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 SOUNDESCS 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.

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 SOUNDESCS dataset which contains paired sound and natural language and by introducing benchmarks for text-audio retrieval.

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, would be easier to search, facilitating

Manuscript received 11 August 2021; revised 14 December 2021 and 27 January 2022; accepted 29 January 2022. Date of publication 8 February 2022; date of current version 13 July 2023. The work of A. Sophia Koepke and Zeynep Akata was supported by the ERC under Grant 853489 - DEXIM, in part by the BMBF under Grant FKZ: 01IS18039A. The work of Andreea-Maria Oncescu was supported by the EPSRC DTG Studentship. The work of João F. Henriques was supported by the Royal Academy of Engineering under Grant RP2018191/18163. The work of Samuel Albanie was supported by EPSRC EP/T028572/1 Visual AI. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin. (A. Sophia Koepke and Andreea-Maria Oncescu contributed equally to this work.) (Corresponding authors: A. Sophia Koepke; Andreea-Maria Oncescu.)

A. Sophia Koepke is with the Explainable Machine Learning Group at the University of Tübingen, 72074 Tübingen, Germany (e-mail: a-sophia.koepke@uni-tuebingen.de).

Zeynep Akata is with the Explainable Machine Learning Group at the University of Tübingen, 72074 Tübingen, Germany, and also with the Max Planck Institute for Intelligent Systems, 66123 Saarbrücken, Germany (e-mail: zeynep.akata@uni-tuebingen.de).

Andreea-Maria Oncescu and João F. Henriques are with the Visual Geometry Group at the University of Oxford, OX1 2JD Oxford, U.K. (e-mail: oncescu@robots.ox.ac.uk; joao@robots.ox.ac.uk).

Samuel Albanie is with the Department of Engineering at the University of Cambridge, CB21TN Cambridge, U.K. (e-mail: albanie@robots.ox.ac.uk).

Code, audio features for all datasets used, and the SOUNDESCS dataset are publicly available at https://github.com/akoepke/audio-retrieval-benchmark.

Digital Object Identifier 10.1109/TMM.2022.3149712

[Online]. Available: https://sounds.bl.uk

1520-9221 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
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 in the diversity of their audio content. Hence, we curated the novel SOUNDDSCS dataset, which was sourced from the BBC Sound Effects database.\(^2\) SOUNDDSCS contains text describing the sounds with significant variation with respect to the audio content and with a relatively large vocabulary used in the descriptions.

We introduce three new benchmarks for text-based audio retrieval, based on our proposed SOUNDDSCS 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. SOUNDDSCS 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 SOUNDDSCS 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 an ensemble of audio experts improves audio retrieval.

In summary, we make three contributions: (1) we introduce the SOUNDDSCS 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 SOUNDDSCS dataset for text-audio retrieval and provide an analysis of its characteristics (in Section III), (ii) we provide more extensive baselines for the audio retrieval task with more detailed ablations across datasets and an additional retrieval architecture (Section 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 to 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]–[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]–[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) [5], [37]–[39]. 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], [6], [7], [43], [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

\(^2\)[Online]. Available: https://sound-effects.bbcrewind.co.uk/
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 re-purposing them for the task of audio retrieval (described in more detail in Section 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]. In 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 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]–[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. 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 AudioCaps [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 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.

3The numbers provided for the AudioCaps dataset in Table I correspond to the subset which does not have an overlap with the VOGSound dataset.
### Comparative Overview of Sound-Language Datasets

| Dataset       | Text source       | language | duration(s) | #audios | #captions | max dur(s) | avg dur(s) | max #words | avg #words | train | val | test |
|---------------|-------------------|----------|-------------|---------|-----------|------------|------------|------------|------------|-------|-----|------|
| AUDIOCAPS     | Human captions   | English  | 135.01      | 50335   | 55512     | 10.08      | 9.84       | 52         | 8.52      | 49291 | 428 | 816  |
| CLOTHO        | Human captions   | English  | 22.55       | 3938    | 3938      | 30.00      | 22.44      | 21         | 11.32     | 2314  | 579 | 1045 |
| Audio Caption | Human captions   | Chinese  | 10.3        | 3707    | 3707      | N/A        | 10.00      | 54         | 11.14     | 3337  | -   | 371  |
| SOUNDDESCS    | Descriptions     | English  | 1060.40     | 32979   | 32979     | 4415.89    | 115.75     | 65         | 15.28     | 23088 | 4947| 4947 |

