Audio Retrieval with Natural Language Queries: A Benchmark Study

A. Sophia Koepke*, Andreea-Maria Oncescu*, João F. Henriques, Zeynep Akata, and Samuel Albanie

Abstract—The objectives of this work are *cross-modal text-audio and audio-text retrieval*, in which the goal is to retrieve the audio content from a pool of candidates that best matches a given written description and vice versa. Text-audio retrieval enables users to search large databases through an intuitive interface: they simply issue free-form natural language descriptions of the sound they would like to hear. To study the tasks of text-audio and audio-text retrieval, which have received limited attention in the existing literature, we introduce three challenging new benchmarks. We first construct text-audio and audio-text retrieval benchmarks from the AUDIOCAPS and CLOTHO audio captioning datasets. Additionally, we introduce the SOUNDDESCS benchmark, which consists of paired audio and natural language descriptions for a diverse collection of sounds that are complementary to those found in AUDIOCAPS and CLOTHO. We employ these three benchmarks to establish baselines for cross-modal text-audio and audio-text retrieval, where we demonstrate the benefits of pre-training on diverse audio tasks. We hope that our benchmarks will inspire further research into audio retrieval with free-form text queries. Code, audio features for all datasets used, and the SOUNDDESCS dataset are publicly available at https://github.com/akoepke/audio-retrieval-benchmark.

Index Terms—Audio Retrieval, Text-based Retrieval, Datasets

I. INTRODUCTION

THE vast and unabated growth of user-generated content in recent years has introduced a pressing need to search ever-growing databases of multimedia. Free-form natural language sentences (i.e. sequences of text that are written as they would be spoken) form an intuitive and powerful interface for composing search queries for these databases since they allow for expressing virtually any concept. Spanning multiple modalities, different retrieval strategies were developed for content as diverse as text (including web pages and books), images [1], and videos [2], [3]. Surprisingly, while search engines currently exist for these modalities (e.g. Google, Flickr and YouTube, respectively), unstructured audio is not accessible in the same way. The aim of this paper is to address this gap by curating the SOUNDDESCS dataset which contains paired sound and natural language and by introducing benchmarks for text-audio retrieval.

* Equal contribution.

A. S. Koepke and Z. Akata are with the Explainable Machine Learning group at the University of Tübingen. Z. Akata is also affiliated with the Max Planck Institute for Intelligent Systems, Tübingen and the Max Planck Institute for Informatics, Saarbrücken.

A.-M. Oncescu and J. Henriques are with the Visual Geometry Group at the University of Oxford.

S. Albanie is with the Department of Engineering at the University of Cambridge.

It is important to distinguish content-based retrieval from that based on metadata, such as the title of a video or song or an audio tag. Metadata retrieval is feasible for manually-curated databases such as song or movie catalogues. However, content-based retrieval is more important in user-generated data, which often has little structure or metadata. There are methods to search for audio which matches an audio query [4], [5], but satisfying the requirement to input an example audio query can be difficult for a human (e.g. making convincing frog sounds is difficult). We, on the other hand, propose a framework which enables the searching of a sound database using detailed free-form natural language queries of the desired sound (e.g. "A man talking as music is playing followed by a frog croaking."). This enables the retrieval of audio data which will ideally match the temporal sequence of events in the query instead of just a single class tag. Furthermore, natural language queries are a familiar user interface widely used in current search engines. Therefore, our proposed audio retrieval with free-form text queries could be a first step towards more natural and flexible audio-only search.

Text-based audio retrieval could also be beneficial for video retrieval. The majority of recent works that address the text-based video retrieval task focus heavily on the visual and text domains [1], [2], [6], [7]. Since audio and visual information inherently have natural semantic alignment for a significant portion of video data, text-based audio retrieval could also be used for querying video databases by only considering the audio stream of the video data. This would allow for video retrieval in the audio domain at reduced computational cost for cases in which audio and visual information correspond, with applications for low-power IoT devices, such as microphones in natural habitats, of particular interest for conservation and biology. Historical archives with extensive sound collections, such as the British Library Sounds[1], would be easier to search, facilitating historical research and public access. Furthermore, text-based retrieval could enable appealing creative applications, such as automatically finding (non-music) background sounds which correspond to input text. This could be especially useful given the growing popularity of audio podcasts and audiobooks which are often supplemented with (background) sounds that match their content.

Learning to retrieve audio, given natural language queries, requires data with paired text and sound. Audio captioning datasets naturally lend themselves to this task, since they contain audio and a matching text description for the sound. However, existing captioning datasets are limited in size and

[1]https://sounds.bl.uk

in the diversity of their audio content. Hence, we curated the novel SOUNDDESCS dataset, which was sourced from the BBC Sound Effects database[2]. SOUNDDESCS contains text describing the sounds with significant variation with respect to the audio content and 2with a relatively large vocabulary used in the descriptions.

We introduce three new benchmarks for text-based audio retrieval, based on our proposed SOUNDDESCS dataset, and the AUDIOCAPS [8] and CLOTHO [9] audio captioning datasets. AUDIOCAPS consists of a subset of 10-second audio clips from the AudioSet dataset [10] with additional human-written audio captions, while CLOTHO contains sounds sourced from the Freesound platform [11], varying between 15 and 30 seconds in duration, and accompanied by crowd-sourced text captions. SOUNDDESCS is considerably more varied in duration and audio content than the other two benchmarks. However, the text descriptions are of mixed quality, since they are obtained automatically from descriptions provided with the data.

In contrast to sound event class labels, audio captions contain detailed information about the sounds. A user searching for a particular sound would usually describe the sound using text similar to an audio caption. AUDIOCAPS [8], CLOTHO [9], and SOUNDDESCS allow to leverage the matching text-audio pairs to train text-based audio retrieval frameworks. To establish baselines for this task, we adapt existing video retrieval frameworks for audio retrieval. We employ multiple pre-trained audio expert networks and show that using an ensemble of audio experts improves audio retrieval.

In summary, we make three contributions: (1) we introduce the SOUNDDESCS dataset for text-based audio retrieval; (2) we introduce three new benchmarks for audio retrieval with natural language queries—to the best of our knowledge, these represent the first public benchmarks for this task; (3) we provide baseline performances with existing multi-modal video retrieval models that we adapt to text-based audio retrieval and show the benefits of using multiple datasets for pre-training.

This paper extends an initial Interspeech 2021 conference version of our work [12] in two ways: (i) we introduce the new SOUNDDESCS dataset for text-audio retrieval and provide an analysis of its characteristics (in Sec. III), (ii) we provide more extensive baselines for the audio retrieval task with more detailed ablations across datasets and an additional retrieval architecture (Sec. V). In particular, we explore the use of the Multi-modal Transformer architecture [7].

II. RELATED WORK

Our work relates to several themes in the literature: *sound event recognition*, *audio captioning*, *audio-based retrieval*, *text-based video retrieval* and *text-domain audio retrieval*. We discuss each of these next.

Sound event recognition. There is a rich literature addressing the task of sound event recognition, which seeks to assign a given segment of audio with a corresponding semantic label. Examples include detecting audio events associated with sports [13], urban sounds [14], and distinguishing vocal and

nonvocal events [15]. Research in this area has been driven by challenges, such as DCASE [16], [17], and by the collection of sound event datasets. These include TUT Acoustic scenes [18], CHIME-Home [19], ESC-50 [20], FSDKaggle [21], and AudioSet [10]. Of relevance to our approach, a number of prior works have employed deep learning for audio comprehension [22], [23], [24], [25], [26]. Our work differs from theirs in that we focus instead on the task of retrieval with natural language queries, rather than audio recognition.

