applied sciences

MDPI

Article

Self-Supervised Video Representation and Temporally Adaptive Attention for Audio-Visual Event Localization

Yue Ran [1,2], Hongying Tang [1,2], Baoqing Li [1,2] and Guohui Wang [1,2,*]

[1] Science and Technology on Microsystem Laboratory, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 201800, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wgh0415@163.com

Abstract: Localizing the audio-visual events in video requires a combined judgment of visual and audio components. To integrate multimodal information, existing methods modeled the cross-modal relationships by feeding unimodal features into attention modules. However, these unimodal features are encoded in separate spaces, resulting in a large heterogeneity gap between modalities. Existing attention modules, on the other hand, ignore the temporal asynchrony between vision and hearing when constructing cross-modal connections, which may lead to the misinterpretation of one modality by another. Therefore, this paper aims to improve event localization performance by addressing these two problems and proposes a framework that feeds audio and visual features encoded in the same semantic space into a temporally adaptive attention module. Specifically, we develop a self-supervised representation method to encode features with a smaller heterogeneity gap by matching corresponding semantic cues between synchronized audio and visual signals. Furthermore, we develop a temporally adaptive cross-modal attention based on a weighting method that dynamically channels attention according to the time differences between event-related features. The proposed framework achieves state-of-the-art performance on the public audio-visual event dataset and the experimental results not only show that our self-supervised method can learn more discriminative features but also verify the effectiveness of our strategy for assigning attention.

Keywords: audiovisual event; temporal localization; fusion; representation learning; self-supervised learning

Citation: Ran, Y.; Tang, H.; Li, B.; Wang, G. Self-Supervised Video Representation and Temporally Adaptive Attention for Audio-Visual Event Localization. *Appl. Sci.* **2022**, *12*, 12622. https://doi.org/10.3390/app122412622

Academic Editors: Min Yang, Hao Liu, Shanxiong Chen and Yinong Chen

Received: 9 November 2022
Accepted: 7 December 2022
Published: 9 December 2022

1. Introduction

Teaching machines to use captured signals, such as visual and acoustic signals, to understand their surroundings is essential for constructing artificial intelligence. At present, studies in scene perception such as action recognition [1–3] and sound event detection [4,5], are mainly based on the use of unimodal signals. However, in more realistic situations, it may not be sufficient to characterize certain scenarios merely with unimodal information. To cope with the ambiguity caused by the unimodal data, the task of audio-visual event (AVE) localization is proposed to investigate how to understand video content by jointly leveraging audio and visual information in neural networks [6]. Specifically, an AVE refers to an event that is both visible and audible in a video segment. Localizing an AVE in a video requires predicting its temporal boundary and identifying its content. For example, as shown in Figure 1, a car appears in the beginning but can only be heard in the latter half of the video after its engine is started, thereby only the last three segments are categorized as 'car' while the other segments are *background*.

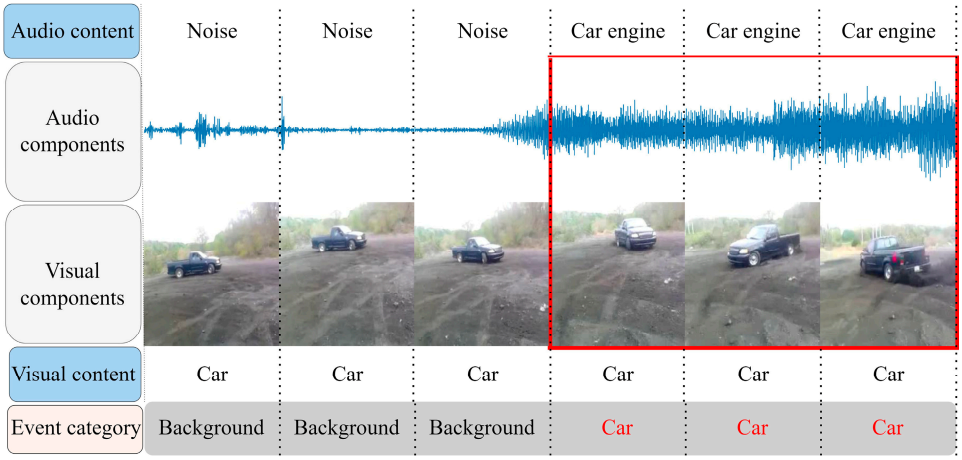

Figure 1. An illustration of the AVE localization task.

