Audio-visual Generalised Zero-shot Learning with Cross-modal Attention and Language

Otniel-Bogdan Mercea¹, Lukas Riesch¹,², A. Sophia Koepke¹, Zeynep Akata¹,³,⁴

¹University of Tübingen ²Robert Bosch GmbH ³MPI for Informatics ⁴MPI for Intelligent Systems

{otniel-bogdan.mercea, a-sophia.koepke, zeynep.akata}@uni-tuebingen.de
lukas.riesch@de.bosch.com

Abstract

Learning to classify video data from classes not included in the training data, i.e., video-based zero-shot learning, is challenging. We conjecture that the natural alignment between the audio and visual modalities in video data provides a rich training signal for learning discriminative multi-modal representations. Focusing on the relatively underexplored task of audio-visual zero-shot learning, we propose to learn multi-modal representations from audio-visual data using cross-modal attention and exploit textual label embeddings for transferring knowledge from seen classes to unseen classes. Taking this one step further, in our generalised audio-visual zero-shot learning setting, we include all the training classes in the test-time search space which act as distractors and increase the difficulty while making the setting more realistic. Due to the lack of a unified benchmark in this domain, we introduce a (generalised) zero-shot learning benchmark on three audio-visual datasets of varying sizes and difficulty, VGGSound, UCF, and ActivityNet, ensuring that the unseen test classes do not appear in the dataset used for supervised training of the backbone deep models. Comparing multiple relevant and recent methods, we demonstrate that our proposed AVCA model achieves state-of-the-art performance on all three datasets. Code and data are available at https://github.com/ExplainableML/AVCA-GZSL.

1. Introduction

Most zero-shot learning (ZSL) methods developed for image classification [5, 6, 60, 61, 72, 82] and action recognition [12, 13, 31, 84] only use unimodal input, e.g. images. However, humans leverage multi-modal sensory inputs in their everyday activities. Imagine the situation in which the sound of a dog barking is audible but the dog is visually occluded. In this case, we cannot understand the scene when relying on visual information alone. Using multiple modalities, such as vision and sound, allows to gather context and capture complementary information. Similarly, using both visual and audio information allows for a richer training signal for learning frameworks. This paper investigates the challenging task of (generalised) ZSL with multi-modal audio-visual data by leveraging the natural alignment of audio and visual information in videos.

Recently, [42, 54] have explored the task of zero-shot video recognition using multi-modal visual and audio information as inputs. However, the AudioSetZSL dataset [54] used for this, contains an overlap between the classes used for validation and testing. This results in learning stronger representations for classes overlapping with the training and validation sets (which covers all the classes in this dataset) and hinders the model’s capability to learn sufficiently gen-

Figure 1. Our audio-visual (generalised) ZSL framework aligns an audio-visual embedding with the corresponding textual label embedding via cross-modal attention. It can classify videos from previously unseen classes (e.g., elephant trumpeting) by predicting the class (red) whose textual label embedding (purple cross) is closest to the audio-visual embedding (blue star).
eralisable representations that allow information transfer. In real-world applications, such models perform well on seen classes, but poorly on previously truly unseen classes. In this work, we propose three benchmarks of varying size and difficulty curated from the VGGSound [18], UCF101 [62], and ActivityNet [25] datasets that could act as a unified and challenging playground for Generalised ZSL (GZSL) and ZSL research in the audio-visual domain. We suggest using audio and visual features extracted using SeLaVi [9] pretrained using self-supervision. Throughout this work, we use features that were obtained from training in a self-supervised fashion to reduce the information leakage from supervised pre-training to the zero-shot task which has been identified as a problem in other ZSL benchmarks [13].

We tackle the audio-visual generalised zero-shot learning task with our Audio-Visual Cross-Attention (AVCA) framework which is trained to align a rich learnt audio-visual representation with textual label embeddings. Our multi-stream architecture contains an audio and a visual branch which exchange information using cross-attention between the two modalities. AVCA is computationally lightweight and efficient since it uses audio and visual features extracted from pretrained networks as inputs instead of raw audio and image data. Our proposed framework is trained using multiple novel loss functions that are based on triplet losses and a regularisation loss that ensures that salient unimodal information is preserved in the learnt multi-modal representations. Our experiments show that AVCA achieves state-of-the-art performance on the three introduced benchmark datasets. We show that using multi-modal input data leads to stronger (G)ZSL performance than using unimodal data.

To summarise, our contributions are as follows: (1) We introduce three novel benchmarks for audio-visual (generalised) zero-shot learning curated from the VGGSound, UCF101, and ActivityNet datasets; (2) We propose AVCA, a cross-modal model for audio-visual (G)ZSL which leverages cross-modal attention between audio and visual information; (3) We show that AVCA yields state-of-the-art performance on all proposed audio-visual (G)ZSL benchmarks, outperforming the state-of-the-art unimodal and multi-modal zero-shot learning methods. Furthermore, we provide a qualitative analysis of the learnt multi-modal embedding space, demonstrating well-separated clustering for both seen and unseen classes.

