Information Fusion 129 (2026) 103966

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Self-attention and cross-modal attention for audio-visual zero-shot learning

Jing Yang [a,b], Xiaoyong Li [a,*], Yuankai Wu [c], Yuling Chen [a], Xiaoli Ruan [a], Chengjiang Li [a], Qing Hou [d,*]

[a] Guizhou University, State Key Laboratory of Public Big Data, Guiyang, 550025, Guizhou, China
[b] Shanghai Jiao Tong University, Department of Computer Science and Engineering, Shanghai, 201100, Shanghai, China
[c] Sichuan University, National Key Laboratory of Fundamental Science on Synthetic Vision, Chengdu, 610065, Sichuan, China
[d] Guizhou Communication Industry Service Co, Ltd, Guiyang, 550025, Guizhou, China

ARTICLE INFO

Keywords:
Zero-shot learning
Audio-visual learning
Video classification
Self-attention
Cross-modal attention

ABSTRACT

Audio-visual generalized zero-shot learning is aimed at learning good representations from audio-visual data, enabling the recognition of unseen classes during testing. The existing embedding and generation methods have made significant progress. However, these methods do not fully extract the internal features of each modality. Moreover, there is insufficient information interaction between different modalities. To address these issues, we propose an audiovisual zero-shot learning method based on self-attention and cross-modal attention (SACMA). Specifically, we use a self-attention mechanism to obtain information within a single modality and a cross-modal cross-attention mechanism to capture the relationships between different modalities. To establish the connections between different modalities and minimize the gap between their features, we introduce a combined contrastive loss function and a cosine similarity loss function. We evaluated the proposed method on three benchmark datasets, VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL, and compared it with eleven state-of-the-art methods. Code and data available at https://github.com/ybyangjing/SACMA.

1. Introduction

Generalized zero-shot learning (GZSL) has become a hot research area in computer vision because it can generalize knowledge from known classes to unknown classes while retaining known class knowledge. In the early zero-shot learning approaches, single-modal data, such as images [1–7] and text [8], were typically utilized. However, relying solely on data derived from a single modality is not sufficient. For example, using both auditory and visual senses when one is watching a movie can enhance one's overall viewing experience. Moreover, audio-visual generalized zero-shot learning can alleviate the limitations of single-modal data and fuse auditory and visual data to obtain powerful representations.

Owing to the semantic differences between audio and visual information, samples belonging to the same category are represented differently by different modalities. Therefore, ensuring semantic consistency between auditory and visual information is a key issue to be solved in audio-visual zero-shot learning tasks. Currently, two types of commonly used solutions are available: embedding-based methods and generation-based methods. Embedding-based methods embed audio and visual information in a shared space and use loss functions to align the features of different modalities. Generative-based methods simulate the features of unknown classes and learn the feature differences between the known and unknown classes. However, these approaches ignore the importance of single-modal information, and different modalities exhibit insufficient information interaction. We found that video data are affected by environmental noise and various changes, such as background noise, illumination changes, and occlusion, which is important for understanding the semantic consistency between the auditory and visual modalities. Therefore, we believe that separately learning the key features contained in audio and video information is beneficial for the subsequent understanding of the semantic associations between auditory information and visual information.

To overcome the limitations of the above approaches, we propose a novel framework termed self-attention and cross-modal attention for audio-visual zero-shot learning. Our audio-visual generalized zero-shot learning method, which focuses on the connections between various modalities, is shown in Fig. 1. Compared with the previously developed methods, the audio-visual generalized zero-order learning method proposed in this study, which fuses intramodal information and intermodal correlations, can simultaneously focus on the information contained within a single modality and the relationships between modalities. Our

* Corresponding authors.
E-mail addresses: jyang23@gzu.edu.cn (J. Yang), gs.xiaoyongli22@gzu.edu.cn (X. Li), wuyk0@scu.edu.cn (Y. Wu), ylchen3@gzu.edu.cn (Y. Chen), xlruan@gzu.edu.cn (X. Ruan), cjli3@gzu.edu.cn (C. Li), houqing.gz@chinaccs.cn (Q. Hou).

https://doi.org/10.1016/j.inffus.2025.103966
Received 3 April 2025; Received in revised form 25 September 2025; Accepted 17 November 2025
Available online 20 November 2025

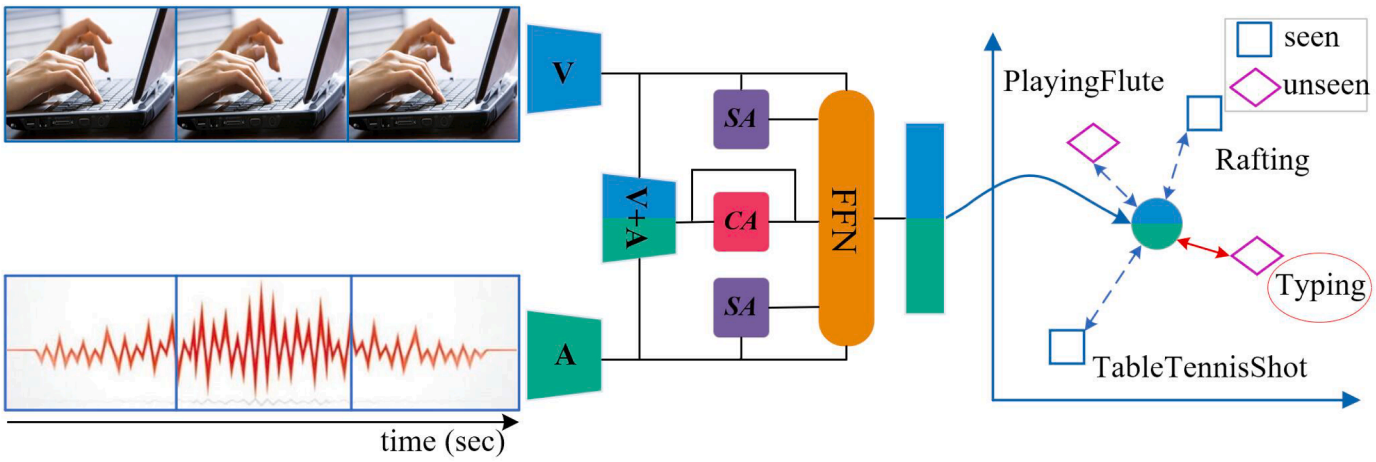


Fig. 1. Illustration of the proposed SACMA framework for audio-visual zero-shot learning. SACMA employs self-attention (SA) to capture intramodal audio and video information and cross-attention (CA) to model the semantic associations between modalities, aligning audio-visual and text embeddings. Knowledge transfer to unseen classes is achieved by predicting the text label embedding that is closest to the given audio-visual embedding.

method uses an attention mechanism to learn the respective internal features of audio and video data and a cross-modal cross-attention mechanism to learn the relationships between the audio and video modalities, allowing the model to more comprehensively utilize the information acquired from different modalities. Thus, the model can learn more general features from audio-visual data and improve its generalizability to unseen classes.

To better train our model, we construct a combined contrastive loss function and a cosine similarity loss. The combined contrastive loss is used to learn the relationships among the audio, video, and text modalities. Thus, the combination of different input modes is fully considered during the modeling process. By considering multimodal combinations, our method not only captures the information between each pair of modalities in a more comprehensive manner but also helps improve the adaptability of the model to complex correlation structures. This function not only enables the model to better learn and utilize the rich information contained in multimodal data but also significantly improves its ability to model the correlations of multimodal data in actual complex scenarios. The cosine similarity loss is used to measure the similarity between the representations learned by the model from samples belonging to the same category. Through this loss function, the model is encouraged to more effectively distinguish between similar samples during the learning process, thereby more accurately capturing the inherent similarities between samples. This function helps to better reflect the categorical associations between samples, providing stronger support enhancing for the generalizability and accuracy of the model. The main contributions of this article are as follows:

- We introduce a dual-input audiovisual GZSL framework, namely, SACMA, which includes both unimodal and multimodal inputs to ensure semantic consistency between data derived from different modalities.
- To extract features from a single modality and the interactions between different modalities, we use a self-attention (SA) mechanism and a cross-modal attention (CA), respectively.
- We construct a combined contrastive loss function and a cosine similarity loss to reduce the distances between different modalities and help the model align the representations of differ-ent modalities in a common multimodal space.
- A comprehensive experimental evaluation of our met-hod is conducted on three datasets: UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL. The proposed method achieves a 7.8 % improvement in GZSL performance and a 0.88 % improvement in zero-shot learning (ZSL) performance on the UCF-GZSL data-set.