Due to the natural heterogeneity gap that exists between visual and acoustic signals, the key challenge in AVE localization is how to effectively fuse the information contained in the two modalities. To address the challenge, previous works [6–13] regarded primary visual and audio features extracted by unimodal backbones as tokens and fed them into attention modules to model the cross-modal relationships. However, the features they used are encoded in different unimodal spaces since they are extracted separately by convolutional neural networks (CNNs) trained solely on single-modal datasets [14,15]. Thus, there is a large heterogeneity gap between these features, which makes it difficult to model the relationship between modalities with only a few sets of scalar weights in previous attention networks. Unlike these deep learning techniques that use primary features learned on separate audio and visual data, the studies in [16–18] show that the human perceptual system naturally receives signals from multiple sensory streams and associates multimodal information collaboratively at an early stage of the learning process. Therefore, our approach is motivated to train feature encoders with both audio and visual data and associate their features in the same semantic space to narrow down the heterogeneity gap.

Another concern laid in existing localization frameworks is the attention module, which plays a pivotal role in learning cross-modal relationships. Generally, using the attention module to obtain a joint representation fusing audio and visual information can facilitate localization accuracy [8,10–13]. Nevertheless, these existing approaches often ignored the interference caused by the interaction between event-unrelated visual and audio features that are generated from *background* segments. Therefore, the approach in [19] applied a threshold on the cross-modal connections to filter out semantically similar feature pairs that share event-related information and used them to obtain a better-fused representation. However, the thresholding method treats all relevant feature pairs equally, ignoring that some of these pairs are temporally asynchronous. We argue that the asynchronous audio and visual feature pairs should also be distinguished since they inherently describe different states of the event. In fact, studies [20,21] on psychophysics have already proven that the temporal misalignment between vision and hearing can lead to the misinterpretation of one modality by another. Furthermore, the study [22] on multisensory representation learning also shows that judging whether visual and audio signals are synchronized can be used to learn representative multimodal features. Hence, our idea is to adaptively assign attention to event-related audio-visual connections by considering the time difference between modalities.

In light of the above analysis, this paper proposes a novel AVE localization framework consisting of two parts: a self-supervised audio-visual representation method called audio-visual Barlow Twins (AV-BT), and temporally adaptive cross-modal attention (TACMA). On the one hand, AV-BT takes both audio and visual data as inputs and trains the feature encoders using the intrinsic relation between audio and visual signals as supervision. It

helps to narrow the heterogeneity gap between the encoded features by matching the corresponding semantic information between synchronized audio and visual embeddings. On the other hand, a weighting method is introduced in TACMA to adaptively model the cross-modal relations by assigning different weights to audio-visual pairs with different time differences and semantic similarities. Specifically, the most attention is assigned between the synchronized visual and audio features which describe the same state of the event since they are the most relevant, while the association between those asynchronous signals (even if both contain the information of the same event) is relatively weakened because they are less relevant. The main contributions of this work are concluded as follows:

- We propose a self-supervised audio-visual representation method that encodes features in the same latent space and matches the semantic information contained in the audio and visual modalities. Such a design can narrow the heterogeneity gap between different modal features and benefit subsequent modeling of cross-modal relationships;
- We propose a weighting-based cross-modal attention module that dynamically weakens the connections between different modal features that are unrelated to events or temporally asynchronous;
- The proposed methods are combined in a framework to perform the event localization task on the public audio-visual event dataset. When directly classifying features learned by our self-supervised method for event localization, significant improvements are achieved in both unimodal and multimodal cases. When further combining these features with the proposed cross-modal attention, our overall approach achieves state-of-the-art localization accuracy.

The rest of this paper is organized as follows: Section 2 briefly introduces several related works about our method. In Section 3, the detailed system flow and discussions of AV-BT and TACMA are presented. Both the quantitative and qualitative results of our methods are given in Section 4. Conclusions and future directions are presented in Section 5.

2. Related Works

Based on the task and method of this paper, related works are briefly reviewed in three aspects: (1) Self-supervised audio-visual representation learning. (2) Multimodal fusion. (3) AVE localization.

2.1. Self-Supervised Audio-Visual Representation Learning

Self-supervised representation learning aims to train representation networks in the absence of human labeling. Existing self-supervised methods [23–28] fed two randomly distorted versions of an image into a backbone and maximized the similarity of their projected features. Their ideas of learning representations that are both invariant to random augmentations and distinctive to semantically different targets, inspire many works on self-supervised audio-visual learning [22,29–34].

Mimicking the self-supervised method for single-modal representation learning, some existing methods focus on designing pretext tasks, trying to exploit the cross-modal correlation effectively. Considering the co-occurrence of hearing and sounding, Owens et al. [22] proposed to predict whether video frames and audio are temporally aligned based on fused multisensory representations. By solving the pretext problem of audio-visual synchronization, Cheng et al. [29] trained a co-attention network in a self-supervised manner. Sarkar et al. [30] find that by relaxing the temporal synchronicity between modalities, more generalized representations can be learned. Different from defining a pretext task, Patrick et al. [31] shifted their attention to imposing transformations on multimodal data and combining them with existing noise-contrastive learning methods and achieved astonishing results on several downstream tasks. These studies show that a carefully designed self-supervised framework can help the network effectively uncover the joint representation of multimodal information by exploiting the general connection between different

modal data. In our work, we develop a framework that matches the corresponding semantic information between synchronized audio and visual signals to obtain multimodal features that have a smaller heterogeneity gap.

