot performance. Zero-shot ablation for different LMs and pooling methods is shared in Supplementary.

The choice of prompts also has a significant impact on zero-shot performance. We show results from a few different prompts for scene classification in Table 6 to highlight the impact of using appropriate prompts. The best-performing prompt - "a person standing in a <scene>", is chosen to have a similar sentence structure as produced by the visual captioning model, which are primarily entity-driven. Therefore, prompts have to be carefully crafted depending on the nature of captions seen during pretraining.

5.1.2 Action Recognition on Kinetics-Sounds. We perform similar zero-shot classification on the kinetics-sounds dataset [1], which



Figure 4: Representative frames from retrieved clips for different audio event prompts; Green border - relevant sound present in clip, Red border - relevant sound not present in clip

includes audio-centric action classes from kinetics. We obtain a zero-shot performance of 12% compared to random baseline of 3.5%.

In order to investigate the utility of the learned representations for different action classes, class-wise zeroshot performance (top-1 and top-5 accuracy) are shown in Fig. 3. We can see that human expressed actions of laughing and singing perform best. This could be due to the prevalence of these actions in visual captions seen during pretraining. Furthermore, our model performs variably on the different musical instrument classes, being able to detect roughly 20% of keyboard, piano, saxophone and guitar/bass guitar sounds, while completely failing for certain instruments such as xylophone, trombone, accordion and harmonica. This discrepancy can also be attributed to the exposure to these different instruments. Further substantiation of this claim is shown in Supplementary based on frequency of actions seen in captions during pretraining.

5.2 A qualitative analysis of SEAR

We use the pretrained encoders to retrieve samples for different sound classes on movie samples from LVU. We do so by using sound tags that frequently occur in our captions as prompts and obtain the top-5 retrieved clips. In Fig. 4, we have shown representative frames from each of the retrieved clips. Each of the clips was manually tagged with the presence/absence of the corresponding sound class and shown with green/red border in the figure.

On the left, we show sound classes that are mostly correctly retrieved. A common characteristic across these classes is the strong audiovisual co-occurrence of the object-sound pairs. For example, rain shown on screen is typically accompanied by water sounds. Similarly, if a musical instrument is shown on screen, you can typically hear the corresponding instrument playing and so on. While we trained the model for semantic similarity using visual captions, the model implicitly learns to cluster in the acoustic space as well. This is supported by the fact that "helicopter" retrieves clips with "engine" sounds such as tractors, trains and heavy machinery. It is also interesting to note that "telephone" tagged clips do pick up telephonic and other atypical speech but not necessarily telephone rings. This also means that it does pick up incorrect samples with unnatural/muffled speech, which is not necessarily a telephone conversation.

There are some commonly occurring sound classes that however

fail to be retrieved correctly most of the time, some of which are shown on the right column of Fig. 4. These are sounds that typically do not exhibit strong audiovisual co-occurrence. For example, even though dogs and cars are commonly observed in movies, they are less likely to be heard during their active screen time.

6 CONCLUSION

In this work, we propose a method of using visual captions to semantically ground audio representations (SEAR) in a self-supervised setup. We use state-of-the-art vision-language models to augment movie datasets with visual captions and demonstrate the effectiveness of our approach on several benchmark movie understanding tasks, including genre, scene, and speaking-style classification. We further show competitive performance in general classification tasks, outperforming previous methods in the domain of action recognition. Finally we uncover the information encoded by the learned representations by performing zero-shot and qualitative analysis. In the future, we hope to leverage large-language models to augment information from auxiliary acoustic models.

Limitations: A limitation of our method is on the dependence of audiovisual activations, i.e., an object being heard whenever it is seen on screen. A potential way to mitigate this could be to use weak information from AED models to filter instances where relevant sounds are not heard (e.g., car is shown on screen but audio contains only speech).

Fairness considerations: In our approach, we do not control for any inherent biases in the visual captioning system or in the movie-training data. Therefore, it is possible that biases that arise due to character depictions in movies (e.g., men typically have higher on-screen time than women) could also seep into learned representation using SEAR. We hope to investigate potential biases and address mitigation strategies in future work.

REFERENCES

- [1] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*. 609–617.
- [2] Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv abs/2006.11477* (2020).
- [3] Max Bain, Arsha Nagrani, A. Brown, and Andrew Zisserman. 2020. Condensed Movies: Story Based Retrieval with Contextual Embeddings. *ArXiv abs/2005.04208* (2020).