Audio captioning. Audio captioning consists of generating a natural language description for a sound [27]. This requires a more detailed understanding of the sound than simply mapping the sound to a set of labels (sound event recognition). Recently, several audio captioning datasets have been introduced, such as CLOTHO [9] which was used in the DCASE automated audio captioning challenge 2020 [28], Audio Caption [29], and AUDIOCAPS [8]. Drawing inspiration from work on video captioning [30], [31], multiple works have addressed automatic audio captioning on the AUDIOCAPS and CLOTHO datasets [32], [33], [34], [35], [36]. In this work, we use the AUDIOCAPS and CLOTHO datasets for cross-modal retrieval.

Audio-based retrieval. Multiple content-based audio retrieval frameworks, in particular query by example methods, leverage the similarity of sound features that represent different aspects of sounds (e.g. pitch, or loudness) [37], [38], [39], [5]. More recently, [4] use a twin neural network framework to learn to encode semantically similar audio close together in the embedding space. [40] address multimedia event detection using only audio data, while [41] tackle near-duplicate video retrieval by audio retrieval. These are purely audio-based methods that are applied to video datasets, but without using visual information. [42] propose a two-step approach for video retrieval which uses audio (coarse) and visual (fine) information together.

Text-based video retrieval. More closely related to our work, a number of methods showed that embedding video and text jointly into a shared space (such that their similarity can be computed efficiently) is an effective approach [1], [3], [2], [6], [43], [7], [44] (though other formulations, such as computing similarities directly in visual space have also been explored [45]). One particular trend has been to combine cues from several "experts"—pre-trained models that specialise in different tasks (such as object recognition, action classification etc.) to inform the joint embedding. Recently, transformer-based architectures have demonstrated impressive results for text-based video retrieval [7], [46], [47]. In this work, we propose to adapt three expert-based methods: the Mixture of Embedded Experts method of [2], the Collaborative Experts model of [6], and the Multi-Modal Transformer [7] by repurposing them for the task of audio retrieval (described in more detail in Sec. IV).

Cross-domain audio retrieval. Methods that retrieve audio by matching associated text, such as metadata or sound event labels, have the implicit assumption that the text is relevant [48]. 2In contrast, [49] is an early work that proposes to link audio and text representations in hierarchical semantic and acoustic spaces. [50] builds on this using mixture-of-probability-expert models for each of the modalities. Chechik et al. [51] propose
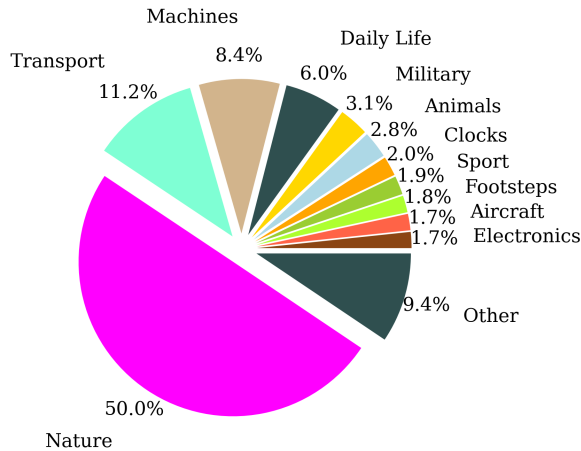
Fig. 1: Pie chart showing the distribution of audio files in the SOUNDDESCS dataset over different categories.

Fig. 2: Histogram of audio files lengths for the AUDIOCAPS, CLOTHO and SOUNDDESCS datasets.

a text-based sound retrieval framework which uses single-word audio tags as queries rather than caption-like natural language. Similarly, [52], [53] learn shared latent spaces between onomatopoeias (words that mimic non-speech sounds) and sound for searching audio using onomatopoeia queries and for generating sound words from audio. The creative approach of [54] learns to align visual, audio, and text representations to enable cross-modal retrieval. Their framework is trained with captioned images and paired image-sound data (sourced from videos) and evaluated using the soundtrack of captioned videos. Other works have explored using images [55] or video data [56], [57], [58], [59], [60] as queries for retrieving audio. More recently, [61] use a twin network to learn a shared latent text and sound space for cross-modal retrieval. While they use class labels as text labels, we study unconstrained text descriptions as queries. Another highly related, concurrent line of works has explored the task of grounding sounds given a text description. [62], [63] presented results for the grounding task on the AudioGrounding dataset, proposed by [62], which augments a subset of the AUDIOCAPS dataset (approximately 10% of the full AUDIOCAPS dataset) by adding fine-grained temporal grounding labels. In contrast to the audio grounding task, text-based audio retrieval does not require expensive temporal annotations and we can therefore leverage large databases which contain immensely varied content.

III. SOUNDDESCS DATASET

In this section, we introduce the SOUNDDESCS dataset for text-audio retrieval. The SOUNDDESCS dataset consists of 32,979 audio files accompanied by natural language descriptions. We present an overview of the SOUNDDESCS dataset by describing how it was collected in Section III-A, and by providing an analysis and comparison to related datasets which contain pairs of audio files and text descriptions in Section III-B. Furthermore, we show some examples of the data in Section III-C.
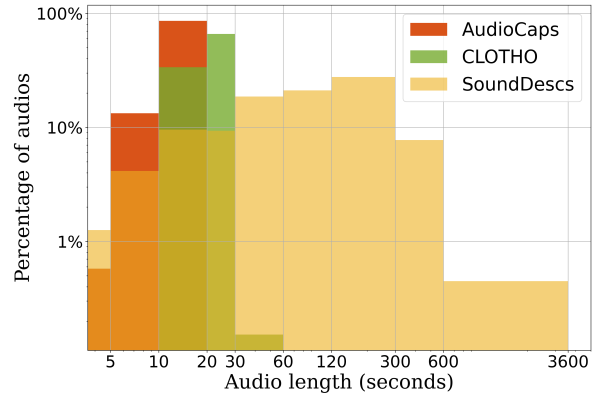
A. Dataset collection

The SOUNDDESCS data was sourced from the BBC Sound Effects webpage[2]. It contains audio files and corresponding textual descriptions of a wide range of sounds from the BBC Radiophonic workshop, the Blitz in London, special effects made for the BBC, and recordings from the Natural History Unit archive. The sounds are of high quality and were recorded for professional applications, such as radio and TV special effects. In some cases, additional information is provided which contains the size and number of channels of the audio file together with the date it was recorded and other sound tags. The audio files are associated with 23 categories, including but not limited to *nature*, *clocks*, *fire*, etc. We show the proportion of files from the different categories in Fig. 1. For this, we use the text tags accompanying the collected audio files. Since some of the files contain multiple text tags, we collected all of them in a bag-of-words fashion and then provided the frequencies with which they appear. We obtained 32,979 audio files sampled at 44.1 Hz which have a non-empty textual description (out of the 33,066 audio files on the BBC website). We propose to split the SOUNDDESCS dataset into training/validation/test subsets by randomly selecting 70% of the files for training, and 15% each for validation and testing.

B. Data analysis

We compare the SOUNDDESCS dataset to related datasets which contain matching sound and language data in Table I. The two main novelties of the SOUNDDESCS dataset compared to related sound-text datasets are the wide variation in duration of the audio files and the size of the vocabulary used for the descriptions. The audio captioning datasets AUDIO-CAPS [8][3] and CLOTHO [9] contain audio files that are only 10-30 seconds long. As can be seen in Fig. 2, SOUNDDESCS contains sounds with a much wider range of audio durations with 109 files lasting longer than 10 minutes. Processing audio files with such variations in duration is challenging, but it enables the detailed analysis of the performance of current text-audio and audio-text retrieval methods with respect to

---

[3]The numbers provided for the AUDIOCAPS dataset in Table I correspond to the subset which does not have an overlap with the VGGSound dataset.