2. Related Work

We review audio-visual learning, ZSL with image, video and audio data, and audio-visual ZSL.

Audio-visual learning. Audio-visual learning has enabled tremendous progress for numerous applications, such as for separating and localising sounds in videos [1, 4, 8, 17, 28, 51, 58, 67, 68, 81, 86, 87, 89], audio-visual synchronisation [16, 22, 24, 36], person-clustering in videos [14], (visual) speech and speaker recognition [2, 5, 46], spotting of spoken keywords [45, 57], audio synthesis using visual information [27, 29, 39, 40, 49, 64, 65, 88], and audio-driven image synthesis [34, 74]. Additionally, the natural alignment between audio and visual data in videos has been leveraged to learn powerful audio-visual representations for video or audio classification [7, 9, 10, 19, 20, 41, 47, 52, 53, 55, 78]. In contrast to those methods, we consider the ZSL setting for classification.

ZSL with images, videos and audio. Recently, numerous image-based generative ZSL methods have been proposed [50, 61, 72, 76, 77, 90, 92]. Their drawback is that the unseen classes need to be known a priori. In contrast, non-generative methods [5, 6, 26, 38, 60, 75, 82, 83] learn a mapping from input features to semantics of the classes (e.g. textual class label embeddings). Our AVCA model also learns to map its inputs to textual embeddings, but it leverages cross-attention between the audio and visual input modalities rather than using only visual inputs.

Video-based ZSL has been addressed by multiple recent works [12, 13, 30, 31, 59, 73, 84]. Using features extracted from pretrained networks results in computationally more feasible frameworks [12, 31, 73] than training end-to-end [13]. Our model also takes pre-extracted audio and visual features as inputs, resulting in a computationally efficient framework. In order to consider a pure ZSL setting when using pre-extracted features, the overlap between classes used for supervised pre-training of the feature extractors and unseen classes has to be removed [13, 30, 59]. This was not done in some of the previous works (e.g. [12, 31, 73, 91]). In contrast, we propose three benchmarks for audio-visual (G)ZSL on multi-modal audio-video data with no overlap between classes used for supervised pre-training and unseen classes.

Methods for zero-shot audio classification [79, 80] also used textual sound class embeddings (e.g. word2vec [43], BERT [23], or GloVe [56]) or descriptions. [21] investigate zero-shot music classification and tagging with word2vec embeddings and human-labeled attribute information (e.g. the presence or absence of musical instruments). For our AVCA model, we do not use any attribute information, but instead leverage the semantic alignment between audio and visual information in addition to textual label embeddings.

Audio-visual ZSL. Recently, [42, 54] proposed frameworks that consider the task of GZSL from audio-visual data. AVGZSLNet [42] uses late fusion on the AudioSetZSL dataset [54] to combine information from the two modalities. Instead, and also different to other audio-visual frameworks [66, 85] that use a simple dot-product operation for cross-attention, we use a transformer-based cross-attention mechanism. This allows for early and efficient sharing of multi-modal information, which is further encouraged by
our proposed loss functions. Furthermore, the AudioSet-ZSL dataset [54] does not include a validation split with unseen validation classes. Hence, [42, 54] select the GZSL hyperparameters directly on the (unseen) test classes. Furthermore, the AudioSetZSL dataset is comparatively small; those for the visual branch are similar. At test time, the class prediction is obtained by determining the class for which \( \theta_w \) is closest to \( \theta_v \).

### 3. Audio-Visual Cross Attention (AVCA)

The goal of audio-visual ZSL from video data is to learn to recognise videos from unseen classes (U), i.e. classes that were not seen during training. In the GZSL setting, the test set contains not only samples from unseen classes, but also from seen classes (S). This makes GZSL more challenging and more closely aligned with real-world learning tasks.

More formally, we denote the training set consisting only of samples from seen classes by \( S = (v_i^s, a_i^s, y_i^s)_{i \in \{1, \ldots, N\}} \), where \( v_i^s \) and \( a_i^s \) are visual and audio features respectively, \( y_i^s \) is the corresponding ground-truth class \( j \), and \( N \) is the number of samples in the training set. We refer to the class-level text embedding for class \( j \) as \( w_j^s \). The goal is to learn a function \( h : (v_i^s, a_i^s) \rightarrow w_j^s \) which can then also be applied to samples from unseen classes \( h(v_i^u, a_i^u) = w_j^u \), where \( (v_i^u, a_i^u, y_i^u) \in U \) for the set of test samples from unseen classes \( U = (v_i^u, a_i^u, y_i^u)_{i \in \{1, \ldots, M\}} \) with \( M \) samples.

### 3.1. Model Architecture

Our AVCA model architecture is visualised in Fig. 2. For easier readability, we dropped the subscripts \( i, j \), indicating the \( i \)-th dataset sample and the ground-truth class \( j \).