2. Related work

2.1. Single-modal learning

Single-modal learning is an important research direction computer vision, focusing on processing and analyzing information from single-modal data (such as images and text). How to obtain powerful feature representations from single-modal data is a crucial challenge. The convolutional neural network (CNN) [9,10] utilizes local context information and translation invariance properties and is widely used in single-modal learning. The attention mechanism is a process of adaptive selection based on input features [11], which helps the model obtain powerful feature representations. In this section, we introduce the research progress on methods for mining intramodal information in single-modal learning.

CNN-based methods: An understanding of biological visual systems serves as the inspiration for the CNN design. The main purpose is to effectively capture local features through convolution operations and to achieve downsampling and position invariance of features through pooling operations. Since Krizhevsky et al. [12] proposed the AlexNet network, CNNs have rapidly emerged as a mainstream framework in computer vision for learning features within single-modal data. As the number of network layers has increased, VGGNet [13] and MSRNet [14] have also emerged. However, the continuous increase in the number of network layers is detrimental to practical applications. As a result, CNN lightweight frameworks such as MobileNet [15] and ShuffleNet [16] have emerged. By strengthening the performance of the convolution module, a series of new architectures have emerged, including the network in network (NiN) [17] and GoogLeNet [18] architectures. In addition, derivative models such as ResNet [19] and Inception ResNet [20] have also been introduced. Since data such as images lack temporal information compared to video data, Tran et al. [21] proposed a deep three-dimensional convolutional network trained on large-scale supervised video datasets to learn spatiotemporal features, helping the model better extract important features within single-modal data. Gao et al. [22] suggested TS-GCN, a dual-stream graph convolutional network framework designed to model relationships between actions and attributes, between actions, and between attributes and actions. TS-GCN can learn from single modalities and obtain more information from data.

Attention mechanism-based methods: Attention mechanisms simulate human perception, enable the model to focus on the salient parts of specific features, and are widely used in various single-modal learning tasks [23,24]. By introducing an attention mechanism, the network can learn autonomously and selectively by focusing on key information in a

single modality, thereby improving the model performance and generalizability. The soft attention mechanism developed by Chen et al. [25] and Max et al. [26] can be trained in an end-to-end manner for convolutional networks. By inputting the features extracted by convolution into the attention block, the model can focus on the internal features of the modality. Wang et al. [27] proposed a residual attention network to integrate soft attention into a rapidly developing feedforward network structure using an encoder-decoder attention module. Different attention modules capture different types of attention to guide the model to learn features in single-modal data. The self-attention mechanism is an attention mechanism originating from natural language processing (NLP) [28,29]. Because of its excellent performance in effectively capturing long-range dependencies and adaptability, the self-attention mechanism plays an important role in single-modal learning.

CNNs automatically learn hierarchical feature representations of single-modal data through structures such as convolutional layers and pooling layers and have achieved great success in single-modal tasks such as image classification [30–33] and object detection [34]. The self-attention mechanism establishes relationships between different positions within a single sequence, focuses on long-distance dependencies within the sequence, and attains good performance with respect to the long-distance dependencies contained in video data. Therefore, in this study, our method uses a self-attention mechanism to learn important audio and visual information from videos.

2.2. Audio-visual multimodal learning

In recent years, audio-visual multimodal learning has received increasing attention because it can simulate human cognitive methods of integrating multisensory input and provide richer and more comprehensive information. In audio-visual multimodal learning, capturing the connection between audio and video using the natural alignment between them is a challenging task. In this section, we introduce research progress on methods for capturing intermodal correlations in audio-visual multimodal learning.

Transformer-based audio-visual learning: The transformer model was initially successful in NLP and was later applied to computer vision. In audio-visual multimodal learning, transformers have been widely used in tasks such as video retrieval [35], audio-visual localization [36], and audio-visual source separation [37]. Cheng et al. [38] proposed a novel self-supervised framework for learning universal cross-modal representations from unlabeled videos and explored three different collaborative attention modules to focus on sound-related discriminative visual areas and to introduce interactions between them. Nagrani et al. [39] studied a variety of audio-visual fusion strategies for the multimodal bottleneck transformer (MBT), aiming to improve the traditional transformer architecture, such as early fusion and midstage fusion. Specifically, the MBT was proposed to limit the cross-modal flow to the later layer of the network. The layers adopt single-modal learning and focus on single-modal features through midterm fusion. In addition, the MBT limits cross-modal attention between tokens within the layer by introducing an attention bottleneck layer to capture more cross-modal features. Mercea et al. [40] proposed a temporal cross-attention framework for the audio-visual generalized zero-shot learning task that learns the relationship between different modalities through a cross-modal attention mechanism.

Non-transformer audio-visual learning: Owens et al. [41] used convolutional neural networks to predict whether a given pair of audio and video clips was temporally aligned in a self-supervised manner. The learned representations were then used to perform sound source localization and audio-visual action recognition. Arandjelovic et al. [42] introduced the audio-visual correspondence learning task. Here, training includes using visual and audio subnetworks to learn semantic the correspondence between audio and visual data. Gao et al. [43] proposed a multi-instance multilabel learning framework to solve the audio-visual source separation problem; in this framework, different audio components are extracted and associated with visual objects in videos.

Our method benefits from a transformer and adopts its cross-modal cross-attention mechanism to learn the correlation between audio and video. Compared with nontransformer-based methods, the use of transformer-based cross-attention mechanisms can enable one sequence to pay attention to another sequence, thereby achieving cross-modal associations between different sequences.

3. Methods

3.1. Problem definition

In the audio-visual ZSL task, the aim is to learn to recognize previously unseen (U) classes, i.e., classes that the model is not exposed to during training, from video data. In a more challenging GZSL scenario, the test set contains not only samples from unseen classes but also samples from seen (S) classes. Therefore, the model needs to have the ability to generalize knowledge from seen to unseen classes during testing while retaining knowledge about visible training classes. The model will be able to better simulate learning tasks in the real world. However, it is challenging for the model to cope with complex scenarios and generalizations. We divide the audio-visual dataset classes into seen and unseen classes, and the corresponding labels are denoted Y_S and Y_U, respectively. $Y_S \cap Y_U = \emptyset$. The dataset consists of audio features a_i,video features v_i and real labels y_i.

3.2. Overall framework

Self-attention mechanisms focus on information within modalities, while cross-attention mechanisms enable interactions between different modalities. To simultaneously focus on the information within audio and visual modalities and inter-modal relationships, we propose a novel framework based on self-attention and cross-modal attention. As shown in Fig. 2. The input data modalities of the SACMA framework are audio, visual, and textual features, i.e., a, v, and t, respectively. Notably, I_a and I_v are passed through the audio encoder A_{enc} and the video encoder V_{enc}, respectively, and a new input I_m is obtained by splicing I_a and I_v. I_a, I_v, and I_m are input into a transformer. Each transformer layer consists of a multi-head self-attention layer, two normalization layers, a fully connected feedforward layer, and residual connections. Self-attention is used to learn the internal features of the modality for I_a and I_v, and cross-modal attention is used to interact with the information seen in different modalities for I_m. The outputs of the model are projected to the multimodal common space through the projection modules A_{pro}, V_{pro}, and M_{pro}. t does not need to be input to the transformer, but it needs to be projected to the multimodal common space through T_{pro}.

3.3. Joint embedding

The input data of our multimodal transformer framework consist of three modal datasets: audio, video, and text. These features are extracted from the original data through pretraining. Two encoders, an audio encoder and a video encoder, denoted A_{enc} and V_{enc}, respectively, are set up by using A_{enc} and V_{enc} to project the audio and video features into the same feature dimension. These features preserve semantic information in the same dimensions as text features t.

$$I_a = A_{enc}(a), I_v = V_{enc}(v) \tag{1}$$

To allow the model to better learn and understand the information within a given modality and the correlations between data belonging to different modalities, we propose a dual-input method that uses single-modal and multimodal inputs. With single-modal inputs, the model is able to independently learn and understand the characteristics of different data modalities, gaining insights into the unique properties of each modality, such as the sound characteristics of audio and the visual content of video. This independent learning scheme helps the model more accurately capture the information of each modality, allowing it to better understand audio and video information in multimodal tasks and achieve a more comprehensive multimodal understanding.