TABLE I: Comparative overview of sound-language datasets. Comparing the number of files in the different datasets, including their training, validation, and test subsets, audio duration (dur.) and caption lengths (#words) measured in seconds and words respectively. Text is sourced from human caption annotations or audio descriptions provided with the sound data.

| Dataset	Text source	language	duration(h)	#audios	#captions	max dur.(s)	avg dur.(s)	max #words	avg #words	train	val	test
AUDIOCAPS [8]	Human captions	English	135.01	50535	55512	10.08	9.84	52	8.80	49291	428	816
CLOTHO [9]	Human captions	English	22.55	3938	3938	30.00	22.44	21	11.32	2314	579	1045
Audio Caption [29]	Human captions	Chinese	10.3	3707	3707	N/A	10.00	54	11.14	3337	-	371
SOUNDDESCS	Descriptions	English	1060.40	32979	32979	4475.89	115.75	65	15.28	23085	4947	4947
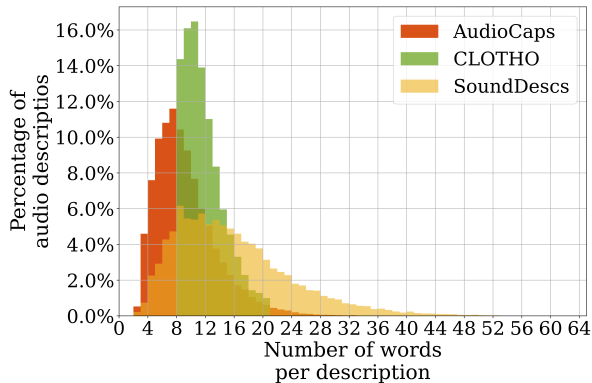
Fig. 3: Distribution of the number of words per caption for the AUDIOCAPS, CLOTHO and SOUNDDESCS datasets.
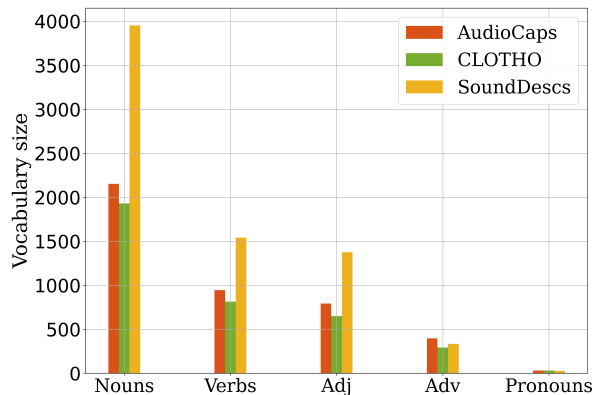

Fig. 4: Vocabulary size of descriptions in the SOUNDDESCS dataset, compared to the AUDIOCAPS and CLOTHO datasets, divided into nouns, verbs, adjectives, adverbs, and pronouns.

the audio duration. In addition to this, the SOUNDDESCS dataset is larger than all related sound-language datasets with a total duration of 1060 hours, compared to 135 hours for AUDIOCAPS. Since SOUNDDESCS contains fewer audio files with associated captions than AUDIOCAPS, the SOUNDDESCS dataset presents a more challenging retrieval benchmark dataset. Furthermore, the average audio duration and average length of the text descriptions in SOUNDDESCS is significantly higher than for AUDIOCAPS or CLOTHO. The word length distributions for these datasets can be seen in Fig. 3. Interestingly, for the CLOTHO and SOUNDDESCS dataset which contain audio and descriptions with varied lengths, there is no strong correlation between the audio and description lengths.

Lastly, the descriptions in the SOUNDDESCS dataset have a larger vocabulary with respect to nouns, verbs, adjectives used, compared to the AUDIOCAPS and CLOTHO datasets
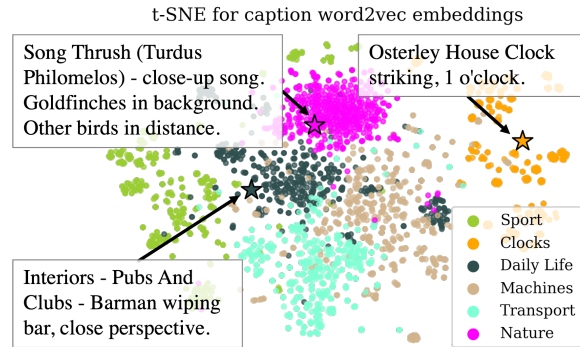

Fig. 5: t-SNE visualisation for word2vec description embeddings on SOUNDDESCS. Example descriptions corresponding to embeddings marked with a star are shown in text boxes.

(see Fig. 4). In particular, the descriptions in SOUNDDESCS contain almost 4000 distinct nouns as apposed to around 2000 in AUDIOCAPS and CLOTHO. This is consistent with the wide array of topics (Fig. 1), which reflects a high diversity of environments in which the sound effects were recorded, and thus hugely varied sources of sounds are identified in the descriptions. A large contributor to this diversity in nouns is the high proportion of Nature sounds (Fig. 1), for which species names are often specified (an example is shown in Fig. 5).

Further details about the related datasets that we use in this work can be found in Section V. A more in depth analysis of the SOUNDDESCS's text descriptions can be found in the Appendix.

C. Dataset examples

In Fig. 5 we visualise the distribution of the descriptions in SOUNDDESCS, showing the full descriptions for some examples. The averaged word2vec [64] vectors extracted from each description are embedded using t-SNE [65], and colour-coded according to the category (Fig. 1). We show 500 randomly chosen samples per class for 6 out of the 23 categories present in SOUNDDESCS. We can observe that the descriptions cluster smoothly according to the categories, and that the descriptions are fairly specific and of high quality, despite not always resembling full sentences.

D. Rights to use

Data collected for generating the new SOUNDDESCS dataset is protected by the BBC RemArc[4] licence. This allows the

---

[4]https://sound-effects.bbcrewind.co.uk/licensing

data to be used for non-commercial, personal or research purposes. We have released the list with urls along with code for downloading and constructing the SOUNDDESCS dataset with our proposed train/val/test split.

IV. METHODS, DATASETS, AND BENCHMARK

In this section, we first formulate the problem of audio retrieval with natural language queries. Next, we describe three cross-modal embedding methods that we adapt for the task of audio retrieval (Section IV-A). Finally, we describe the five datasets used in our experimental study (Section IV-B), and the three benchmarks that we propose for evaluating performance on the audio retrieval task (Section IV-C).

Problem formulation. Given a natural language query (*i.e.* a written description of an audio event to be retrieved) and a pool of audio samples, the objective of text-audio (abbreviated to `t2a`) retrieval is to rank the audio samples according to their similarity to the query. We also consider the converse `a2t` task, viz. retrieving text with audio queries.

A. Methods

To tackle the problem of text-audio retrieval, we propose to learn cross-modal embeddings. Specifically, given a collection of N audio samples with corresponding textual descriptions, $\{(a_i, t_i) : i \in \{1, \ldots, N\}\}$, we aim to learn embedding functions, ψ_a and ψ_t, that project each audio sample a_i and text sample t_i into a shared space, such that $\psi_a(a_i)$ and $\psi_t(t_i)$ are close when the text describes the audio, and far apart otherwise. Writing s_{ij} for the cosine similarity of the audio embedding $\psi_a(a_i)$ and the text embedding $\psi_t(t_j)$, we learn the embedding functions by minimising a contrastive ranking loss [66]:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1, j \neq i}^{B} [m + s_{ij} - s_{ii}]_+ + [m + s_{ji} - s_{ii}]_+ \quad (1)$$

where B denotes the batch size, m the *margin* (set as a hyperparameter), and $[\cdot]_+ = max(\cdot, 0)$ the hinge function.

We consider three recent state-of-the-art frameworks for learning such embedding functions ψ_a and ψ_t: Mixture-of-Embedded Experts (MoEE) [2], Collaborative-Experts (CE) [6], and the Multi-modal Transformer (MMT) [7]. All three frameworks were originally designed for text-video retrieval and construct their video encoder from a collection of "experts" (features extracted from networks pre-trained for object recognition, action classification, sound classification, etc.) which are computed for each video.