AVCA takes audio and visual features \( a, v \in \mathbb{R}^{k_{input}} \) as inputs which are extracted using pretrained feature extractors. Those are passed through two different encoder blocks \( A_{enc} \) and \( V_{enc} \) for the audio and visual modality respectively, giving embeddings

\[
A_{enc}(a) = \phi_a \quad \text{and} \quad V_{enc}(v) = \phi_v
\]  

(1)

with \( \phi_a, \phi_v \in \mathbb{R}^{k_f} \). The encoder blocks each consist of a sequence of two linear layers \( f_1^m, f_2^m \) for \( m \in \{a, v\} \), where \( f_1^m : \mathbb{R}^{k_{input}} \rightarrow \mathbb{R}^{k_{hid}} \) and \( f_2^m : \mathbb{R}^{k_{hid}} \rightarrow \mathbb{R}^{k_f} \). \( f_1^m, f_2^m \) are each followed by batch normalisation [33], a ReLU [48], and dropout [63] with dropout rate \( r_{enc} \).

**Cross-attention block.** We propose to use a cross-attention block to share information between the audio and visual representations \( \phi_a \) and \( \phi_v \). It consists of a multi-head self-attention layer, followed by a fully-connected feed-forward block. Similar to [71], we use a residual connection for the two layers, followed by layer normalisation [11].

The feed-forward blocks for the audio and visual branch each consist of a linear projection layer \( f_3^m : \mathbb{R}^{k_f} \rightarrow \mathbb{R}^{k_{hid}} \) for \( m \in \{a, v\} \), followed by GELU [32], dropout with dropout rate of \( r_{enc} \), another linear projection layer \( f_4^m : \mathbb{R}^{k_{hid}} \rightarrow \mathbb{R}^{k_f} \) for \( m \in \{a, v\} \) and finally a dropout with dropout rate of \( r_{enc} \). The outputs of the cross-attention block are \( \phi_a^{att}, \phi_v^{att} \in \mathbb{R}^{k_f} \).
A residual connection around the cross-attention block and subsequent projection blocks \( A_{proj} \) and \( V_{proj} \) give

\[
A_{proj}(\phi^a + \phi_k) = \theta_a \quad \text{and} \quad V_{proj}(\phi^v + \phi_k) = \theta_v,
\]

where \( \theta_a, \theta_v \in \mathbb{R}^{k_{proj}} \). The projection blocks each consist of a sequence of two linear layers \( f^m_n \) for \( m \in \{a, v\} \), where \( f^m_5 : \mathbb{R}^{k_{hid}} \rightarrow \mathbb{R}^{k_{hid}} \) and \( f^m_0 : \mathbb{R}^{k_{hid}} \rightarrow \mathbb{R}^{k_{proj}} \). \( f^m_5, f^m_0 \) are each followed by batch normalisation, a ReLU, and dropout with dropout rate \( r_{proj} \).

Furthermore, the word2vec class label embeddings \( w_j \) for class \( j \) are passed through the projection block \( W_{proj}(w_j) = \theta_w \), where \( \theta_w \in \mathbb{R}^{k_{proj}} \) (in Fig. 2 shown without the superscript \( j \)). \( W_{proj} \) consists of a sequence of one linear projection layer, batch normalisation, ReLU, and dropout with dropout rate \( r_{dec} \).

At test time, class predictions \( \hat{c} \) are obtained by determining the class \( c \) that corresponds to the textual class label embedding that is closest to the multi-modal representation \( \theta_t \) (in our experiments we found that using \( \theta_a \) gave slightly weaker results):

\[
\hat{c} = \arg\min_j(\|\theta_w^j - \theta_t\|_2).
\]

### 3.2. Loss Functions

We train our AVCA model using a loss function \( l \) consisting of a base triplet loss \( l_t \), a composite triplet and reconstruction loss \( l_c \), and a regularisation loss \( l_r \):

\[
l = l_t + l_c + l_r.
\]

We use the triplet loss function \( t(a, p, n) = \max(\|a - p\|_2 - \|a - n\|_2 + \mu) \), where \( a \) is the anchor embedding, \( p \) and \( n \) are embeddings for positive samples and negative samples respectively, and \( \mu \) is the margin hyperparameter. For triplet losses, we use the superscript \( t \) to denote positive samples that match the anchor and \( n \) for negative samples that do not semantically match the anchor. For all other losses, we only use matching pairs.

**Base triplet loss.** In our base triplet loss \( l_t \):

\[
l_t = t(\theta^+_a, \theta^+_w, \theta^-_a) + t(\theta^+_v, \theta^-_w) + t(\theta^-_a, \theta^+_w, \theta^-_w) + t(\theta^+_w, \theta^-_w, \theta^-_w),
\]

where \( \theta^+_m \) and \( \theta^-_m \) correspond to positive and negative samples respectively for \( m \in \{a, v, w\} \), ensuring that the projected visual and audio features \( \theta_a \) and \( \theta_v \) are aligned with the projected textual features \( \theta_w \). This is essential, since at test time, the proximity of \( \theta_t \) (which, despite being the output of the visual branch of AVCA, is a multi-modal embedding containing both audio and visual information) to \( \theta_w \) for different classes is used to determine the output class.