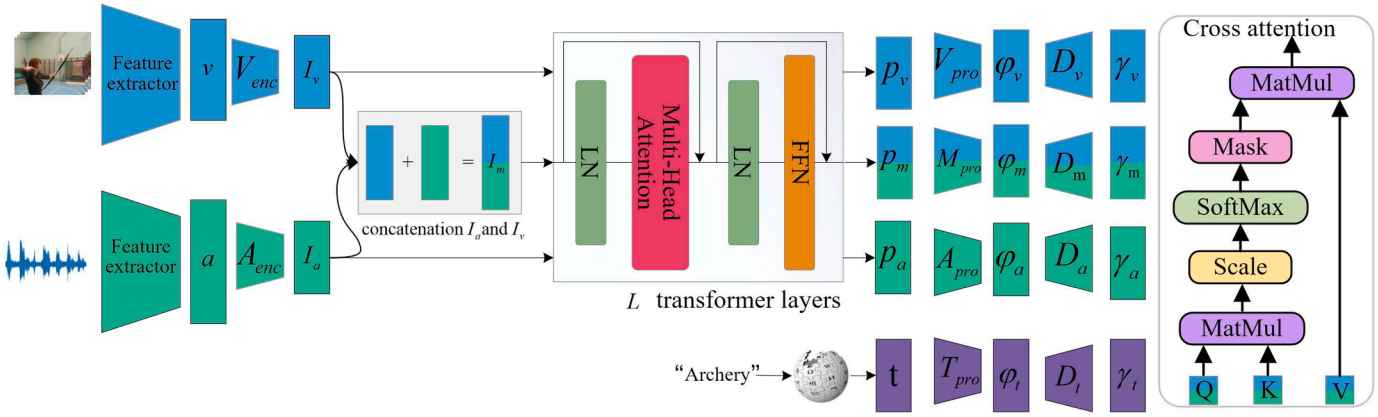


Fig. 2. SACMA takes audio and visual features extracted from video data as input. After concatenation, the fused features are passed to transformer layers, where self-attention captures intra-modal information and cross-attention models inter-modal interactions. The resulting classification outputs (p_a, p_v, p_m) are projected into a multimodal common space shared with text embeddings, where the loss function is applied. Projection modules (A_{pro}, V_{pro}, M_{pro}, T_{pro}) and reconstruction modules ($D_{(a)}$, $D_{(v)}$, $D_{(m)}$, $D_{(t)}$) correspond to different modalities.

In multimodal tasks, the model needs to learn the characteristics of each modality independently, and it also needs to understand and pay attention to the relationships between different modalities. For example, in audio and video sentiment analysis tasks, the model needs to understand not only the sound characteristics contained in the input audio and the facial expression characteristics of the input video but also the relationship between these two types of characteristics because sounds and expressions are often related to each other when emotions are expressed. To encourage the model to pay attention to the relationships between modalities, we use early fusion to integrate the information derived from different modalities. Early fusion can fuse the information acquired from different modalities into a common representation early in the data processing pipeline, allowing the data elements of different modalities to interact at the input level; this helps the model to more comprehensively understand multimodal inputs and jointly participate in feature learning.

$$I_m = concat(I_a, I_v) \tag{2}$$

where concat represents a splicing operation and a and v represent the embedding of audio and visual features, respectively.

3.4. Multimodal fusion transformer

The proposed method contains L-stacked transformer layers; each transformer layer consists of a multihead attention (MHA) layer and a feedforward neural network (FFN) layer, and layer normalization (LN) is applied before each layer. After layer normalization is applied, the input features are first sent to the multihead attention layer. After residual connection and layer normalization are applied, the signals are sent to the feedforward neural network layer. The output of the l-th feedforward neural network is used as the input of the $l+1$-th multihead attention layer. The FFN contains two linear layers, a gaussian error linear unit (GELU), which serves as the nonlinear activation function, and two random deactivation functions. Specifically, a self-attention mechanism is applied to single-modal inputs I_a and I_v, whereas a cross-modal attention mechanism is employed for multimodal input I_m. The self-attention mechanism enables the model to focus on the information contained within the input sequence, which helps the model more comprehensively understand the intrinsic structure of single-modal data. The cross-modal cross-attention mechanism can establish dependencies between different data modalities, thereby effectively merging multimodal information and integrating the important features of multimodal data. Combining the two mechanisms can help the model learn more powerful audio-visual representations. We transform the input embedding x into query (Q), key (K), and value (V) vectors by means of three linear transformations: $Q = XW_Q$, $Q = XW_K$, and $Q = XW_V$. W_Q, W_K, and W_V are the weight matrices. The attention score can be obtained through Eq. (3). Since multimodal inputs concatenate audiovisual features, the cross-attention mechanism pays more attention to the information interaction between different modalities, and the attention scores of the same modality are masked by the mask matrix.

$$Score(Q,K) = \frac{QK^T}{\sqrt{d_k}} mask(opt.) \tag{3}$$

$$Attention(Q,K,V) = soft\max[(Score(Q,K)]V \tag{4}$$

where Attention(.) denotes the attention output and mask (opt.) denotes that the mask is optional. The formula for the l-th transformer layer is as follows:

$$Z_l' = MHA(LN(Z_{l-1})) + Z_{l-1} \tag{5}$$

$$Z_l = FFN\left(LN\left(Z_l'\right)\right) + Z_l' \tag{6}$$

where l represents the number of layers, the value range is $\{1 \dots L\}$, and Z_{l-1} represents the output of the $l-1$ layer.

3.5. Projection to the multimodal common space

In multimodal data, information from different modalities has different semantic granularities. In each training iteration, for each input of I_a, I_v, and I_m, the SACMA model is called once to obtain the corresponding output classification token. So there are three outputs p_a, p_v, and p_m. To obtain the final embedding for each modality, we map the output and text label embeddings of the model together into a multimodal common space. In the prediction phase, we obtain class predictions by determining the projected word2vec embedding that is closest to the output embedding.

$$p = \arg\min_c \left(\|\varphi_t^c - \varphi_m\|_2\right) \tag{7}$$

where φ_t^c represents the representation of the word2vec embedded class label of class c after performing embedding and mapping, and φ_m represents the representation of the audio-visual output after the mapping process.

3.6. Loss function

We train our model using a loss function l consisting of a combined contrastive loss [44] l_{ccl}, a cosine similarity loss $l_{\cos}$, a regression loss l_{reg}, and a reconstruction loss l_{rec}.

$$l = l_{ccl} + \lambda_{\_\cos} l_{\cos} + l_{reg} + l_{rec} \tag{8}$$

where $\lambda_{\_\cos}$ represents the weight of the cosine similarity loss $l_{\cos}$.

3.6.1. Combinatorial contrastive loss

Our combined contrastive loss consists of a single-modal contrastive loss l_s and a multimodal contrastive loss l_m.

$$l_{ccl} = l_s + l_m \tag{9}$$

Unlike traditional contrastive learning methods that rely on explicit negative sampling, our approach implicitly leverages intra-batch negatives: within each mini-batch, non-matching sample pairs are naturally treated as negative examples. This design encourages the model to draw semantically aligned cross-modal samples closer in the shared embedding space while simultaneously pushing apart dissimilar pairs. To this end, we train three pairwise contrastive losses, namely the contrastive loss l_{a_v} between audio and video, the contrastive loss between text and audio l_{t_a}, and the contrastive loss between text and video l_{a_v}.

$$l_s = \lambda_{t\_a} l_{t\_a} + \lambda_{t\_v} l_{t\_v} + \lambda_{a\_v} l_{a\_v} \tag{10}$$

where λ_{α_β} represents the weighting coefficient of (α, β). Our single-modal combination loss considers all possible and available modal combinations and can be generalized to any set of modalities $\mathrm{M} = \{m_1, m_2, m_3, \ldots, m_n\}$.

$$l_s = \sum_{u,v \subset M, u \cap v = \emptyset} \lambda_{uv} l_{uv} \tag{11}$$

where l_{uv} is the contrastive loss between modality u and modality v, and λ_{uv} is the weighting coefficient. Moreover, we encourage class tokens to exchange information between a single modality and multiple modalities, that is, the contrastive loss l_{a_av} is between a and av, the contrastive loss l_{v_av} is between v and av, and the contrastive loss l_{t_av} is between t and av. The multimodal contrastive loss l_m is a combination of the above losses.

$$l_m = \lambda_{t\_av} l_{t\_av} + \lambda_{a\_av} l_{a\_av} + \lambda_{v\_av} l_{v\_av} \tag{12}$$

The complete representation of our combined comparison loss l_{ccl} is as follows:

$$\begin{aligned} l_{ccl} =& \lambda_{t\_v} l_{t\_v} + \lambda_{t\_v} l_{t\_v} + \lambda_{a\_v} l_{a\_v} \\ &+ \lambda_{t\_av} l_{t\_av} + \lambda_{a\_av} l_{a\_av} + \lambda_{v\_av} l_{v\_av} \end{aligned} \tag{13}$$

We use noise contrastive estimation [45] to calculate the combined contrastive loss:

$$\begin{aligned} l_{u\_v} =& -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp\left(u_i^{\mathrm{T}} v_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(u_i^{\mathrm{T}} v_j/\tau\right)}\right) - \\ & -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp\left(u_i^{\mathrm{T}} v_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(u_j^{\mathrm{T}} v_i/\tau\right)}\right) \end{aligned} \tag{14}$$

where τ represents the temperature parameter and B is the batch size.