MoEE and CE aggregate the experts along their temporal dimension with NetVLAD [67], and project them to a lower dimension via a self-gated linear map followed by L2-normalisation. Their text encoder first embeds each word token with word2vec [64] and aggregates the results with NetVLAD [67]. The result is projected by a sequence of self-gated linear maps (one for each expert) into a shared embedding space with the outputs of the video encoder. Finally, a scalar-weighted sum of the embedded experts in each joint space is used to compute the overall cosine similarity

between the video and text (see [2] for more details). CE adopts the same text encoder as MoEE and similarly makes use of multiple video experts. However, rather than projecting them directly into independent spaces against the embedded text, CE first applies a "collaborative gating" mechanism, which filters each expert with an element-wise attention mask that is generated with a small MLP that ingests all pairwise combinations of experts (see [6] for further details).

MMT, on the other hand, uses a multi-modal transformer encoder which refines the expert embeddings by passing through multiple multi-headed self-attention layers. Differently from the temporal aggregation of expert features used by MoEE and CE, expert features are instead passed as input directly to the transformer encoder. This allows to iteratively focus on the relevant information from each expert at multiple time steps by comparing content from different experts, rather than down-projecting the expert embeddings in a single step (as is the case for MoEE and CE). Each input expert uses an aggregated embedding which collects the information for that expert through the layers and serves as the output expert representation. The iterative attention across different modalities and time steps enables MMT to combine information across input experts, rather than comparing each expert individually to the text query embeddings. MMT leverages a pre-trained BERT model [68] to extract text embeddings. Gated embedding functions are learned to obtain a text embedding for each video expert. The similarity s_{ij} between the aggregated audio and text embeddings is computed as the weighted sum of the similarity between each expert and the text embedding. Further information about MMT can be found in [7].

To adapt the MoEE, CE, and MMT frameworks for audio retrieval, we use the same text encoder structures as used for video retrieval ψ_t. We build audio encoders ψ_a by mimicking the structure of their video encoders, replacing the "video experts" with "audio experts" (described in Sec. V).

B. Datasets

As the primary focus of our work, we study three *audio-centric datasets*—these are datasets which comprise audio streams (sometimes with accompanying visual streams) paired with natural language descriptions that focus explicitly on the content of the audio track. To explore differences between audio retrieval and video retrieval, we also consider two *visual-centric datasets*, that comprise audio and video streams paired with natural language which focus primarily (though not always exclusively) on the content of the video stream. Details of the five datasets we employ are given next.

1. SOUNDDESCS *(audio-centric)* consists of sounds sourced from the BBC Sound Effects webpage and annotated with free-form text descriptions. It contains 23,085 training, 4947 validation and 4947 test samples. [5] Further information on our newly introduced SOUNDDESCS dataset is provided in Section III.

2. AUDIOCAPS [8] *(audio-centric)* is a dataset of sounds with event descriptions which was introduced for the task

---

[5]The sample list is publicly available at the project page [69].

of audio captioning, with sounds sourced from the AudioSet dataset [10]. Annotators were provided the audio tracks together with category hints (and with additional video hints if needed). We use a subset of the data, excluding a small number of samples for which either: (i) the YouTube-hosted source video is no longer available, (ii) the source video overlaps with the training partition of the VGGSound dataset [70]. Filtering to exclude samples affected by either issue leads to a dataset with 49,291 training, 428 validation and 816 test samples.[5]

3. CLOTHO [9] *(audio-centric)* is a dataset of described sounds that was also introduced for the task of audio captioning, with sounds sourced from the Freesound platform [11]. During labelling, annotators only had access to the audio stream (i.e. no visual stream or meta tags) to avoid their reliance on contextual information for removing ambiguity that could not be resolved from the audio stream alone. The descriptions are filtered to exclude transcribed speech. The publicly available version of the dataset includes a `dev` set of 2893 audio samples and an `evaluation` set of 1045 audio samples. Every audio sample is accompanied by 5 written descriptions. We used a random split of the `dev` set into a training and validation set with 2,314 and 579 samples, respectively.

4. ACTIVITYNET-CAPTIONS [71] (visual-centric) consists of videos sourced from YouTube and annotated with dense event descriptions. It allocates 10,009 videos for training and 4,917 videos for testing (we use the public `val_1` split provided by [71]). For this dataset, descriptions also tend to focus on the visual stream.

5. QUERYD [72] (visual-centric) is a dataset of described videos sourced from YouTube and the YouDescribe [73] platform. It is accompanied by *audio descriptions* that are provided with the explicit aim of conveying the video content to visually impaired users. Therefore, the provided descriptions focus heavily on the visual modality. We use the version of the dataset comprising trimmed videos with 9,114 training, 1,952 validation, and 1,954 test samples.

C. Benchmark

To facilitate the study of the text-based audio retrieval task, we introduce the SOUNDDESCS dataset and propose to re-purpose the two *audio-centric* datasets described above, AUDIOCAPS and CLOTHO, to provide benchmarks for text-based audio retrieval. The approach is inspired by precedents in the vision and language communities, where datasets, such as [74], that were originally introduced for the task of video captioning, have become popular benchmarks for text-based video retrieval [2], [43], [6], [7].

V. EXPERIMENTS

In this section, we first compare text-audio retrieval (`t2a`) and audio-text retrieval (`a2t`) performance on audio-centric and visual-centric datasets. Next, we perform an ablation study on the contributions of different experts and present our baselines for the proposed AUDIOCAPS, CLOTHO, and SOUNDDESCS benchmarks. Finally we perform experiments to assess the influence of pre-training, audio segment duration and training dataset size, and give qualitative examples of

retrieval results. Throughout the section, we use the standard retrieval metrics: recall at rank k (R@k) which measures the percentage of targets retrieved within the top k ranked results (higher is better), along with the median (medR) and mean (meanR) rank. For all metrics, we report the mean and standard deviation of three different randomly seeded runs.

Implementation details. We use pre-trained feature extractors to obtain audio and visual expert features. To encode the audio signal, we use two pre-trained audio feature extractors[6] which we refer to as VGGish and VGGSound. We explain both in more detail in the following.

VGGish. These audio features are obtained with a VGGish model [75], trained for audio classification on the YouTube-8M dataset [76]. To produce the input for the VGGish model, the audio stream of each video is re-sampled to a 16kHz mono signal, converted to an STFT with a window size of 25ms and a hop size of 10ms with a Hann window, then mapped to a 64 bin log mel spectrogram. Finally, the features are parsed into non-overlapping 0.96s collections of frames (each collection comprises 96 frames, each of 10ms duration), which are mapped to a 128-dimensional feature vector.

VGGSound. These features are extracted using a ResNet-18 model [77] that has been pre-trained on the VGGSound dataset (model H) [70]. We modify the last average pooling layer to aggregate along the frequency dimension, but keep the full temporal dimension. This results in features of dimension t x 512, where t denotes the number of time steps.

For the results with visual experts in Table IV, we employed a subset of visual feature extractors used in [6] which we refer to as Inst, Scene, and R2P1D. We we will describe each of those in more detail below.

Inst. These features are extracted using a ResNeXt-101 model [78] that has been pre-trained on Instagram hashtags [79] and finetuned on ImageNet [80] for the task of image classification. Features are extracted from frames extracted at 25 fps, where each frame is resized to 224×224 pixels. Embeddings are 2048-dimensional.

Scene. Scene features are extracted from 224×224 pixel centre crops with a DenseNet-161 model [81] pre-trained on Places365 [82]. Embeddings are 2208-dimensional.

R2P1D. Features are extracted with a 34-layer R(2+1)D model [83] trained on IG-65M [84] which processes clips of 8 consecutive 112×112 pixel frames, extracted at 30 fps. Embeddings are 512-dimensional.

Training All models were trained using the contrastive ranking loss (Eqn. 1), with m set to 0.2 for CE and MoEE, and 0.05 for MMT (those parameters were taken from [6] and [7] respectively). CE and MoEE models were trained with a batch size of 128 for 20 epochs and the models that gave the best performance on the geometric mean of R@1, R@5, and R@10 were chosen as the final models. MMT was trained with a batch size of 32 for 50K steps.