**Composite triplet and reconstruction loss.** Inspired by [42], we additionally use a composite triplet and reconstruction loss and explain its components in more detail below:

\[
l_c = l_{rec} + l_{ct} + l_w.
\]

We use a decoder \( D : \mathbb{R}^{k_{proj}} \rightarrow \mathbb{R}^{k_{w2v}} \), such that \( D(\theta_m) = \rho_m \) for \( m \in \{a, v, w\} \). \( D \) consists of a sequence of one linear projection layer, batch normalisation, a ReLU, and dropout with dropout rate \( r_{dec} \). We employ the mean squared error metric \( d(b, c) = \frac{1}{n} \sum_{i=1}^n (b_i - c_i)^2 \). The reconstruction loss \( l_{rec} \) can then be written as:

\[
l_{rec} = d(\rho_a, w) + d(\rho_v, w) + d(\rho_w, w).
\]

This ensures that AVCA is able to decode the pre-extracted textual label embeddings \( w \) from the embeddings \( \theta_a, \theta_v, \theta_w \).

The triplet loss \( l_{ct} \) is defined as follows:

\[
l_{ct} = t(\rho^+_a, \rho^+_v, \rho^-_a) + t(\rho^+_w, \rho^-_v, \rho^-_w),
\]

where \( \rho^+ \) and \( \rho^- \) correspond to positive and negative examples respectively. \( l_{ct} \) further encourages the decoded audio and visual features \( \rho_a, \rho_v, \rho_w \) to be aligned with the textual features \( \rho_w \) that were obtained using the same decoder (with shared weights). The third component \( l_w \) is similar to the base triplet loss in Eq. (5) and compares the audio and visual embeddings \( \theta_a, \theta_v \) to \( \theta_w \):

\[
l_w = t(\theta^+_a, \theta^-_a, \theta^+_w) + t(\theta^-_a, \theta^+_w, \theta^-_w) + t(\theta^+_w, \theta^-_w, \theta^-_w).
\]

**Regularisation loss.** The final component of our loss \( l \) consists of regularisation loss terms which directly encourage the alignment of the audio and visual embeddings with the text embeddings while preserving the information from their respective input modality. For this, we add two reconstruction blocks \( A_{rec} \) and \( V_{rec} \), such that \( \phi^a_{rec} = A_{rec}(\theta_a) \) and \( \phi^v_{rec} = V_{rec}(\theta_v) \). \( A_{rec} \) and \( V_{rec} \) each consist of a linear projection layer followed by batch normalisation, ReLU, and dropout with dropout rate \( r_{dec} \):

\[
l_r = d(\phi^a_{rec}, \phi_a) + d(\phi^v_{rec}, \phi_v) + d(\theta_v, \theta_a) + d(\theta_a, \theta_w).
\]

### 4. Experiments

We apply our AVCA model to audio-visual GZSL and ZSL for video classification. In this section, we first describe our proposed benchmark (Section 4.1). We discuss implementation details (Section 4.2), and then ablate the choice of different model components and loss functions (Section 4.5). Finally, we compare AVCA to state-of-the-art baseline methods for (G)ZSL (Section 4.3), and provide a detailed qualitative analysis of the learnt multi-modal embeddings (Section 4.4).
Table 1. Statistics for our VGGSound, UCF, and ActivityNet (G)ZSL datasets, showing the number (\#) of classes and videos in our splits (tr: train, v: validation, ts: test; S: seen, U: unseen).

| Dataset           | \# classes | \# videos       |
|-------------------|------------|-----------------|
| VGGSound-GZSL     | 276/51     | 2676/300        |
| UCF-GZSL          | 51/200     | 1534/999        |
| ActivityNet-GZSL  | 138/3174   | 353/276         |

4.1. Audio-Visual GZSL Benchmark

In this section, we propose three benchmark datasets for audio-visual GZSL curated from the VGGSound [18], UCF101 [62], and ActivityNet [25] datasets (summarised in Table 1), and introduce our training and evaluation protocol.

Dataset statistics. For our proposed audio-visual GZSL splits, we include classes contained in the Sports1M [35] dataset only in our seen subsets to allow the use of feature extractors pretrained on Sports1M without leakage of information to unseen classes.

Our GZSL splits for the three datasets consist of a training set (tr), a validation set which is divided into a subset with samples from seen classes (v(S)) and another one with unseen classes (v(U)). Finally, we provide a test set consisting of seen classes (ts(S)) and unseen classes (ts(U)). The training set and the seen validation subset share the same classes with a ratio of 0.9/0.1 with respect to the number of videos. The subsets \{tr \cup v(U) \cup v(S)\} and ts(S) share the same classes and were split to have a ratio of 0.9/0.1 with respect to the number of videos.