3.6.2. Cosine similarity loss

Our cosine similarity loss computes the similarity between the three outputs φ_a, φ_m, and φ_v of the model and the text label φ_t, for which we train three pairwise similarity losses l_{a_t}, l_{v_t}, and l_{m_t}. Moreover, we implement an additional cosine similarity loss l_{a_v} to calculate the similarity between φ_a and φ_v.

$$l_{\cos} = l_{a\_t} + l_{v\_t} + l_{m\_t} + l_{v\_a} \tag{15}$$

where l_{x_y} represents the cosine similarity loss between φ_x and φ_y. In deep learning, the cosine similarity loss l_{x_y} between x and y is calculated as follows:

$$l_{x\_y} = 1 - \cos(x, y) \tag{16}$$

where $x, y \subset \{\varphi_a, \varphi_v, \varphi_m, \varphi_t\}$, and $\cos(x, y)$ represent the cosine similarity between x and y. When the cosine similarity is close to 1, the loss value is close to 0, indicating that the similarity between the model prediction and the target is high; however, when the cosine similarity is far from 1, the loss value increases, indicating that the similarity is low. The value range of cosine similarity is [-1, 1], where 1 indicates complete similarity, -1 indicates complete dissimilarity, and 0 indicates orthogonality, i.e., no similarity or independence. The values in between represent intermediate similarities or dissimilarities. The cosine similarity between x and y is calculated as follows:

$$\cos(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{17}$$

3.6.3. Regression loss

The goal of the regression loss is to reduce the distance between the single-modal output embedding of a sample, the multimodal output embedding, and the corresponding word2vec embedding. As in [40], our regression loss is also based on the mean square error metric. However, we supplement this approach and focus not only on the distance between the multimodal output embedding and the text label embedding but also on the single-modal distance between the static input and the text label embedding. The regression loss is expressed as follows:

$$\begin{aligned} l_{reg} =& \frac{1}{n}\sum_{i=1}^{n}\left(\varphi_{m_i} - \varphi_{t_i}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\varphi_{a_i} - \varphi_{t_i}\right)^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left(\varphi_{v_i} - \varphi_{t_i}\right)^2 \end{aligned} \tag{18}$$

where φ_{m_i}, φ_{a_i}, and φ_{v_i} are the audio-visual embedding, audio embedding, and visual embedding, respectively.

3.6.4. Reconstruction loss

The reconstruction loss of the SACMA method is also based on the mean squared error metric and complements the process of reconstructing single-modal output embeddings. The goal of the reconstruction loss is to ensure that our model is able to decode the semantic information contained in the pre-extracted text label embeddings from embeddings φ_a, φ_m, φ_v, and φ_{v_i}. The reconstruction loss is expressed as follows:

$$\begin{aligned} l_{rec} =& \frac{1}{n}\sum_{i=1}^{n}\left(\gamma_{m_i} - t_i\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\gamma_{a_i} - t_i\right)^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left(\gamma_{v_i} - t_i\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\gamma_{t_i} - t_i\right)^2 \end{aligned} \tag{19}$$

where γ_{m_i}, γ_{a_i}, γ_{v_i}, and γ_{t_i} represent the reconstructed audio-visual features, audio features, visual features, and text features, respectively.

4. Experimental setup

4.1. Dataset

We evaluate our method on the three benchmark datasets (UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL) that were introduced previously [46]. Each dataset is divided according to the total number of classes in the dataset, the total number of seen classes, the total number of unseen classes, and the detailed division of the numbers of seen and unseen classes during training, validation, and testing (as shown in Table 1).

VGGSound-GZSL is a modified version of the large audio-visual dataset VGGSound [47]. This dataset contains a total of 276 classes, including 138 seen classes and 138 unseen classes.

UCF-GZSL is an action recognition dataset for real-life action videos. This dataset is a subset of UCF101 [48] and contains a total of 51 classes of data sources, including 30 seen classes and 21 unseen classes.

ActivityNet-GZSL is an action recognition dataset based on ActivityNet [49]. This dataset contains a total of 200 classes of data, including 99 visible classes and 101 unseen classes.

4.2. Implementation details

The Adam optimizer is used to train our model with a weight decay of $1e^{-5}$. The initial learning rates for VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL are set to $6e^{-5}$, $7e^{-5}$, and $7e^{-5}$, respectively. When we

Table 1
Statistics of the three benchmark datasets, including their total numbers of classes (C), seen classes (S), and unseen classes (U), as well as the division of seen and unseen classes across training (tr_s), validation (val_s, val_u), and test (te_s, te_u) sets.

STAGE				First stage			Second stage		
				Training	Validation		Training	Test	
DATASET	C	S	U	tr_s	val_s	val_u	tr_s	te_s	te_u
UCF-GZSL	51	30	21	30	30	12	42	42	9
VGGSound-GZSL	276	138	138	138	138	138	207	207	69
ActivityNet-GZSL	200	99	101	99	99	51	150	150	50

calculate the loss function l, we set the weight $\lambda_{\_\cos}$ of the loss $l_{\_\cos}$ in Eq. (8) to 0.2. We also calculate the temperature parameter when calculating the contrast loss and normalize the vector before calculating the dot product. Larger weights are set for the loss functions l_{t_a}, l_{t_v}, and l_{t_av} in Eq. (13) because this is beneficial for training the model: $\lambda_{t\_av} = 1$, $\lambda_{t\_a} = \lambda_{t\_v} = 0.8$, and $\lambda_{a\_v} = \lambda_{a\_av} = \lambda_{v\_av} = 0.1$. Furthermore, for the VGGSound-GZSL dataset, the learning rate attenuator is set to true, and the other settings follow those employed in [34]; all the models are trained on a single NVIDIA GeForce GTX 3090Ti GPU.

4.3. Evaluation metrics

GZSL suffers from a common problem in that the model has an inherent bias towards unseen classes, resulting in a generally higher accuracy for seen classes. To overcome this bias and improve the accuracies on both seen and unseen classes simultaneously, we follow [58] and use average class accuracy to evaluate the model, calculate the average accuracies on seen (S) and unseen (U) classes, and use their harmonic mean (HM) as the evaluation metric for our GZSL task. The ZSL performance is obtained by considering only the subset of test samples from the unseen test classes.

4.4. Comparsion methods

We compare the proposed SACMA method with five ZSL approaches and six GZSL approaches. For ZSL methods, we use concatenated image and audio features as input instead of image features alone. The core ideas of the compared methods are as follows:

DEVISE [52] introduces a deep embedding model that leverages textual data to learn semantic relations between labels and explicitly maps images into a rich semantic embedding space, where similarity matching is used for label prediction.

ALE [1] is a label-embedding-based ZSL framework that maps classes into an attribute vector space and performs classification by learning a compatibility function between images and label embeddings.

SJE [53] learns a compatibility function between image and class embeddings, ensuring that matched embeddings receive higher scores than mismatched ones.

APN [6] jointly learns global features and local features through an attribute prototype network, enhancing the localization and disentanglement of attributes for more effective knowledge transfer from seen to unseen classes.

f-VAEGAN-D2 [54] is a unified conditional generative framework that combines VAE and GAN, leveraging unlabeled data to model marginal feature distributions and generate interpretable visual features.

CJME [55] embeds video, audio, and text labels into a shared embedding space, ensuring that embeddings of the same class are closer than those of different classes. It further introduces a modality-attention mechanism to identify the dominant modality.

AVZSLNet [56] extends CJME by employing a cross-modal decoder and composite triplet loss. The decoder enforces reconstruction of textual label features from audio and video embeddings, while the composite triplet loss minimizes distances among audio, video, and textual embeddings.