 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 datasets 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 use a larger vocabulary with respect to nouns, verbs, adjectives used, compared to the AUDIOCAPS and CLOTHO datasets (see Fig. 4). In particular, the descriptions in SOUNDDESCS contain almost 4000 distinct nouns as opposed 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.

### 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 SOUND-DECS. 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 SOUND-DECS dataset is protected by the BBC RemArc licence. This allows the 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 SOUND-DECS 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 $\psi_a$ and $\psi_t$, 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]:

$$L = \frac{1}{B} \sum_{i=1, j \neq i} \left[ m + s_{ij} - s_{ii} \right]_+ + \left[ m + s_{ji} - s_{ii} \right]_+$$

(1)

where $B$ denotes the batch size, $m$ the margin (set as a hyper-parameter), and $[\cdot]_+$ the hinge function.

We consider three recent state-of-the-art frameworks for learning such embedding functions $\psi_a$ and $\psi_t$: 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-head 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 $\psi_t$. We build audio encoders $\psi_a$ by mimicking the structure of their video encoders, replacing the “video experts” with “audio experts” (described in Section 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. Further information on our newly introduced SOUNDDESCS dataset is provided in Section III.

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

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]. 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, AUDIOPACS and CLOTHO, to provide benchmarks for text-based audio retrieval. The approach is inspired by predecessors 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], [6], [7], [43].

### 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 base-lines for the proposed AUDIOPACS, 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 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-8 M dataset [76]. To produce the input for the VGGish model, the audio stream of each video is re-sampled to a 16 kHz mono signal, converted to an STFT with a window size of 25 ms and a hop size of 10 ms with a Hanning window, then mapped to a 64 bin log mel spectrogram. Finally, the features are parsed into non-overlapping 0.96 s collections of frames (each collection comprises 96 frames, each of 10 ms 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 $t \times 512$, 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 will describe each of those in more detail below.

**Inst:** These features are extracted using a ResNetXt-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 obtained from frames extracted at 25 fps, where each frame is resized to $224 \times 224$ pixels. Embeddings are 2048-dimensional.

---

5The sample list is publicly available at the project page [69].
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 MoOE, 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 50 K steps.

For CE and MoEE, we used the Lookahead solver [85] in combination with RAdam [86] (implementation by [87]) with 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 1 K 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 SOUNDSCES, 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 SOUNDSCES and AUDIOCAPS, and it is weakest overall on the visual-centric ACTIVITYNET-CAPCTIONS 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 QUERDY 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 QUERDY 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 SOUNDDESCS 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 (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 dataset for both text-audio and audio-text retrieval. For SOUNDDESCS and CLOTHO, all three models yield comparable results.

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 SOUNDSCS brings a slight boost for AUDIOCAPS and harms the performance on CLOTHO. This might be due to a larger domain gap between SOUNDSCS and CLOTHO compared to AUDIOCAPS. Pre-training on AUDIOCAPS improves the performance on CLOTHO, but is not beneficial for SOUNDSCS. Furthermore, we explored pre-training on CLOTHO and fine-tuning on the AUDIOCAPS and SOUNDSCS 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 SOUNDSCS dataset. The retrieval accuracy for CE on different subsets of the test data according to their audio duration is shown in Fig. 7(a). We show the performance for SOUNDSCS audio files with a duration of 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.

Influence of training scale: Finally, we present experiments using different proportions of the SOUNDSCS dataset for training CE-Audio in Fig. 7(b). 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 SOUNDSCS dataset consisting of paired audio and text descriptions. Furthermore, we proposed three benchmarks for natural language based audio retrieval on the CLOTHO, AUDIOCAPS, and SOUNDSCS 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

The authors would like to thank A. Zisserman for suggestions. Samuel Albanie would also like to thank Z. Novak and S. Carlson for support.