VGGSound [18] is a large audio-visual dataset with 309 classes and over 200k videos. The videos can be grouped into the 9 categories animals, home, music, nature, people, sports, tools, vehicle, and others. For our VGGSound-GZSL split, we exclude videos from the others category and all samples from v(U) and ts(U) that were used to train SeLaVi [9], resulting in 93,752 videos in 276 classes. The 42 classes that overlap with the Sports1M dataset are only used as training classes for GZSL.

UCF101 [62] is a video action recognition dataset which consists of over 13k videos in 101 classes. We use the subset of UCF101 which contains audio information. This results in a total of 6,816 videos for 51 classes. Previous (visual-only) methods repeatedly split the dataset into random seen and unseen classes. The 30 classes contained in the Sports1M dataset are not selected as unseen classes.

ActivityNet [25] is an action recognition dataset with 20k videos in 200 classes of varying duration. Again, we propose the ActivityNet-GZSL split ensuring that the 99 classes contained in the Sports1M dataset are not selected as unseen classes.

Training and evaluation protocol. We introduce a unified training and evaluation protocol for our GZSL benchmarks. We follow this protocol to train and test all models, including AVCA and the baselines that we compare to.

We propose a two-stage training and evaluation protocol for GZSL. In the first stage, we train the models on the training set (tr), using the subsets of seen validation classes (v(S)) and unseen validation classes (v(U)) to determine the GZSL parameters, for instance for calibrated stacking [15].

In the second training stage, we re-train the models on the training (tr) and full validation set \{v(S) \cup v(U)\} using the GZSL parameters determined during the first training stage. Our final models are then evaluated on the test set \{ts(S) \cup ts(U)\}. ts(S) contains samples from the same classes as the training classes with no overlap between training samples for the second stage and the test samples. In particular, there is no class overlap between v(U) and ts(U).

Evaluation metrics. Following [75], we propose to evaluate all models using the mean class accuracy. For GZSL, we evaluate the models on the full test set \{ts(S) \cup ts(U)\}, and report the averaged performance on the unseen (U) and seen (S) classes. Furthermore, we compute their harmonic mean \( H_M = \frac{2 \cdot U \cdot S}{U + S} \). We report the ZSL performance by evaluating only on the subset ts(U).

4.2. Experimental Setting

For each video, we use the self-supervised SeLaVi [9] framework pretrained on VGGSound [18] to extract audio and visual features for each second in a video. In our VGGSound-GZSL split, there is no overlap between videos in the unseen test and unseen validation sets and videos that were used for pre-training SeLaVi. We average the per-second features extracted using SeLaVi prior to the two-layer MLP heads to obtain 512-dimensional per-video audio and visual features. We provide additional results for using features extracted from audio and video classification networks in the supplementary material.

All networks were optimised for GZSL performance (HM) and we do not train separate networks for GZSL and ZSL. The training for the first stage was done for 50 epochs. We selected the number of training epochs for the second stage based on the GZSL performance on the validation set in the first stage. To eliminate the bias that the ZSL methods have towards seen classes, we used calibrated stacking [15] on the interval \([0, 3]\) with a step size of 0.2. For AVCA, \( k_{input} \) was set to 512 and the size of the word2vec embedding, \( k_{w2v} \), was set to 300. We used dropout rates \( r_{dec} = r_{enc} = r_{proj} = 0.5 / 0.2 / 0.3 \) for UCF-GZSL, 0.1 / 0.2 / 0.2 for Activity-GZSL, and 0.1 / 0.0 for VGGSound-GZSL. The layer dimensions were set to \( k_f = 300, k_{fhidd} = 512, k_{attlnhidd} = 64, \) and \( k_{proj} = 64 \). We used 3 heads for
self-attention. The loss margin hyperparameter, \( \mu \), was set to 1. We used a batchsize of 256 for UCF-GZSL and ActivityNet-GZSL, and 64 for VGGSound-GZSL. We used the Adam optimiser [37] with an initial learning rate of 0.001 which was reduced by a factor of 0.1 when the GZSL performance plateaued with a patience of 3 epochs.

### 4.3. Comparing with the State of the Art

**Compared methods.** In our benchmark study, we include four image-based state-of-the-art methods and one generative method for (G)ZSL which we adapt to take audio-visual features as inputs. For this, we concatenate the audio and visual features and use those as inputs instead of image features. Moreover, we compare to current state-of-the-art methods for audio-visual GZSL [42,54]. Here, we describe each of the methods that we compare to in more detail.

**ALE** [5] learns a linear mapping between the input features and the ground-truth embeddings, such that the projection of the input features is close to the ground-truth embedding for the corresponding class. For this, it uses a weighted approximate ranking objective [69]. **SJE** [6] computes the dot product between linearly mapped input features and the ground-truth embedding of all negative classes. The highest dot product for each example is chosen and then minimised. **DEVISE** [26] also computes the dot product between the output of a linear projection and the negative class embeddings and it minimises the sum of these dot products. **APN** [82] is the current non-generative state-of-the-art method for image-based ZSL. APN is based on the assumption that the ground-truth embeddings contain visual class attributes. Prototypes are used to map the attributes from the ground-truth embeddings to relevant locations in the image. **f-VAEGAN-D2** [77] is a generative ZSL method which learns to generate synthetic features for unseen classes. Then, a classifier is trained on real examples from seen classes and synthetic examples from unseen classes. **CJME** [54] proposed the task of audio-visual GZSL for video classification on the AudioSetZSL dataset. It embeds audio, video and text into a joint embedding space and uses proximity in the embedding space to select the classification output at test time. **AVGZSLNet** [42] builds on [54] and is the current state-of-the-art method for audio-visual GZSL for video classification. One of the main strengths of this method is its use of triplet losses to leverage information from negative examples.