AVCA [49] proposes a cross-modal attention framework to integrate audio and visual information and align the resulting audio-visual embeddings with textual label embeddings.

TCAF [42] builds upon AVCA by additionally exploiting temporal information from audio and video inputs and applying cross-attention to capture cross-modal dependencies.

AVFS [59] introduces an audio-visual feature synthesis method that leverages contrastive and discriminative learning to simulate audio-visual features of unseen classes.

AVMST [57] proposes an Audio-Visual Modality-fusion Spiking Transformer network, which extracts temporal features using a spiking neural network, fuses semantic and temporal information through cross-attention, and performs feature reasoning with a transformer, enabling efficient classification of unseen video classes.

Our method is built upon the baseline TCAF model. TCAF primarily adopts a feature-level fusion strategy by introducing a temporal-aware cross-modal attention mechanism, which effectively leverages the temporal correlations between the audio and visual modalities. However, it does not explicitly explore the internal feature structures contained within a single modality. To address this limitation, we propose SACMA, which incorporates a multilevel attention mechanism consisting of self-attention and cross-attention modules to achieve deeper feature modeling and more efficient cross-modal interaction. Specifically, SACMA introduces a "dual embedding" strategy on top of TCAF: self-attention is applied to the given audio and visual sequences to strengthen their intramodal feature representations, while cross-attention is subsequently employed to capture richer and more effective intermodal interactions. Furthermore, we integrate a contrastive loss with a cosine similarity loss to enhance the semantic consistency and discriminability of the features across different modalities.

5. Experimental results

5.1. Comparison with state-of-the-art

To validate the effectiveness of our model, we compare it with the state-of-the-art audiovisual zero-shot learning methods on three benchmark datasets, as shown in Table 2. On the UCF-GZSL dataset, compared with the baseline TCAF model, which achieves 31.72 % HM and 24.81 % ZSL, SACMA achieves state-of-the-art performance, attaining 37.71 % HM and 29.07 % ZSL, respectively. On the VGGSound-GZSL dataset, SACMA achieves 8.29 % HM and 6.46 % ZSL for GZSL, whereas TCAF achieves 7.33 % HM and 6.06 % ZSL.

The classes contained in the ActivityNet-GZSL dataset are constructed on the basis of a semantic category ontology, which provides a rich hierarchical structure for action classes. For example, the class "hand washing clothes" belongs to "laundry" (second level), "household chores" (third level), and "home activities" (fourth level). However, precisely owing to this clear hierarchical organization structure, the performance achieved by SACMA on this dataset does not meet expectations, yielding results that are comparable to those of the baseline. This limitation arises because SACMA focuses primarily on intramodal feature learning and cross-modal feature alignment and does not fully exploit the hierarchical semantic information embedded in the input dataset, thereby constraining its performance. Nevertheless, SACMA produces

Table 2
Experimental results of our proposed method and the state-of-the-art audio-visual GZSL method on the VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL datasets.

Model	Venue	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
		S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
DEVISE [50]	NeurIPS'13	36.22	1.07	2.08	5.59	55.59	14.94	23.56	16.09	3.45	8.53	4.91	8.53
ALE [1]	T-PAMI'15	0.28	5.48	0.53	5.48	57.59	14.89	26.50	18.93	2.63	7.87	3.94	7.90
SJE [51]	CVPR'20	48.33	1.10	2.15	4.06	63.10	16.77	26.50	18.93	4.61	7.04	5.57	7.08
f-VAEGAN-D2 [52]	CVPR'19	12.77	0.95	1.77	1.91	17.29	8.47	11.37	11.11	4.36	2.14	2.87	2.40
APN [6]	IJCV'22	7.48	3.88	5.11	4.49	28.46	16.16	20.61	16.44	9.84	5.76	7.27	6.34
CJME [53]	WACV'20	8.69	4.78	6.17	5.16	26.04	8.21	12.48	8.29	5.55	4.75	5.12	5.84
AVGZSLNet [54]	WACV'21	18.05	3.48	5.83	5.28	52.52	10.90	18.05	13.65	8.93	5.04	6.44	5.40
AVCA [46]	CVPR'22	14.90	4.00	6.31	6.00	51.53	18.43	27.15	20.01	24.86	8.02	12.13	9.13
TCAF [40]	ECCV'22	9.64	5.91	7.33	6.06	58.60	21.74	31.72	24.81	18.07	7.50	10.71	7.91
AVFS [55]	IJCNN'23	15.20	5.13	7.67	6.00	54.87	16.49	25.36	22.37	29.00	9.13	**13.89**	**11.18**
AVMST [56]	ICME'23	14.14	5.28	7.68	**6.61**	44.08	22.63	29.91	28.19	17.75	9.90	12.71	10.37
SACMA	Ours	13.42	6.00	**8.29**	6.46	60.15	27.46	**37.71**	**29.07**	17.12	7.32	10.25	7.52

Table 3
Evaluation of the attention mechanism, showing GZSL and ZSL performance after removing individual components: audio self-attention (A_S), video self-attention (V_S), both ($A_S + V_S$), and cross-modal attention (M_c).

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
w/o A_S	15.25	5.13	7.68	5.62	81.08	15.42	25.91	25.15	9.36	4.49	6.07	4.79
w/o V_S	6.33	3.77	4.73	4.20	66.13	20.17	30.91	25.15	7.42	4.49	5.60	4.78
w/o $A_S + V_S$	4.00	5.59	4.67	5.59	66.42	15.71	25.41	22.40	11.72	4.16	6.14	4.83
w/o M_c	4.85	4.99	4.92	6.38	46.49	23.93	31.60	26.17	7.45	3.15	4.43	3.94
SACMA	13.42	6.00	**8.29**	**6.46**	60.15	27.46	**37.71**	**29.07**	17.12	7.32	**10.25**	**7.52**

Table 4
Impact of different textual embeddings and attention mechanisms.

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
w_{clip}	10.56	4.75	6.56	4.96	46.60	22.52	30.37	24.57	21.70	10.80	**14.42**	**11.59**
w_{clip} + only SA	12.03	3.91	5.91	4.43	27.04	24.94	25.94	25.31	13.56	10.46	11.81	10.72
SACMA	13.42	6.00	**8.29**	**6.46**	60.15	27.46	**37.71**	**29.07**	17.12	7.32	10.25	7.52

Table 5
Evaluation of loss functions by comparing GZSL and ZSL performance when ablating l_{ccl}, $l_{\cos}$, l_{reg}, and l_{rec} individually on VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL.

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
w/o l_{ccl}	0.48	1.45	0.73	1.77	8.24	0.14	0.28	18.62	0.60	0.56	0.58	1.93
w/o $l_{\cos}$	12.83	3.91	6.00	4.26	64.36	22.24	33.06	27.77	7.10	1.81	2.89	2.50
w/o l_{reg}	4.92	4.14	4.50	4.41	54.50	21.85	31.19	24.72	12.84	4.05	6.16	4.31
w/o l_{rec}	9.10	6.06	7.27	6.35	56.43	25.88	35.48	28.68	23.3	5.51	8.91	6.05
SACMA	13.42	6.00	**8.29**	**6.46**	60.15	27.46	**37.71**	**29.07**	17.12	7.32	**10.25**	**7.52**

significantly superior results on the VGGSound-GZSL and UCF-GZSL datasets, providing strong evidence of its effectiveness. By integrating self-attention and cross-attention mechanisms for conducting audiovisual representation learning and by enforcing both contrast and cosine similarity within a shared multimodal space, our method effectively leverages intramodal information and intermodal interactions to ensure semantic consistency across different modalities, thereby enhancing the robustness of the learned audiovisual representations.

5.2. Qualitative results

A qualitative analysis of the learned audio-visual embeddings is presented in Fig. 3. It shows t-SNE [59] visualization of the audio, the visual input features, and the learned audio-visual embedded features from six classes in the UCF-GZSL dataset. As shown in Fig. 3. The clustering of audio and video input features is poor, especially for audio input. In contrast, the audio-visual embeddings are clearly clustered. It appears that our SACMA-learned au-dio-visual features provide improved clustering effects over those of the input audio and visual features.