- [4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 1708–1718.
- [5] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.
- [6] Digbalay Bose, Rajat Hebbar, Krishna Somanepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. 2023. MovieCLIP: Visual Scene Recognition in Movies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2083–2092.
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language Models are Few-Shot Learners. *ArXiv abs/2005.14165* (2020).
- [8] Sourish Chaudhuri, Joseph Roth, Daniel PW Ellis, Andrew Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin Wilson, et al. 2018. Ava-speech: A densely labeled dataset of speech activity in movies. *arXiv preprint arXiv:1808.00606* (2018).
- [9] Shixing Chen, Xiang Hao, Xiaohan Nie, and Raffay Hamid. 2022. Movies2Scenes: Learning Scene Representations Using Movie Similarities. *arXiv preprint arXiv:2202.10650* (2022).
- [10] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. BEATS: Audio Pre-Training with Acoustic Tokenizers. *arXiv preprint arXiv:2212.09058* (2022).
- [11] Vesna Dakic. 2009. Sound design for film and television.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [13] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an Audio Captioning Dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 736–740.
- [14] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. CLAP: Learning Audio Concepts From Natural Language Supervision. *ArXiv abs/2206.04769* (2022).
- [15] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 829–852.
- [16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.
- [17] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The REPERE Corpus: a multimodal corpus for person recognition.. In *LREC*. 1102–1107.
- [18] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).
- [19] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James R. Glass. 2022. SSAST: Self-Supervised Audio Spectrogram Transformer. In *AAAL*.
- [20] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056.
- [21] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 976–980.
- [22] Yoonchang Han, Jeongsoo Park, and Kypogu Lee. 2017. Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification.. In *DCASE*. 46–50.
- [23] Rajat Hebbar, Digbalay Bose, Krishna Somanepalli, Veena Vijai, and Shrikanth Narayanan. 2023. A dataset for Audio-Visual Sound Event Detection in Movies. *arXiv preprint arXiv:2302.07315* (2023).
- [24] Rajat Hebbar, Krishna Somanepalli, and Shrikanth Narayanan. 2019. Robust speech activity detection in movie audio: Data resources and experimental evaluation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4105–4109.
- [25] Rajat Hebbar, Krishna Somanepalli, and Shrikanth S Narayanan. 2018. Improving Gender Identification in Movie Audio Using Cross-Domain Data.. In *Interspeech*, Vol. 282. 286.
- [26] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.
- [27] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV* 16. Springer, 709–727.
- [28] Mahdi M. Kalayeh, Shervin Ardeshtir, Lingyi Liu, Nagendra Kamath, and Ashok Chandrashekar. 2022. On Negative Sampling for Audio-Visual Contrastive Learning from Movies. *ArXiv abs/2205.00073* (2022).
- [29] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL*.
- [30] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. 2021. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069* (2021).
- [31] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
- [32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.
- [33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [34] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.
- [35] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive Audio-Language Learning for Music. *ArXiv abs/2208.12208* (2022).
- [36] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. 2017. DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. 85–92.
- [37] Pascal Mettes, Jan C Van Gemert, and Cees GM Snoek. 2016. Spot on: Action localization from pointly-supervised proposals. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V* 14. Springer, 437–453.
- [38] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 9876–9886.
- [39] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. 2019. Moments in Time Dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–8. <https://doi.org/10.1109/TPAMI.2019.2901464>
- [40] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12475–12486.
- [41] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 407–426.
- [42] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* 34 (2021), 14200–14213.
- [43] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2021. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [44] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. 2022. Moviecuts: A new dataset and benchmark for cut type recognition. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 668–685.
- [45] Karol J. Piczak. [n. d.]. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia* (Brisbane, Australia, 2015-10-13). ACM Press, 1015–1018. <https://doi.org/10.1145/2733373.2806390>
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.
- [47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv abs/2204.06125* (2022).
- [48] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A unified framework for shot type classification based on subject centric lens. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* 16. Springer, 17–34.
- [49] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10155.
- [50] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3875–3879.

- [51] Mattia Soldan, A. Pardo, Juan Le'on Alc'azar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. MAD: A Scalable Dataset for Language Grounding in Videos from Movie Audio Descriptions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 5016–5025.
- [52] Krishna Somandepalli, Tanaya Guha, Victor R Martinez, Naveen Kumar, Hartwig Adam, and Shrikanth Narayanan. 2021. Computational media intelligence: Human-centered machine analysis of media. *Proc. IEEE* 109, 5 (2021), 891–910.
- [53] Krishna Somandepalli, Rajat Hebbar, and Shrikanth Narayanan. 2020. Multi-face: Self-supervised multiview adaptation for robust face clustering in videos. *arXiv preprint arXiv:2008.11289* (2020).
- [54] Janne Spijkervet and John Ashley Burgoyne. 2021. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410* (2021).
- [55] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.
- [56] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. 2012. "Knock! Knock! Who is it?" probabilistic person identification in TV-series. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2658–2665.
- [57] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv abs/1807.03748* (2018).
- [58] Chaoxia Wu and Philipp Krähenbühl. 2021. Towards Long-Form Video Understanding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 1884–1894.
- [59] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning Robust Audio Representations from Clip. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022), 4563–4567.
- [60] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. *arXiv preprint arXiv:2211.06687* (2022).
- [61] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*. 570–586.