**Results.** We compare our AVCA framework to recent methods for (G)ZSL in Table 2 on the VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL datasets. AVCA obtains the best results on all three datasets. On VGGSound-GZSL, AVCA obtains a HM of 6.31% for GZSL and a ZSL performance of 6.00% compared to 6.17% HM for CJME and a ZSL performance of 5.59% for DEVISE. On the UCF-GZSL dataset, our AVCA model outperforms SJE for GZSL with a performance of 27.15% compared to 26.50%, and we obtain a stronger ZSL performance of 20.01% compared to 18.93%. On ActivityNet-GZSL, AVCA outperforms APN, with a GZSL performance of 12.13% compared to 7.27%. For ZSL, AVCA is stronger than DEVISE with a score of 9.13% compared to 8.53%. It can be observed that in some cases U is higher than S. This is due to the use of calibrated stacking [15] as described in [44].

### 4.4. Qualitative Results

We present a qualitative analysis of the learnt multi-modal embeddings in Fig. 3. The t-SNE visualisations [70] for a subset of ActivityNet-GZSL classes show the differences between the audio and visual input features and the learnt multi-modal embeddings. We provide additional qualitative results for VGGSound-GZSL and UCF-GZSL in the supplementary material. It can be seen in Fig. 3a that the input audio features are not as well-separated and clustered as the visual features shown in Fig. 3b. However, the vi-

![Table 2. Evaluating our AVCA model and state-of-the-art audio-visual ZSL methods and adapted ZSL methods for GZSL and ZSL on the VGGSound, UCF, and ActivityNet (G)ZSL benchmarks. We report the mean class accuracy on the seen (S) and unseen (U) test classes, and their harmonic mean (HM) for GZSL performance. The ZSL performance is evaluated on the test subset from unseen classes.](image)
Figure 3. t-SNE visualisation for three seen (scuba diving, playing congas, wakeboarding) and two unseen (camel ride, making a cake) test classes from ActivityNet-GZSL, showing embeddings extracted with SeLaVi [9] for (a) audio and (b) visual features. (c) Learnt audio-visual embeddings of our model. Projected textual class label embeddings are visualised with a cross with black boundary.

Table 3. Influence of training AVCA with different modalities for GZSL and ZSL on the VGGSound-GZSL, UCF-GZSL and ActivityNet-GZSL datasets measuring the harmonic mean (HM) for GZSL and the mean class accuracy for ZSL. Using both modalities yields the strongest GZSL and ZSL performances.

| Model       | VGGSound-GZSL | UCF-GZSL | ActivityNet-GZSL |
|-------------|---------------|----------|------------------|
|             | HM     | ZSL     | HM     | ZSL     | HM     | ZSL     |
| Visual branch | 4.83   | 4.06    | 20.92  | 14.16   | 7.53   | 6.49    |
| Audio branch | 3.84   | 3.83    | 11.78  | 10.78   | 4.19   | 4.06    |
| AVCA        | 6.31   | 6.00    | 27.15  | 20.01   | 12.13  | 9.13    |

Table 4. Using different components of AVCA for GZSL and ZSL on VGGSound-GZSL, UCF-GZSL and ActivityNet-GZSL. Audio (Visual) with x-att uses the visual (audio) modality only for the cross-attention. W/o x-att optimises each branch in isolation and their output predictions are averaged. x-att denotes cross-attention.

| Model       | VGGSound-GZSL | UCF-GZSL | ActivityNet-GZSL |
|-------------|---------------|----------|------------------|
|             | HM     | ZSL     | HM     | ZSL     | HM     | ZSL     |
| W/o x-att   | 6.02   | 4.81    | 26.82  | 18.37   | 6.50   | 5.64    |
| Visual with x-att | 6.63   | 4.78    | 27.11  | 17.22   | 9.50   | 6.89    |
| Audio with x-att | 4.93   | 5.01    | 18.61  | 16.05   | 11.05  | 8.78    |
| AVCA        | 6.31   | 6.00    | 27.15  | 20.01   | 12.13  | 9.13    |

4.5. Ablation Analysis

Here, we analyse how different architectural choices and loss components for AVCA impact the performances on VGGSound-GZSL, ActivityNet-GZSL, and UCF-GZSL. ActivityNet-GZSL dataset. This confirms that the complementary information from the audio and visual inputs is highly beneficial for GZSL and ZSL for video classification. We provide the S/U performances for Table 3 in the supplementary material.