For CE and MoEE, we used the Lookahead solver [85] in combination with RAdam [86] (implementation by [87]) with

---

[6]Since the AudioCaps test set is a subset of the AudioSet training set (unbalanced), we do not use audio experts pre-trained on AudioSet.

TABLE II: Audio retrieval on AUDIOCAPS, CLOTHO, and SOUNDDESCS, using the CE, MoEE, and MMT frameworks. Retrieval performance metrics for text to audio and audio to text retrieval reported are recall at rank k (R@1, R@5, R@10), and the median and mean rank (medR and meanR). Audio experts used are obtained from the VGGish model [10] and from a ResNet18 model pre-trained on VGGSound [70].

| Dataset | Text ⟹ Audio | | | | | | Audio ⟹ Text | | | | | |
	R@1 ↑	R@5 ↑	R@10 ↑	R@50 ↑	medR ↓	meanR ↓	R@1 ↑	R@5 ↑	R@10 ↑	R@50 ↑	medR ↓	meanR ↓
AUDIOCAPS												
CE	$23.6_{\pm0.6}$	$56.2_{\pm0.5}$	$71.4_{\pm0.5}$	$92.3_{\pm1.5}$	$4.0_{\pm0.0}$	$18.3_{\pm3.0}$	$27.6_{\pm1.0}$	$60.5_{\pm0.7}$	$74.7_{\pm0.8}$	$94.2_{\pm0.4}$	$4.0_{\pm0.0}$	$14.7_{\pm1.4}$
MoEE	$23.0_{\pm0.7}$	$55.7_{\pm0.3}$	$71.0_{\pm1.2}$	$93.0_{\pm0.3}$	$4.0_{\pm0.0}$	$16.3_{\pm0.5}$	$26.6_{\pm0.7}$	$59.3_{\pm1.4}$	$73.5_{\pm1.1}$	$94.0_{\pm0.5}$	$4.0_{\pm0.0}$	$15.6_{\pm0.8}$
MMT	$36.1_{\pm3.3}$	$72.0_{\pm2.9}$	$84.5_{\pm2.0}$	$97.6_{\pm0.4}$	$2.3_{\pm0.6}$	$7.5_{\pm1.3}$	$39.6_{\pm0.2}$	$76.8_{\pm0.9}$	$86.7_{\pm1.8}$	$98.2_{\pm0.4}$	$2.0_{\pm0.0}$	$6.5_{\pm0.5}$
CLOTHO												
CE	$6.7_{\pm0.4}$	$21.6_{\pm0.6}$	$33.2_{\pm0.3}$	$69.8_{\pm0.3}$	$22.3_{\pm0.6}$	$58.3_{\pm1.1}$	$7.0_{\pm0.3}$	$22.7_{\pm0.6}$	$34.6_{\pm0.5}$	$67.9_{\pm2.3}$	$21.3_{\pm0.6}$	$72.6_{\pm3.4}$
MoEE	$6.0_{\pm0.1}$	$20.8_{\pm0.7}$	$32.3_{\pm0.3}$	$68.5_{\pm0.5}$	$23.0_{\pm0.0}$	$60.2_{\pm0.8}$	$7.2_{\pm0.5}$	$22.1_{\pm0.7}$	$33.2_{\pm1.1}$	$67.4_{\pm0.3}$	$22.7_{\pm0.6}$	$71.8_{\pm2.3}$
MMT	$6.5_{\pm0.6}$	$21.6_{\pm0.7}$	$32.8_{\pm2.1}$	$66.9_{\pm2.0}$	$23.0_{\pm2.6}$	$67.7_{\pm3.1}$	$6.3_{\pm0.5}$	$22.8_{\pm1.7}$	$33.3_{\pm2.2}$	$67.8_{\pm1.5}$	$22.3_{\pm1.5}$	$67.3_{\pm2.9}$
SOUNDDESCS												
CE	$31.1_{\pm0.2}$	$60.6_{\pm0.7}$	$70.8_{\pm0.5}$	$86.0_{\pm0.2}$	$3.0_{\pm0.0}$	$63.6_{\pm2.2}$	$30.8_{\pm0.8}$	$60.3_{\pm0.3}$	$69.5_{\pm0.1}$	$85.4_{\pm0.2}$	$3.0_{\pm0.0}$	$63.2_{\pm0.6}$
MoEE	$30.8_{\pm0.7}$	$60.8_{\pm0.3}$	$70.9_{\pm0.5}$	$85.9_{\pm0.6}$	$3.0_{\pm0.0}$	$62.0_{\pm3.8}$	$30.9_{\pm0.3}$	$60.3_{\pm0.4}$	$70.1_{\pm0.3}$	$85.3_{\pm0.6}$	$3.0_{\pm0.0}$	$61.5_{\pm3.2}$
MMT	$30.7_{\pm0.4}$	$61.8_{\pm1.0}$	$72.2_{\pm0.8}$	$88.8_{\pm0.4}$	$3.0_{\pm0.0}$	$34.0_{\pm0.6}$	$31.4_{\pm0.8}$	$63.2_{\pm0.7}$	$73.4_{\pm0.5}$	$89.0_{\pm0.3}$	$3.0_{\pm0.0}$	$32.5_{\pm0.4}$

TABLE III: Audio retrieval on audio-centric and visual-centric datasets. Performance is strongest on the audio-centric SOUNDDESCS dataset and weakest on the visual-centric ACTIVITYNET-CAPTIONS (ACTNETCAPS) dataset.

| Dataset | Anno. Focus | Pool | Text ⟹ Audio | | Audio ⟹ Text | |
			R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
AUDIOCAPS [8]	audio	816	$18.5_{\pm0.3}$	$62.0_{\pm0.5}$	$20.7_{\pm1.8}$	$62.9_{\pm0.4}$
CLOTHO [9]	audio	1045	$4.0_{\pm0.2}$	$25.4_{\pm0.5}$	$4.8_{\pm0.4}$	$25.8_{\pm1.7}$
SOUNDDESCS	audio	4947	$25.4_{\pm0.6}$	$64.1_{\pm0.3}$	$24.2_{\pm0.3}$	$62.5_{\pm0.2}$
ACTNETCAPS [71]	visual	4917	$1.4_{\pm0.1}$	$8.5_{\pm0.2}$	$1.1_{\pm0.1}$	$7.9_{\pm0.0}$
QUERYD [72]	visual	1954	$3.7_{\pm0.2}$	$17.3_{\pm0.6}$	$3.8_{\pm0.2}$	$16.8_{\pm0.2}$

TABLE IV: The influence of different experts on AUDIOCAPS. A comparison of audio and visual experts (applied to the video from which the audio was sourced) using CE [6]. Audio features are significantly more effective than visual features (which nevertheless provide some complementary signal as can be seen when jointly using audio and visual features).