5.3. Ablation analysis

5.3.1. Influence of the attention mechanism

Table 3 shows that removing V_S for the VGGSound-GZSL and ActivityNet-GZSL datasets yields lower results than removing A_S. The performance of HM and ZSL on VGGSound-GZSL dramatically decreased from 8.29 % and 6.46 % to 4.73 % and 4.20 %, respectively. On ActivityNet-GZSL, the performance of HM and ZSL dramatically decreased from 10.25 % and 7.52 % to 5.60 % and 4.78 %, respectively. Interestingly, the opposite results are achieved on the UCF-GZSL dataset. On the VGGSound-GZSL and UCF-GZSL datasets, the performance decreased to its lowest value after removing $A_S + V_S$, which suggests that our self-attention mechanism helps to improve the performance of HM vs. ZSL, while on the ActivityNet-GZSL dataset, the performance achieved after removing $A_S + V_S$ is better than that achieved after removing A_S or V_S alone. After replacing the cross-attention mechanism used for combining features with the self-attention mechanism, the performance of the model on all three dataset sets significantly decreased, especially on the ActivityNet-GZSL dataset, where the values of HM and

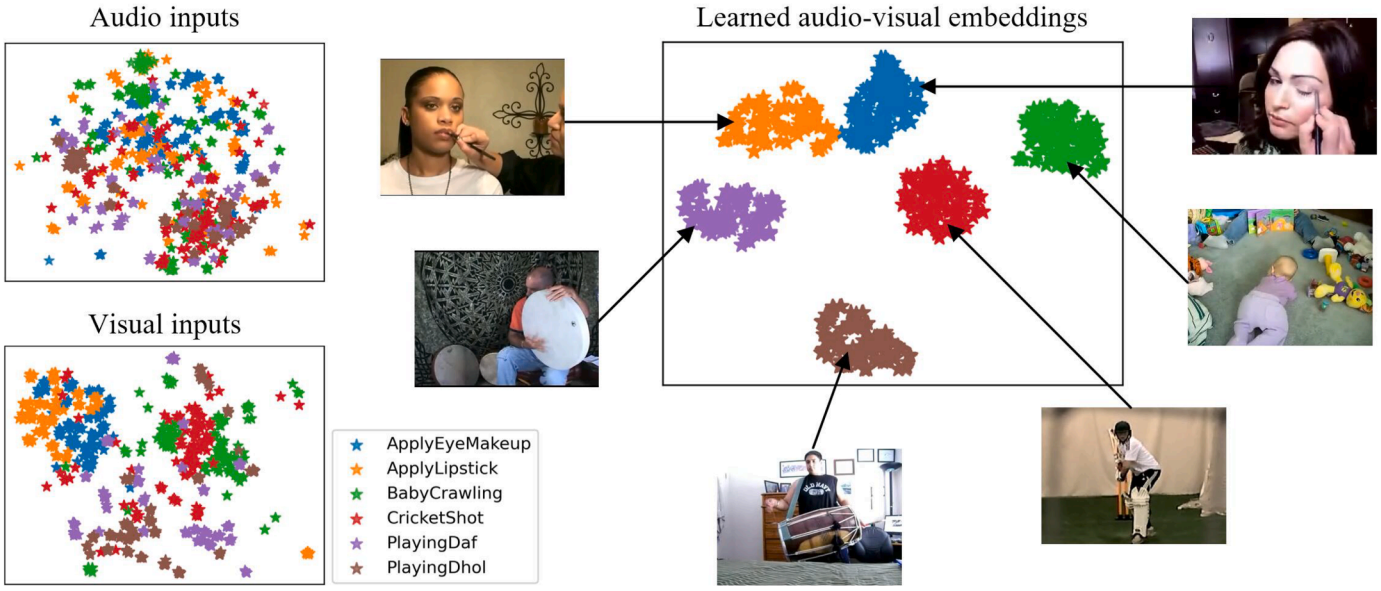


Fig. 3. t-SNE visualization for six seen samples from the UCF-GZSL dataset, showing audio and visual input embeddings extracted with SeLaVi [57], and audio-visual output embeddings learned with SACMA.

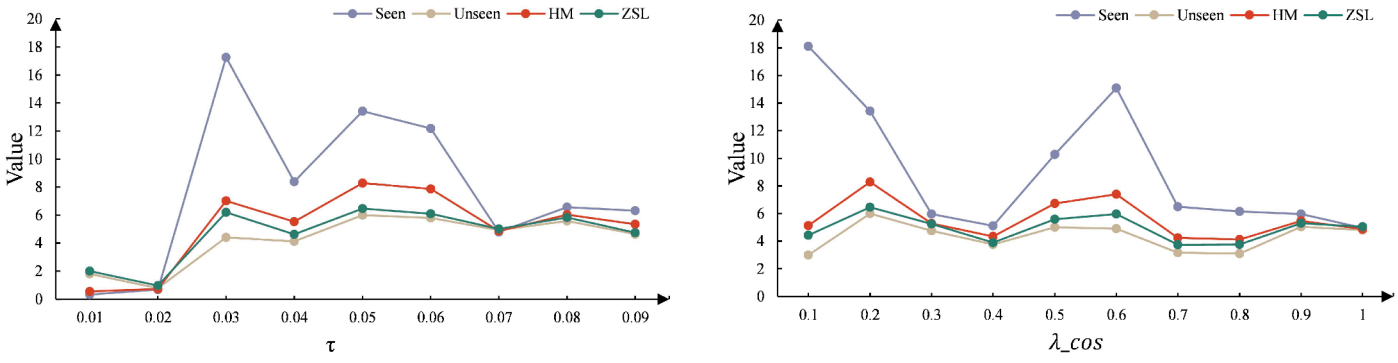
ZSL decreased from 10.25 % and 7.52 % to 4.43 % and 3.94 %, respectively.The interaction of information from different modal data is beneficial for improving the performance of the audiovisual GZSL task-heavy HM and ZSL. Overall, our $A_S + V_S + M_C$ achieved the strongest GZSL and ZSL performances on all three datasets, proving the sophistication of our attentional mechanism selection approach.

5.3.2. *Influence of different text encoders and attention mechanisms*

To further validate the effectiveness of each module in the proposed method, we evaluated the impact of different textual embeddings and attention mechanisms, as summarized in Table 4. When replacing word2vec with Contrastive Language-Image Pre-Training(CLIP) embeddings (w_{clip}), the model achieves improved performance on ActivityNet-GZSL, with HM and ZSL increasing from 10.25 % and 7.32 % to 14.42 % and 11.59 %, respectively, indicating that richer semantic information is beneficial for this dataset. However, on VGGSound-GZSL and UCF-GZSL, the performance metrics decrease, suggesting that although CLIP provides richer semantic information and advantages in visual-text alignment, it does not consistently enhance generalization in audiovisual GZSL scenarios. Furthermore, when using (w_{clip} + only SA), the performance declines across all three datasets, demonstrating that relying solely on intra-modal feature learning is insufficient to effectively promote cross-modal semantic alignment and zero-shot generalization. In contrast, SACMA achieves the best performance on VGGSound-GZSL and UCF-GZSL, providing strong evidence that the proposed combination of self-attention and cross-attention effectively captures both intra-modal key features and inter-modal semantic relationships.

5.3.3. *Influence of hyperparameter selection*

In this section, we experimentally evaluate the effect of the temperature parameters τ and $\lambda_{\_cos}$ on the SACMA performance. The combined contrast loss acts as the distance between samples of the same category, and we set a smaller temperature parameter τ to increase the sensitivity of the model and impose a larger penalty on samples of different categories. The weight of the cosine similarity loss $\lambda_{\_cos}$ can improve the convergence of the model and avoid overfitting. First, we start with the optimal value of the temperature parameter τ, and then we select the value of the weight $\lambda_{\_cos}$ for the cosine similarity loss. Fig. 4 shows the results of our experiments on the VGGSound-GZSL dataset with two hyperparameters chosen. According to Fig. 4, SEEN is highly sensitive to changes in the two hyperparameters. The sensitivities of UNSEEN, HM, and ZSL are relatively small, and the changes are relatively flat. Based on the experimental results, we select a temperature parameter of 0.05 and a cosine similarity loss weight of 0.2 as the optimal values for our model.



(a) Influence of τ parameter selection

(b) Influence of $\lambda_{\_cos}$ parameter selection

Fig. 4. Influence of hyperparameter selection.

Table 6
Evaluation of the influence of different input modalities.

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
I_m	10.88	3.16	4.90	3.45	46.42	20.62	28.56	24.80	15.67	5.46	8.09	5.68
$I\_m + I\_a$	8.21	4.26	5.61	4.87	56.83	18.55	27.97	20.91	1.78	5.10	2.63	5.10
$I\_m + I\_v$	10.37	6.14	7.72	6.38	56.06	21.20	30.77	24.37	6.95	5.36	6.05	5.61
SACMA	13.42	6.00	**8.29**	**6.46**	60.15	27.46	**37.71**	**29.07**	17.12	7.32	**10.25**	**7.52**

Table 7
Evaluation of the performance of the different output representations.