Evaluating the cross-modal attention block. Next, we investigate the effect of using our cross-modal attention block in Table 4. To obtain results without using cross-attention (W/o x-att), each branch is optimised individually. For evaluation, we compute the distances between the outputs of both branches and \( \theta_w \) for each class, and then average the distances computed by both branches. The GZSL and ZSL performances drop dramatically when not using the cross-attention block from 12.13% and 9.13% for AVCA to 6.50% and 5.64% for GZSL and ZSL scores respectively on the ActivityNet-GZSL dataset. The pattern is similar for VGGSound-GZSL and UCF-GZSL, confirming the importance of our cross-modal attention block for sharing information between the input modalities.

Furthermore, we compare optimising our full AVCA model to using only the visual (Visual with x-att) or only the audio branch (Audio with x-att) for training. Using only the visual branch entails removing \( A_{rec} \) and \( A_{proj} \) along with their associated losses from the audio branch but keeping the cross-attention. This experiment is repeated for the audio branch by removing the corresponding components from the visual branch. Jointly optimising both branches provides better results than using only...
| Model output | VGGSound-GZSL | UCF-GZSL | ActivityNet-GZSL |
|--------------|--------------|----------|-----------------|
|              | HM            | ZSL      | HM              | ZSL       | HM            | ZSL       |
| AVCA ($\theta_a$) | 5.18         | 4.87    | 25.98           | 18.25     | 12.54         | 9.23      |
| AVCA ($\theta_v$) | 6.31         | 6.00    | 27.15           | 20.01     | 12.13         | 9.13      |
| AVCA ($\theta_a$, $\theta_v$) | 5.90         | 5.42    | 25.78           | 19.30     | 12.17         | 8.95      |
| AVCA ($\min(\theta_a, \theta_v)$) | 6.10         | 5.36    | 25.86           | 18.39     | 12.45         | 9.08      |

Table 5. Influence of using the outputs of the audio and visual branches $\theta_a$ and $\theta_v$ separately, or using both jointly ($\theta_a$, $\theta_v$) for evaluation on VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL datasets. All models were trained with $\theta_a$ and $\theta_v$.

one of the branches on ActivityNet-GZSL and UCF-GZSL. On ActivityNet-GZSL, we obtain a GZSL performance of 12.13% compared to 11.05% and 9.50% for using only the audio and visual branches respectively. Interestingly, for the VGGSound-GZSL dataset, the Visual with x-att model yields a slightly stronger GZSL performance than our full AVCA model, with a HM of 6.63% compared to 6.31%. This is in line with the Audio branch performing worse than the Visual branch on VGGSound-GZSL (Table 3). However, the joint optimisation of AVCA gives the best results.

Evaluating different modalities as output. In Table 5, we investigate the effect of evaluating our full trained AVCA model using only the output features from the audio ($\theta_a$) or the visual ($\theta_v$) branch, or from both branches together (($\theta_a$, $\theta_v$) and $\min(\theta_a, \theta_v)$). For AVCA($\theta_a$, $\theta_v$), we compute the distance $|\theta_a - \theta_v| + |\theta_v - \theta_w|$. AVCA($\min(\theta_a, \theta_v)$) uses the embedding from the modality that has the smallest distance to a word embedding. The class corresponding to the closest text embedding resembles the class prediction.

Using the visual branch gives the strongest performance on VGGSound-GZSL/UCF-GZSL with a HM of 6.31% / 27.15% vs 5.18% / 25.98% for the audio branch. On ActivityNet-GZSL, the audio branch yields slightly better results (HM of 12.54% vs. 12.13% for the visual branch). Both AVCA($\theta_a$, $\theta_v$) and AVCA($\min(\theta_a, \theta_v)$) obtain lower scores that $\theta_v$. The best results (highest averaged HM) across all three datasets are produced when using the visual branch only. However, as the cross-attention block fuses the audio and visual modalities, both branches contain multimodal information from both input modalities.

Evaluating different loss functions. Finally, we analyse the impact of using different loss functions for training AVCA on the GZSL and ZSL performance in Table 6. We observe that using our full loss $l$ provides the strongest GZSL results (HM) on the UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL datasets by a large margin. On ActivityNet-GZSL, omitting $l_t$ for training our model ($l - l_t$) provides slightly stronger ZSL results than using our full loss $l$ with a mean class accuracy of 9.54% compared to 9.13%. However, the GZSL performance is significantly better when using $l$ with a HM of 12.13% compared to 8.39% when using $l - l_t$. Our loss ablations confirm that our strong overall performance on all three datasets is only obtained when training with our full proposed loss function.

4.6. Limitations and Discussion

Our proposed GZSL benchmark datasets pose an extremely challenging setting, since the underlying datasets span a wide variety of classes (e.g. including wakeboarding and making a cake for the ActivityNet dataset). Our AVCA leverages the varied audio-visual input information effectively, resulting in more robust GZSL performance than the related methods. However, AVCA uses temporally averaged audio-visual input information, and hence does not consider fine semantic details. Furthermore, our model relies on multi-modal input data and cannot be used when only one modality is available.