| Expert | Text ⟹ Audio/Video | | Audio/Video ⟹ Text | |
	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
Visual experts only				
Scene	$6.0_{\pm0.0}$	$35.6_{\pm0.8}$	$6.8_{\pm0.6}$	$31.9_{\pm1.3}$
Inst	$8.2_{\pm0.3}$	$46.2_{\pm0.5}$	$10.1_{\pm0.8}$	$41.3_{\pm0.6}$
R2P1D	$8.1_{\pm0.4}$	$45.8_{\pm0.2}$	$10.7_{\pm0.1}$	$43.4_{\pm1.9}$
Scene + Inst	$8.2_{\pm0.3}$	$47.1_{\pm0.2}$	$10.2_{\pm1.2}$	$41.5_{\pm1.3}$
Scene + R2P1D	$8.6_{\pm0.1}$	$47.4_{\pm0.2}$	$11.6_{\pm0.4}$	$43.5_{\pm0.8}$
R2P1D + Inst (CE-Visual)	$\mathbf{9.5_{\pm0.6}}$	$\mathbf{50.0_{\pm0.5}}$	$\mathbf{11.2_{\pm0.1}}$	$\mathbf{45.2_{\pm1.9}}$
Audio experts only				
VGGish	$18.5_{\pm0.3}$	$62.0_{\pm0.5}$	$20.7_{\pm1.8}$	$62.9_{\pm0.4}$
VGGSound	$22.4_{\pm0.3}$	$69.2_{\pm0.9}$	$27.0_{\pm0.9}$	$72.5_{\pm0.7}$
VGGish + VGGSound (CE-Audio)	$\mathbf{23.6_{\pm0.6}}$	$\mathbf{71.4_{\pm0.5}}$	$\mathbf{27.6_{\pm1.0}}$	$\mathbf{74.7_{\pm0.8}}$
Audio and visual experts				
CE-Visual + VGGish	$24.5_{\pm0.8}$	$74.9_{\pm1.0}$	$31.0_{\pm2.2}$	$78.8_{\pm1.2}$
CE-Visual + VGGSound	$27.6_{\pm0.2}$	$78.0_{\pm0.8}$	$32.7_{\pm0.9}$	$82.4_{\pm0.4}$
CE-Visual + *CE-Audio*	$\mathbf{28.0_{\pm0.5}}$	$\mathbf{80.4_{\pm0.3}}$	$\mathbf{35.8_{\pm0.6}}$	$\mathbf{83.3_{\pm0.6}}$

an initial learning rate of 0.01 and weight decay of 0.001. We use a learning rate decay for each parameter group with a factor of 0.95 every epoch. MMT was trained using Adam and a learning rate of 0.00005, which was decayed by a multiplicative factor 0.95 every 1K optimisation steps.

For the NetVLAD module in CE and MoEE, we used 20 VLAD clusters and one ghost cluster [88] for text, and 16 VLAD clusters for the audio features. On AUDIOCAPS, we used a maximum of 52 word tokens, a maximum of 10 time frames for VGGish and a maximum of 32 time frames for VGGSound features. For CLOTHO, we used a maximum of 21 word tokens, a maximum of 31 time frames for VGGish, and 95 VGGSound features per sample (95 for both audio feature experts for MMT). For SOUNDDESCS, the maximum number of word tokens was set to 46, and audio time frames used to 400 (both for VGGish and VGGSound).

Audio-centric vs. visual-centric queries. We first investigate audio retrieval with audio-centric and visual-centric queries (The audio-centric datasets contain audio-centric queries, and the video-centric datasets contain video-centric queries.). For this experiment, we use CE with a single expert (VGGish audio features). In Table III, we observe that performance is strongest on the audio-centric datasets SOUNDDESCS and AUDIOCAPS, and it is weakest overall on the visual-centric ACTIVITYNET-CAPTIONS dataset. This is expected, since visual-centric queries contain information that is not captured in the audio data. We note that the CLOTHO dataset is particularly challenging, with performance weaker (accounting for pool size) than the visual-centric QUERYD dataset. We hypothesise that this is for two reasons: (1) the significantly smaller training set size of CLOTHO, compared to all other

datasets, (2) CLOTHO was constructed such that the audio tag distribution resulted in varied audio content, making it a potentially more difficult benchmark. We note, however, that the QUERYD experiments suggest that computationally efficient video retrieval using only the audio stream can still be obtained, although at a lower accuracy.

Ablation study. We next conduct an ablation study to investigate the effectiveness of different audio and visual experts for audio retrieval on the AUDIOCAPS dataset. We perform this experiment on the AUDIOCAPS dataset, since the SOUND-DESCS and CLOTHO datasets are audio-only datasets which implies that we cannot use any visual experts. We present the results in Table IV, where we observe that audio experts significantly outperform visual experts (pre-trained models for visual tasks like scene classification, which we compute from the video from which the audio was sourced). We note that the combination of audio and visual experts performs strongest overall, suggesting that the audio-centric queries contain information that is more accessible from the visual modality. The strongest audio-only retrieval is achieved by combining VGGish and VGGSound features—we therefore adopt this setting for the remaining experiments.

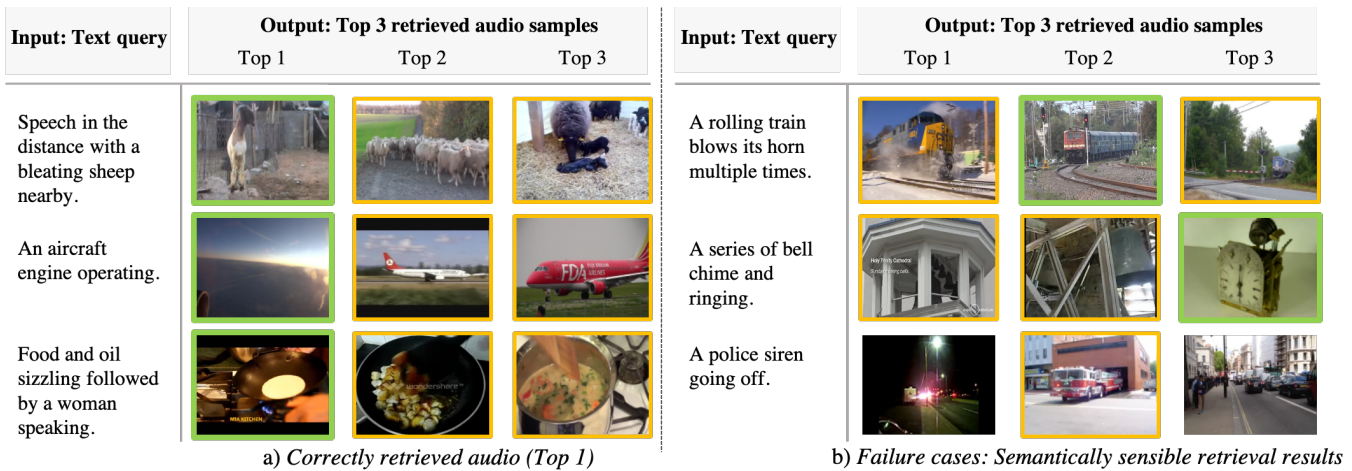Additionally, we experimented with using speech features

Input: Text query	Output: Top 3 retrieved audio samples			Input: Text query	Output: Top 3 retrieved audio samples		
	Top 1	Top 2	Top 3		Top 1	Top 2	Top 3
Speech in the distance with a bleating sheep nearby.				A rolling train blows its horn multiple times.			
An aircraft engine operating.				A series of bell chime and ringing.			
Food and oil sizzling followed by a woman speaking.				A police siren going off.			

a) *Correctly retrieved audio (Top 1)* — b) *Failure cases: Semantically sensible retrieval results*

Fig. 6: Qualitative results. Text-based audio retrieval results on AUDIOCAPS using CE with VGGish and VGGSound features. For an input text query, we visualise the top 3 retrieved audio samples using a video frame from the corresponding videos (the audio can be heard at the project webpage [69]). We mark the audio samples which correspond to the query with green boxes. Successful retrievals are shown in a), failures in b). Note, in particular, the examples in b), where the model's top 1 retrieved audio is not the correct one, but the retrieved results nevertheless sound reasonable (visually convincing results are marked with yellow boxes).

TABLE V: Pre-training for audio retrieval. Text-audio and audio-text retrieval results for CE [6] with VGGish and VGGSound features on the proposed SOUNDDESCS, AUDIO-CAPS, and CLOTHO retrieval benchmarks. Pre-training on AUDIOCAPS improves the performance on CLOTHO, and pre-training on SOUNDDESCS slightly boosts the performance on AUDIOCAPS.