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
φ_a	3.55	3.13	3.33	3.57	36.60	17.73	23.89	19.40	5.27	3.56	4.25	3.86
φ_v	4.43	3.83	4.19	3.94	30.91	18.84	23.41	24.37	6.27	4.29	5.09	4.77
φ_m	13.42	6.00	**8.29**	**6.46**	60.15	27.46	**37.71**	**29.07**	17.12	7.32	**10.25**	**7.52**

5.3.4. Analysis of the impact of different loss functions

In this section, we conduct an ablation experiment analysis on the impact of changes in the loss function components (l_{ccl}, $l_{\cos}$, l_{reg}, and l_{rec}). In the experiment, we eliminate one of the components at a time to observe the model performance on the VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL datasets. We observe that on the three datasets, eliminating one of the components of the loss function negatively affects the performance of the model, demonstrating the importance of each of these components. On the UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL datasets, the results decreased sharply when l_{ccl} was removed. The performances of HM and ZSL on VGGSound-GZSL decreased from 8.29 % and 6.46 % to 0.73 % and 1.77 %, respectively; for UCF-GZSL, the performances of GZSL and ZSL decreased from 37.71 % and 29.07 % to 0.28 % and 18.62 %, respectively, whereas on ActivityNet-GZSL, the performances of GZSL and ZSL decreased from 10.25 % and 7.52 % to 0.58 % and 1.93 %, respectively. The three datasets are extremely sensitive to changes in l_{ccl}, which shows that l_{ccl} can greatly improve the performance of the model and is well suited for the GZSL and ZSL tasks. After eliminating $l_{\cos}$ and l_{reg}, the performance of the model on the VGGSound-GZSL and ActivityNet-GZSL datasets decreased significantly, which shows that $l_{\cos}$ and l_{reg} are better for larger and more complex datasets. The impacts of eliminating l_{rec} on the results on the three datasets are relatively similar, indicating that using l_{rec} to constrain the representation learned by the model to contain information in text label information is beneficial and can steadily improve the performance of the model. After using our full loss function, the best GZSL and ZSL performance were achieved on the three datasets, demonstrating the effectiveness of this loss function. Table 5

5.4. Evaluating different modalities

5.4.1. Influence of modal inputs

We compared the performances of training SACMA using different input modalities, and the results are shown in Table 6. In the three datasets, adding visual input performed better than adding audio input. This indicates that visual input features contain more comprehensive video content information than audio input features. For VGGSound-GZSL, after adding the audio mode, the performance of HM increased from 4.9 % to 5.61 %, and the performance of ZSL increased from 3.45 % to 4.87 %. Interestingly, the opposite result was observed for UCF-GZSL. After adding the audio mode, the HM and ZSL scores decreased from 28.56 % and 24.8 % to 27.97 % and 20.91 %, respectively. For ActivityNet-GZSL, adding audio or visual input alone leads to performance degradation; however, adding audio and visual information simultaneously improves the model's performance, indicating that exploiting complementary information in audio and visual inputs is very beneficial for GZSL and ZSL in video classification.

5.4.2. Influence of different modal output

The results of the evaluation using different output representations are shown in Table 7. For VGGSound-GZSL and ActivityNet-GZSL, using φ_v as the output evaluation representation led to slightly better performance than did using φ_a, suggesting that they learn visual information better. Specifically, for VGGSound-GZSL, the HM and ZSL scores obtained using φ_a as the output evaluation representation are 3.33 % and 3.57 %, respectively, while the HM and ZSL scores obtained using φ_v are 4.19 % and 3.94 %, respectively. On ActivityNet-GZSL, the HM and ZSL scores obtained using φ_a as the output evaluation are 4.25 % and 3.86 %, respectively, while the HM and ZSL scores obtained using φ_v are 5.09 % and 4.77 %, respectively. The HM performance obtained by using φ_a as the output evaluation representation for UCF-GZSL was slightly better than that obtained by using φ_v, and the opposite trend was observed in terms of the ZSL method, with HM scores of 23.89 % and 23.41 % and ZSL scores of 19.4 % and 24.37 %, respectively. Overall, using φ_m as the output evaluation representation on the three datasets led the best performance representation, better than that obtained using φ_a and φ_v as the output performance representation. Therefore, we adopt φ_m as the output evaluation representation of our model.

6. Conclusion

This study proposes an attention-based audio-visual generalized zero-shot learning method to improve the model's ability to learn from audio-visual data, obtain better audio-visual representations, and achieve knowledge transfer from seen classes to unseen classes. This method processes single-modal input through a self-attention mechanism, captures key features within each modality, and optimizes the utilization of single-modal information. Moreover, a cross-attention mechanism is used to process multimodal input, allowing the model to more comprehensively understand and integrate multimodal information and explore the interrelationships between modalities in detail. During training, a combined contrast loss function is introduced to analyze the combination of model input modalities during the training process, thereby strengthening the model's understanding of the relationships between different modalities, improving the generalizability of the model, and making the model better suited for complex audio-visual multimodal tasks. The introduction of the cosine similarity loss at the same time improves the model's ability to accurately capture the intrinsic similarity of the samples and allows the model to complete the classification task by optimizing the similarity between the representations learned by the model between samples of the same category. The experiments show that our method achieves state-of-the-art performance on three benchmark datasets. For example, on the UCF-GZSL dataset, the performance of HM reaches 37.71 %, and the performance of ZSL reaches 29.07 %, which are better than those of the existing advanced methods.

The proposed SACMA model focuses on features within a single modality and feature alignment between different modalities. It leads overall on the VGGSound-GZSL and UCF-GZSL datasets but performs commonly on the ActivityNet-GZSL dataset, which has a distinct hierarchical structure. In the future, further research is recommended to explore how to make the model learn the intrinsic hierarchical structure of the data to improve its generalization ability on hierarchical datasets. We will also continue to investigate more effective multimodal feature fusion strategies to ensure that information from different modalities can better complement each other.

Data availability

The code and dataset are available at: https://github.com/ybyangjing/SACMA.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jing Yang: Writing – original draft, Investigation, Conceptualization; **Xiaoyong Li:** Writing – original draft, Visualization, Methodology, Conceptualization; **Yuankai Wu:** Writing – review & editing, Data curation; **Yuling Chen:** Writing – review & editing, Supervision, Methodology, Conceptualization; **Xiaoli Ruan:** Writing – review & editing, Methodology, Investigation; **Chengjiang Li:** Writing – review & editing, Supervision; **Qing Hou:** Writing – review & editing, Supervision, Methodology.

Acknowledgment

This work was supported by the national natural science foundation of China 62441608, 62166005, the developing objects and projects of scientific and technological talents in Guiyang city (No.ZKHT[2023]48-8), Guizhou University Basic Research Fund ([2024]08), Open project of State Key Laboratory of Public Big Data (PBD2023-09,PBD2023-16), Guizhou High-level Innovative Talents (GCC[2023]101).