5. Conclusion

We introduced three new benchmarks for audio-visual (generalised) zero-shot learning for video classification on the VGGSound, UCF, and ActivityNet datasets. We proposed a framework for (G)ZSL from audio-visual data which learns to align the audio-visual embeddings with textual label embeddings. Furthermore, we provided baseline performances for seven (G)ZSL methods, and show that our model outperforms them for GZSL and ZSL on our new benchmarks. Finally, we provided a qualitative analysis of the learnt multi-modal embeddings. We hope that our proposed benchmarks will enable and encourage further research into audio-visual zero-shot learning.

Acknowledgements

This work was supported by BMBF FKZ: 01IS18039A, DFG: SFB 1233 TP 17, project number: 276693517, by the ERC (853489 - DEXIM), and by EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting O.-B. Mercea. The authors would like to thank M. Mancini for helpful suggestions and feedback.
References

[1] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In ECCV, 2020. 2

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. IEEE TPAMI, 2018. 2

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Ask is all you need: Cross-modal distillation for lip reading. In ICASSP, 2020. 2

[4] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In ECCV, 2020. 2

[5] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. IEEE TPAMI, 2015. 1, 2, 6

[6] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In CVPR, 2015. 1, 2, 6

[7] Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In NeurIPS, 2020. 2

[8] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In ECCV, 2018. 2

[9] Yuki M. Asano, Mandel Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In NeurIPS, 2020. 2, 5, 7

[10] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabelled video. In NeurIPS, 2016. 2

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 3

[12] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In BMVC, 2019. 1, 2

[13] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In CVPR, 2020. 1, 2

[14] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, body, voice: Video person-clustering with multiple modalities. In ICCV, 2021. 2

[15] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In ECCV, 2016. 5, 6

[16] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronisation in the wild. In BMVC, 2021. 2

[17] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In CVPR, 2021. 2

[18] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In ICASSP, 2020. 2, 5

[19] Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In CVPR, 2021. 2

[20] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In ACM MM, 2020. 2

[21] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. Zero-shot learning for audio-based music classification and tagging. In ISMIR, 2019. 2

[22] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In ACCV, 2016. 2

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019. 2

[24] Joshua P Ebeneze, Yongjun Wu, Hai Wei, Srimat Sethuraman, and Zongyi Liu. Detection of audio-video synchronization errors via event detection. In ICASSP, 2021. 2

[25] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In CVPR, 2015. 2, 5

[26] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In NeurIPS, 2013. 2, 6

[27] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In ECCV, 2020. 2

[28] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In ICCV, 2019. 2

[29] Shir Goldstein and Yael Moses. Guitar music transcription from silent video. In BMVC, 2018. 2

[30] Shreyank N Gowda, Laura Sevilla-Lara, Kiyoon Kim, Frank Keller, and Marcus Rohrbach. A new split for evaluating true zero-shot action recognition. arXiv preprint arXiv:2107.13029, 2021. 2

[31] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A cross-modal embedding approach to action learning. arXiv preprint arXiv:1901.00484, 2019. 1, 2

[32] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016. 3

[33] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICLR, 2015. 3

[34] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. In ICML, 2013. 2, 6

[35] Mikolov. Devise: A deep visual-semantic embedding model. In NeurIPS, 2013. 2, 6

[36] Andrew Senior, Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Frank Keller, and Marcus Rohrbach. A new split for evaluating true zero-shot action recognition. arXiv preprint arXiv:2107.13029, 2021. 2

[37] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A cross-modal embedding approach to action learning. arXiv preprint arXiv:1901.00484, 2019. 1, 2

[38] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016. 3

[39] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICLR, 2015. 3

[40] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. In IJCV, 2019. 2

[41] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014. 5

[42] Naji Khosravan, Shervin Ardeshir, and Rohit Puri. On attention modules for audio-visual synchronization. In NeurIPS Workshop, 2019. 2

[43] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In CorRR, 2015. 6
[75] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI*, 2018. 2, 5

[76] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 2

[77] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vae-gan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019. 2, 6

[78] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[79] Huang Xie, Okko Räsänen, and Tuomas Virtanen. Zero-shot audio classification with factored linear and nonlinear acoustic-semantic projections. In *ICASSP*, 2021. 2

[80] Huang Xie and Tuomas Virtanen. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. 2

[81] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *ACM MM*, 2020. 2

[82] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 1, 2, 6

[83] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *CVPR*, 2022. 2

[84] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV*, 2016. 1, 2

[85] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*, 2020. 2

[86] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 2

[87] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 2

[88] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *ICCV*, 2019. 2

[89] Lingyu Zhu and Esa Rahtu. V-slowfast network for efficient visual sound separation. In *WACV*, 2022. 2

[90] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 2

[91] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *CVPR*, 2018. 2

[92] Yizhe Zhu, Jianwen Xie, Bingchen Liu, and Ahmed Elgammal. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In *ICCV*, 2019. 2