	Text $\implies$ Audio		Audio $\implies$ Text	
Pre-training	R@1 ↑	R@10 ↑	R@1 ↑	R@10 ↑
AUDIOCAPS				
None	$23.6_{\pm 0.6}$	$71.4_{\pm 0.5}$	$27.6_{\pm 1.0}$	$74.7_{\pm 0.8}$
SOUNDDESCS	$24.6_{\pm 0.1}$	$72.2_{\pm 0.8}$	$27.8_{\pm 0.6}$	$75.2_{\pm 0.4}$
CLOTHO				
None	$6.7_{\pm 0.4}$	$33.2_{\pm 0.3}$	$7.0_{\pm 0.3}$	$34.6_{\pm 0.5}$
AUDIOCAPS	$9.1_{\pm 0.3}$	$39.7_{\pm 0.4}$	$11.1_{\pm 1.1}$	$39.6_{\pm 1.1}$
SOUNDDESCS	$6.4_{\pm 0.5}$	$32.5_{\pm 1.7}$	$6.1_{\pm 0.7}$	$31.4_{\pm 1.8}$
SOUNDDESCS				
None	$31.1_{\pm 0.2}$	$70.8_{\pm 0.5}$	$30.8_{\pm 0.8}$	$69.5_{\pm 0.1}$
AUDIOCAPS	$23.3_{\pm 0.7}$	$63.9_{\pm 0.5}$	$22.2_{\pm 0.4}$	$63.3_{\pm 0.3}$

(word2vec [64] encodings of speech-to-text transcriptions [89] of the audio stream). However, this did not improve the retrieval performance. Upon further investigation, we found that the audio captions and the spoken words in the AUDIOCAPS dataset do not have a significant overlap (corresponding to a METEOR [90] score of only 0.03 – perfect agreement between text sentences would give a score of 1).

Benchmark results. Incorporating the strongest combination of experts from the ablation study, we report our final baselines for text-audio and audio-text retrieval for three methods on the SOUNDDESCS, AUDIOCAPS, and CLOTHO datasets in Table II. We observe that MMT outperforms CE and MoEE on the AUDIOCAPS and CLOTHO datasets for both text-audio and audio-text retrieval. For SOUNDDESCS, all three models yield comparable results, with CE and MoEE being slightly

stronger than MMT.

We also report the performance after pre-training the CE model for retrieval on the AUDIOCAPS or SOUNDDESCS datasets and then fine-tuning on CLOTHO, AUDIOCAPS, or SOUNDDESCS in Table V. Here, we observe that pre-training on the SOUNDDESCS brings a slight boost for AUDIOCAPS and harms the performance on CLOTHO. This might be due to a larger domain gap between SOUNDDESCS and CLOTHO compared to AUDIOCAPS. Pre-training on AUDIOCAPS improves the performance on CLOTHO, but is not beneficial for SOUNDDESCS. Furthermore, we explored pre-training on CLOTHO and fine-tuning on the AUDIOCAPS and SOUND-DESCS datasets, but found negligible change in performance (likely due to the fact that the AUDIOCAPS training set is significantly larger than that of CLOTHO).

Qualitative results. The qualitative results in Fig. 6 show examples in which the CE model with VGGish and VGGSound expert modalities (CE-Audio) is used to retrieve audio with natural language queries. The retrieved results mostly contain audio that is semantically similar to the input text queries. Observed failure cases arise from audio samples sounding very similar to one another despite being semantically distinct (e.g. the siren of a fire engine sounds very similar to a police siren).

Influence of audio segment duration. Next, we investigate the influence of audio segment duration on retrieval accuracy on the SOUNDDESCS dataset. The retrieval accuracy for CE on different subsets of the test data according to their audio duration is shown in Fig. 7a. We show the performance for SOUNDDESCS audio files with a duration up to 30 seconds, those between 30 and 120 seconds, and those that are longer than 120 seconds. We observe that the retrieval performance is slightly weaker for longer audio segments than for those that last less than 30 seconds. However, the performance for very long audio files (longer than 120 seconds) is still solid.

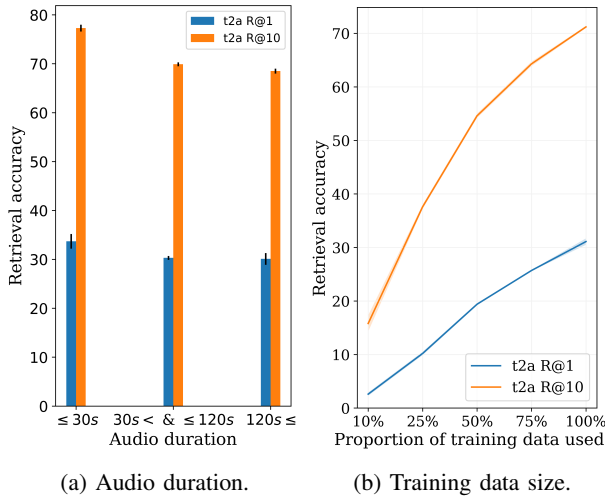(a) Audio duration.   (b) Training data size.

Fig. 7: The influence of audio duration and training data scale on the audio retrieval performance on SOUNDDESCS. Performance for the CE model is shown for the subsets of the SOUNDDESCS test set according to different audio durations in a), and for different proportions of available training data in b).

Influence of training scale. Finally, we present experiments using different proportions of the SOUNDDESCS dataset for training CE-Audio in Fig. 7b. As expected for deep learning frameworks, we observe that as more training data becomes available, the performance increases monotonically. We also observe that there is still clear room for improvement in terms of retrieval results simply by collecting additional training data, motivating further dataset construction work to support future research on this important task.

VI. CONCLUSION

We introduced the novel SOUNDDESCS dataset for natural language based audio retrieval. Furthermore, we proposed three benchmarks for natural language based audio retrieval on the CLOTHO, AUDIOCAPS, and SOUNDDESCS datasets, and provided baseline results by adapting strong multi-modal video retrieval methods. Our results show that these methods are relatively well-suited for the audio retrieval task, however there is room for improvement, as expected for an under-explored problem. We hope that our proposed benchmarks will facilitate the development of future audio search engines, and make this large fraction of the world's produced media available for public use.

ACKNOWLEDGMENT

APPENDIX

In this appendix, we provide additional information about the SOUNDDESCS dataset.

In Fig. 3, we presented a comparison of the description length distributions for the AUDIOCAPS, CLOTHO, and SOUNDDESCS datasets. Since descriptions in SOUNDDESCS can contain one or more sentences, we also provide the distribution of sentence lengths in Fig. 8.
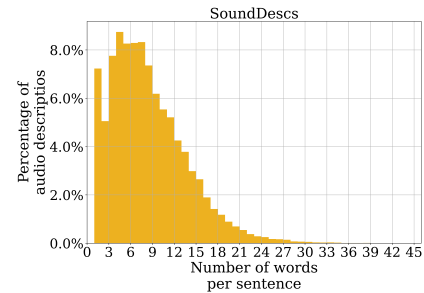

Fig. 8: Sentence length distribution for SOUNDDESCS.

To provide further insights into the text descriptions in the SOUNDDESCS dataset, Figures 9 and 10 show the distribution of unique nouns and verbs per description (0 unique nouns/verbs indicates that none were present). SOUNDDESCS contains noticeably more descriptions without any nouns and/or any verbs than AUDIOCAPS and CLOTHO. There are many Nature instances in SOUNDDESCS which contain names of species of birds that are labelled as Proper Nouns instead of Nouns. Descriptions shown to not contain any verbs are either wrongly tagged, or they do not describe actions, e.g. *Fountains: Rome - Sound of fountains, with street atmosphere.*
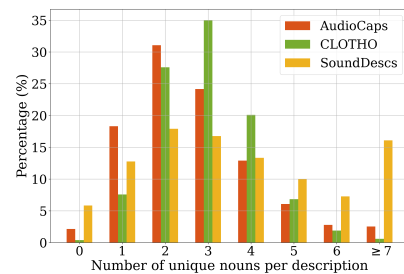

Fig. 9: Distribution of unique nouns for descriptions in the SOUNDDESCS, AUDIOCAPS, and CLOTHO datasets.
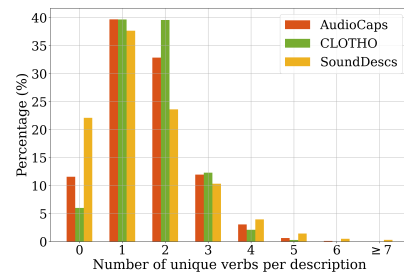

Fig. 10: Distribution of unique verbs for descriptions in the SOUNDDESCS, AUDIOCAPS, and CLOTHO datasets.