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (7) (2015) 1425–1438.
[2] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936. https://doi.org/{10.1109/cvpr.2015.7298911}.
[3] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero-and few-shot learning via aligned variational autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8247–8255. https://doi.org/{10.1109/cvpr.2019.00844}.
[4] B. Romera-Paredes, P.H.S. Torr, An embarrassingly simple approach to zero-shot learning, Vis. Attributes (2017) 11. https://doi.org/{10.1007/978-3-319-50077-5_2}.
[5] V.K. Verma, G. Arora, A. Mishra, P. Rai, Generalized zero-shot learning via synthesized examples, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4281–4289. https://doi.org/{10.1109/cvpr.2018.00450}.
[6] W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata, Attribute prototype network for zero-shot learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21969–21980.
[7] Y. Luo, X. Wang, F. Pourpanah, Dual VAEGAN: a generative model for generalized zero-shot learning, Appl. Soft Comput. 107 (2021) 107352. https://doi.org/{10.1016/j.asoc.2021.107352}.
[8] Y. Xiong, W.-C. Chang, C.-J. Hsieh, H.-F. Yu, I. Dhillon, Extreme zero-shot learning for extreme text classification, Recall 40 (50) (2022) 60. https://doi.org/{10.18653/v1/2022.naacl-main.399}.
[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. Inst. Electr. Electron. Eng. Inc. 86 (11) (1998) 2278–2324. https://doi.org/{10.1109/5.726791}.
[10] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551. https://doi.org/{10.1162/neco.1989.1.4.541}.
[11] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, S.-M. Hu, Visual attention network, Comput. Vis. Media 9 (4) (2023) 733–752.
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 84–90. https://doi.org/{10.1145/3065386}.
[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, (2014). arXiv preprint arXiv:1409.1556, https://api.semanticscholar.org/CorpusID:14124313
[14] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, J. Ma, MSR-net: low-light image enhancement using deep convolutional network, (2017). arXiv preprint arXiv:1711.02488, https://api.semanticscholar.org/CorpusID:23231130
[15] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: efficient convolutional neural networks for mobile vision applications, (2017).arXiv preprint arXiv:1704.04861, https://api.semanticscholar.org/CorpusID:12670695
[16] X. Zhang, X. Zhou, M. Lin, J. Sun, ShuffleNet: an extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856. https://doi.org/{10.1109/cvpr.2018.00716}.
[17] M. Lin, Q. Chen, S. Yan, Network in network, (2013). arXiv preprint arXiv:1312.4400, https://api.semanticscholar.org/CorpusID:16636683
[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9. https://doi.org/{10.1109/cvpr.2015.7298594}.
[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. https://doi.org/{10.1109/cvpr.2016.90}.
[20] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 31, 2017, pp. 4278–4284. https://doi.org/{10.1609/aaai.v31i1.11231}.
[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497. https://doi.org/{10.1109/iccv.2015.510}.
[22] J. Gao, T. Zhang, C. Xu, I know the relationships: zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8303–8311. https://doi.org/{10.1609/aaai.v33i01.33018303}.
[23] Y. Liu, R. Dian, S. Li, Low-rank transformer for high-resolution hyperspectral computational imaging, Int. J. Comput. Vis. 133 (2024) 809–824. https://doi.org/{10.1007/s11263-024-02203-7}.
[24] J. Yang, X. Ma, Y. Wu, C. Li, Z. Su, J. Xu, Y. Feng, AOGN-CZSL: an attribute- and object-guided network for compositional zero-shot learning, Inf. Fusion 120 (2025) 103096. https://doi.org/{10.1016/j.inffus.2025.103096}.
[25] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille, Attention to scale: scale-aware semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3640–3649. https://doi.org/{10.1109/cvpr.2016.396}.
[26] M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015) 7–12.
[27] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017). https://api.semanticscholar.org/CorpusID:13756489
[29] J.D. M.-W.C. Kenton, L.K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, (2019), pp. 4171–4186. https://api.semanticscholar.org/CorpusID:52967399
[30] M. Zhang, L. Liu, Y. Jin, Z. Lei, Z. Wang, L. Jiao, Tree-shaped multiobjective evolutionary CNN for hyperspectral image classification, Appl. Soft Comput. 152 (2024) 111176. https://doi.org/{10.1016/j.asoc.2023.111176}.
[31] S. Ding, X. Ruan, J. Yang, J. Sun, S. Li, J. Hu, LSSMA: lightweight spectral-spatial neural architecture with multiattention feature extraction for hyperspectral image classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17 (2024) 6394–6413. https://doi.org/{10.1109/JSTARS.2024.3371536}.
[32] P. Jiang, Y. Xue, F. Neri, Convolutional neural network pruning based on multi-objective feature map selection for image classification, Appl. Soft Comput. 139 (2023) 110229. https://doi.org/{10.1016/j.asoc.2023.110229}.
[33] R. Shang, J. Wang, L. Jiao, X. Yang, Y. Li, Spatial feature-based convolutional neural network for PolSAR image classification, Appl. Soft Comput. 123 (2022) 108922. https://doi.org/{10.1016/j.asoc.2022.108922}.
[34] W. Yang, Z. Li, C. Wang, J. Li, A multi-task faster R-CNN method for 3D vehicle detection based on a single image, Appl. Soft Comput. 95 (2020) 106533. https://doi.org/{10.1016/j.asoc.2020.106533}.
[35] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, C. Feichtenhofer, VideoCLIP: contrastive pre-training for zero-shot video-text understanding, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6787–6800. https://doi.org/{10.18653/v1/2021.emnlp-main.544}.
[36] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, A. Zisserman, Localizing visual sounds the hard way, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16867–16876. https://doi.org/{10.1109/cvpr46437.2021.01659}.

[37] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, A. Torralba, The sound of pixels, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 570–586.
[38] Y. Cheng, R. Wang, Z. Pan, R. Feng, Y. Zhang, Look, listen, and attend: co-attention network for self-supervised audio-visual representation learning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3884–3892. https://doi.org/{10.1145/3394171.3413869}.
[39] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun, Attention bottlenecks for multimodal fusion, Adv. Neural Inf. Process. Syst. 34 (2021) 14200–14213.
[40] O.-B. Mercea, T. Hummel, A.S. Koepke, Z. Akata, Temporal and cross-modal attention for audio-visual zero-shot learning, in: European Conference on Computer Vision, Springer, 2022, pp. 488–505. https://doi.org/{10.1007/978-3-031-20044-1_28}.
[41] A. Owens, J. Wu, J.H. McDermott, W.T. Freeman, A. Torralba, Ambient sound provides supervision for visual learning, in: Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 801–816. https://doi.org/{10.1007/978-3-319-46448-0_48}.
[42] R. Arandjelovic, A. Zisserman, Objects that sound, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 435–451.
[43] R. Gao, R. Feris, K. Grauman, Learning to separate object sounds by watching unlabeled video, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 35–53. https://doi.org/{10.1007/978-3-030-01219-9_3}.
[44] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. Feris, D. Harwath, J. Glass, H. Kuehne, Everything at once - multi-modal fusion transformer for video retrieval, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 19988–19997. https://doi.org/{10.1109/CVPR52688.2022.01939}.
[45] A.V.D. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, (2018). arXiv preprint arXiv:1807.03748, https://api.semanticscholar.org/CorpusID:49670925
[46] O.-B. Mercea, L. Riesch, A. Koepke, Z. Akata, Audio-visual generalised zero-shot learning with cross-modal attention and language, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10553–10563. https://doi.org/{10.1109/cvpr52688.2022.01030}.
[47] H. Chen, W. Xie, A. Vedaldi, A. Zisserman, Vggsound: a large-scale audio-visual dataset, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 721–725. https://doi.org/{10.1109/icassp40776.2020.9053174}.
[48] K. Soomro, A.R. Zamir, M. Shah, A dataset of 101 human action classes from videos in the wild, Center Res. Comput. Vis. 2 (11) (2012) 1–7. https://api.semanticscholar.org/CorpusID:7197134
[49] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, ActivityNet: a large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970. https://doi.org/{10.1109/cvpr.2015.7298698}.
[50] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, DeVISE: a deep visual-semantic embedding model, Adv. Neural Inf. Process. Syst. 26 (2013) 2121-2129.
[51] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936. https://doi.org/{10.1109/cvpr.2015.7298911}.
[52] Y. Xian, S. Sharma, B. Schiele, Z. Akata, F-VAEGAN-D2: a feature generating framework for any-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10275–10284. https://doi.org/{10.1109/cvpr.2019.01052}.
[53] K. Parida, N. Matiyali, T. Guha, G. Sharma, Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 3251–3260. https://doi.org/{10.1109/wacv45572.2020.9093438}.
[54] P. Mazumder, P. Singh, K.K. Parida, V.P. Namboodiri, AVGZSLNet: audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3090–3099. https://doi.org/{10.1109/wacv48630.2021.00313}.
[55] Q. Zheng, J. Hong, M. Farazi, A generative approach to audio-visual generalized zero-shot learning: combining contrastive and discriminative techniques, in: 2023 International Joint Conference on Neural Networks (IJCNN), IEEE, 2023, pp. 1–8. https://doi.org/{10.1109/ijcnn54540.2023.10191705}.
[56] W. Li, Z. Ma, L.-J. Deng, H. Man, X. Fan, Modality-fusion spiking transformer network for audio-visual zero-shot learning, in: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2023, pp. 426–431. https://doi.org/{10.1109/icme55011.2023.00080}.
[57] Y. Asano, M. Patrick, C. Rupprecht, A. Vedaldi, Labelling unlabelled videos from scratch with multi-modal self-supervision, Adv. Neural Inf. Process. Syst. 33 (2020) 4660–4671.
[58] Y. Xian, C.H. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2018) 2251–2265. https://doi.org/{10.1109/cvpr.2017.328}.
[59] L. van der Maaten, G.E. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a