APPENDIX

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

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

To provide further insights into the text descriptions in the SOUNDSCS dataset, Figs. 9 and 10 show the distribution of unique nouns and verbs per description (0 unique nouns/verbs indicates that none were present). SOUNDSCS contains noticeably more descriptions without any nouns and/or any verbs than AUDIOCAPS and CLOTHO. There are many Nature instances in SOUNDSCS 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.
et al., “Jointly discovering visual objects and spoken words from raw sensory input,” in Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 649–665.

M. Yasuda, Y. Ohishi, Y. Koizumi, and N. Harada, “Crossmodal sound retrieval based on specific target co-occurrence denoted with weak labels,” in Proc. InterSpeech, 2020, pp. 1446–1450.

D. Suris, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto, “Cross-modal embeddings for video and audio retrieval,” in Proc. Eur. Conf. Comput. Vis. Workshops, 2018, pp. 711–716.

D. Zeng, Y. Yu, and K. Oyama, “Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 16, no. 3, pp. 1–23, 2020.

C. Jin et al., “Cross-modal deep learning applications: Audio-visual retrieval,” in Proc. Int. Conf. Pattern Recognit., 2021, pp. 301–313.

A. Nagrani, S. Albanie, and A. Zisserman, “Learnable pins: Cross-modal embeddings for person identity,” in Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 71–88.

B. Elizalde, S. Zarar, and B. Raj, “Cross modal audio search and retrieval with joint embeddings based on text and audio,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 4095–4099.

X. Xu, H. Dinkel, M. Wu, and K. Yu, “Text-to-audio grounding: Building correspondence between captions and sound events,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 606–610.

H. Tang, J. Zhu, Q. Zheng, and Z. Cheng, “Query-graph with cross-gating attention model for text-to-audio grounding,” 2021, arXiv:2106.14136.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, arXiv:1301.3781.

L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 2579–2605, 2008.

R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, “Grounded compositional semantics for finding and describing images with sentences,” Trans. Assoc. Comput. Linguistics, vol. 2, pp. 207–218, 2014.

R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN architecture for weakly supervised place recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5297–5307.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

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

H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “VggSound: A large-scale audio-visual dataset,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 721–725.

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. CarlosNiebles, “Dense-captioning events in videos,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 706–715.

A.-M. Onesescu, J. F. Henriques, Y. Liu, A. Zisserman, and S. Albanie, “QuerYD: A video dataset with high-quality textual and audio narrations,” in Proc. Int. Conf. Acoust., Speech Signal Process., 2021, pp. 2265–2269.

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

J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5288–5296.

S. Hershey et al., “CNN architectures for large-scale audio classification,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 131–135.

A. Abu-El-Haija et al., “YouTube-8M: A large-scale video classification benchmark,” 2016, arXiv:1609.08675.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015, arXiv:1512.03385.

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 Conf. Artif. Intell., 2015, pp. 2346–2352.

D. Mahajan et al., “Exploring the limits of weakly supervised pretraining,” in Proc. IEEE Eur. Conf. Comput. Vis., 2018, pp. 181–196.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700–4708.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

D. Tran et al., “A closer look at spatiotemporal convolutions for action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6450–6459.

D. Gadiyaram, D. Tran, and D. Mahajan, “Large-scale weakly-supervised pre-training for video action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 12046–12055.

M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, “Lookahead optimizer: K steps forward, 1 step back,” in Proc. Neural Inf. Process. Syst., 2019, pp. 9597–9607.

L. Liu et al., “On the variance of the adaptive learning rate and beyond,” 2019, arXiv:1908.03265.

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

Y. Zhong, R. Arandjelovic, and A. Zisserman, “Ghostvlad for set-based face recognition,” in Proc. Asian Conf. Comput. Vis., 2018, pp. 35–50.

Google, “Speech-to-text API.” [Online]. Available: https://cloud.google.com/speech-to-text

S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in Proc. ACL Workshop, 2005, pp. 65–72.