REFERENCES

[1] J. Dong, X. Li, and C. G. Snoek, "Word2visualvec: Image and video to sentence matching by visual feature prediction," *arXiv:1604.06838*, 2016.

[2] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv:1804.02516*, 2018.

[3] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proc. ACM ICMR*, 2018.

[4] P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde, and B. Raj, "Content-based representations of audio using siamese neural networks," in *Proc. ICASSP*, 2018.

[5] I. Lallemand, D. Schwarz, and T. Artières, "Content-based retrieval of environmental sounds by multiresolution analysis," in *Proc. SMC*, 2012.

[6] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *Proc. BMVC*, 2019.

[7] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Proc. ECCV*, 2020.

[8] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. NACCL*, 2019.

[9] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. ICASSP*, 2020.

[10] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017.

[11] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACM Multimedia*, 2013.

[12] A.-M. Oncescu, A. S.. Koepke, J. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," in *INTERSPEECH*, 2021.

[13] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proc. IEEE ICME*, 2003.

[14] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *JASA*, 2007.

[15] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *Proc. ICASSP*, 2006.

[16] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, 2015.

[17] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017*, 2017.

[18] A. Mesaros et al., "TUT Acoustic scenes 2017, Dev. dataset," Mar 2017. [Online]. Available: https://doi.org/10.5281/zenodo.400515

[19] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *IEEE WASPAA*, 2015.

[20] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Multimedia*, 2015.

[21] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," *arXiv:1906.02975*, 2019.

[22] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *Proc. ICASSP*, 2018.

[23] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," *arXiv:1803.02353*, 2018.

[24] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM TASLP*, 2019.

[25] L. Ford, H. Tang, F. Grondin, and J. R. Glass, "A deep residual network for large-scale acoustic scene analysis." in *INTERSPEECH*, 2019.

[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, 2020.

[27] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *IEEE WASPAA*, 2017.

[28] "DCASE2020 challenge task 6: Automated audio captioning," 2020. [Online]. Available: http://dcase.community/challenge2020/task-automatic-audio-captioning

[29] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *Proc. ICASSP*, 2019.

[30] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. ICCV*, 2015.

[31] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[32] Y. Koizumi, Y. Ohishi, D. Niizumi, D. Takeuchi, and M. Yasuda, "Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval," *arXiv:2012.07331*, 2020.

[33] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," *arXiv:2102.11457*, 2021.

[34] A. Ö. Eren and M. Sert, "Audio captioning based on combined audio and semantic embeddings," in *IEEE ISM*, 2020.

[35] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," *arXiv preprint arXiv:2107.09817*, 2021.

[36] 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.

[37] J. T. Foote, "Content-based retrieval of music and audio," in *Multimedia Storage and Archiving Systems II*. International Society for Optics and Photonics, 1997.

[38] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, 1996.

[39] M. Helén and T. Virtanen, "Query by example of audio signals using euclidean distance between gaussian mixture models," in *Proc. ICASSP*, 2007.

[40] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Proc. ISCA*, 2012.

[41] P. Avgoustinakis, G. Kordopatis-Zilos, S. Papadopoulos, A. L. Syme-onidis, and I. Kompatsiaris, "Audio-based near-duplicate video retrieval with audio similarity learning," *arXiv:2010.08737*, 2020.

[42] S. Hou and S. Zhou, "Audio-visual-based query by example video retrieval," *Mathematical Problems in Engineering*, 2013.

[43] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *Proc. ICCV*, 2019.

[44] I. Croitoru, S.-V. Bogolin, Y. Liu, S. Albanie, M. Leordeanu, H. Jin, and A. Zisserman, "Teachtext: Crossmodal generalized distillation for text-video retrieval," in *ICCV*, 2021.

[45] J. Dong, X. Li, and C. G. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018.

[46] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," *ICCV*, 2021.

[47] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval," *arXiv preprint arXiv:2104.08860*, 2021.

[48] B. Elizalde, R. Badlani, A. Shah, A. Kumar, and B. Raj, "Nels-never-ending learner of sounds," *arXiv:1801.05544*, 2018.

[49] M. Slaney, "Semantic-audio retrieval," in *Proc. ICASSP*, 2002.

[50] ——, "Mixtures of probability experts for audio retrieval and indexing," in *Proc. IEEE ICME*, 2002.

[51] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proc. ACM ICMIR*, 2008.

[52] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds," in *Proc. DCASE Workshop*, 2018.

[53] ——, "Generating sound words from audio signals of acoustic events with sequence-to-sequence model," in *Proc. ICASSP*, 2018.

[54] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," *arXiv:1706.00932*, 2017.

[55] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. ECCV*, 2018.

[56] M. Yasuda, Y. Ohishi, Y. Koizumi, and N. Harada, "Crossmodal sound retrieval based on specific target co-occurrence denoted with weak labels." in *INTERSPEECH*, 2020.

[57] D. Surís, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto, "Cross-modal embeddings for video and audio retrieval," in *Proc. ECCVW*, 2018.

[58] D. Zeng, Y. Yu, and K. Oyama, "Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval," *ACM TOMM*, 2020.

[59] C. Jin, T. Zhang, S. Liu, Y. Tie, X. Lv, J. Li, W. Yan, M. Yan, Q. Xu, Y. Guan *et al.*, "Cross-modal deep learning applications: audio-visual retrieval," in *Proc. ICPR*, 2021.

[60] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. ECCV*, 2018.

[61] B. Elizalde, S. Zarar, and B. Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *Proc. ICASSP*, 2019.

[62] X. Xu, H. Dinkel, M. Wu, and K. Yu, "Text-to-audio grounding: Building correspondence between captions and sound events," in *Proc. ICASSP*, 2021.

[63] H. Tang, J. Zhu, Q. Zheng, and Z. Cheng, "Query-graph with cross-gating attention model for text-to-audio grounding," *arXiv preprint arXiv:2106.14136*, 2021.

[64] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, 2013.

[65] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, 2008.

[66] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, 2014.

[67] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. CVPR*, 2016.

[68] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[69] "Project page." [Online]. Available: https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/

[70] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *Proc. ICASSP*, 2020.

[71] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proc. ICCV*, 2017.

[72] A.-M. Oncescu, J. F. Henriques, Y. Liu, A. Zisserman, and S. Albanie, "QuerYD: A video dataset with high-quality textual and audio narrations," in *Proc. ICASSP*, 2021.

[73] Video Description Research and Development Center, "Youdescribe," 2013. [Online]. Available: http://youdescribe.ski.org

[74] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proc. CVPR*, 2016.

[75] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, 2016.

[76] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[78] R. Xu, C. Xiong, W. Chen, and J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proc. AAAI*, 2015.

[79] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. ECCV*, 2018.

[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[81] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017.

[82] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE PAMI*, 2017.

[83] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018.

[84] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proc. CVPR*, 2019.

[85] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, "Lookahead optimizer: k steps forward, 1 step back," in *NeurIPS*, 2019.

[86] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[87] "Ranger optimiser." [Online]. Available: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer

[88] Y. Zhong, R. Arandjelović, and A. Zisserman, "Ghostvlad for set-based face recognition," in *Proc. ACCV*, 2018.

[89] Google, "Speech-to-Text API," https://cloud.google.com/speech-to-text.

[90] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005.