2.2. Multimodal Fusion

In the real world, many realistic tasks involve inputs of multimodal information that are complementary to each other. For example, the combination of speaking tone, the content of speech, and facial expressions are critical to properly judging a person's emotion. Therefore, integrating cues from different modalities, known as multimodal fusion, is necessary for building a stronger multimodal neural network to tackle real-world problems. Here we mainly review some of the commonly used fusion techniques in neural networks.

For information fusion in neural networks, many simple operations can be exploited, such as direct concatenation [35,36], element-wise multiplication [37,38], and weighted sums [39]. Since useful information about the data is often expressed in a highly abstract fashion in neural networks (which can be linearly related to the output predictions), applying these simple operations before the classification layer to directly combine high-level information from different sources can be effective. However, these simple operations often play limited roles because the complex relationships that may exist between different modalities can affect the representation of the data. Hence, modules that utilize the attention mechanism attract more interest [40,41]. Computed from multimodal cues, attention blocks that employ sets of scalar weights are more capable when modeling both inter-modal and intra-modal relationships. Lu et al. [42] proposed a co-attention model to extract correlated information from both visual and text modalities. Dou et al. [43] developed a merged-attention mechanism to ease the computational cost while also avoiding modeling the redundant information between modalities. In our solution, we do not restrict the information fusion process to a particular module. In early fusion, we encode informative multimodal features by self-supervised representation learning. For late fusion, we modify the process of cross-modal attention to make it more attentive to useful parts. In the classification layer, we directly employ summation for decision-level fusion.

2.3. AVE Localization

Nowadays, with the development of machine learning and deep learning, there are more and more cases of neural network techniques being used in practical applications [44]. As single-modal deep learning is becoming increasingly mature in applications [45–48], the machine learning community has turned to multimodal cases. For video content understanding based on visual and audio modalities, Tian et al. [6] first defined the AVE localization task, which is to detect whether an event that is both visible and audible happened in a video segment. Other than task definition, they also propose a complete set of solutions and initially study several basic fusion strategies of features. As the encoded feature of a video is sequential, Lin et al. [9] employ a bidirectional long short-term memory (Bi-LSTM) [49] network to better fuse the concatenated audio-visual feature. Although the localization task is built on the segment level, Wu et al. [7] consider the potential intra-modal misalignment within the segment so they propose a dual attention-matching module to model high-level global information over a longer range. Note that in their method, to compute global features, *background* segments need to be additionally supervised during training. Furthermore, in [10,11], Ramaswamy introduced several attention-based modules to interact between audio and visual both locally and globally. Xuan et al. [12] presented a different network with multiple attention modules to deal with the task under temporal inconsistency. Duan et al. [13] took a further step on multimodal attention and utilized joint co-attention for modeling both inter-modal and intra-modal relations, aiming to learn robust fused features by recursively stacking attention blocks. Lin et al. [50] complete the encoding of temporal features and information fusion simultaneously through their newly proposed audio-visual transformer. The above works emphasize a variety of attention-based designs, resulting in a large increase in computational cost when improv-

ing performance. Zhou et al. [19] revisit the localization task from a more concise perspective and propose a positive sample propagation module to only aggregate information from high-relevance audio and visual features. This cross-modal attention-based module selects positive audio-visual pairs that are of high attention scores through threshold operation. However, it cannot be ignored that most of the features that carry event-related information are inconsistent in time. Since this inconsistency reflects the fact that features from different modalities describe different states of the event, these asynchronous features should be distinguished from those feature pairs that are both synchronous and associated. Therefore, we apply weights on the attention map based on the time difference to integrate information more effectively by weakening the less relevant connections caused by inter-modal asynchrony.

3. The Proposed Framework

The whole scheme of our framework is depicted in Figure 2, in which our network is trained in two stages. First, the two-stream representation network is trained in a self-supervised manner to encode spatial–temporal features with the proposed AV-BT. At this point, the heterogeneity gap between features from different modalities is narrowed. Then, the extracted features are fed into the proposed TACMA to adaptively integrate useful information from each modality. To start with the description, the task statement is outlined in Section 3.1. Next, the detailed system flow and discussion of the AV-BT and TACMA are described in Sections 3.2 and 3.3, respectively.

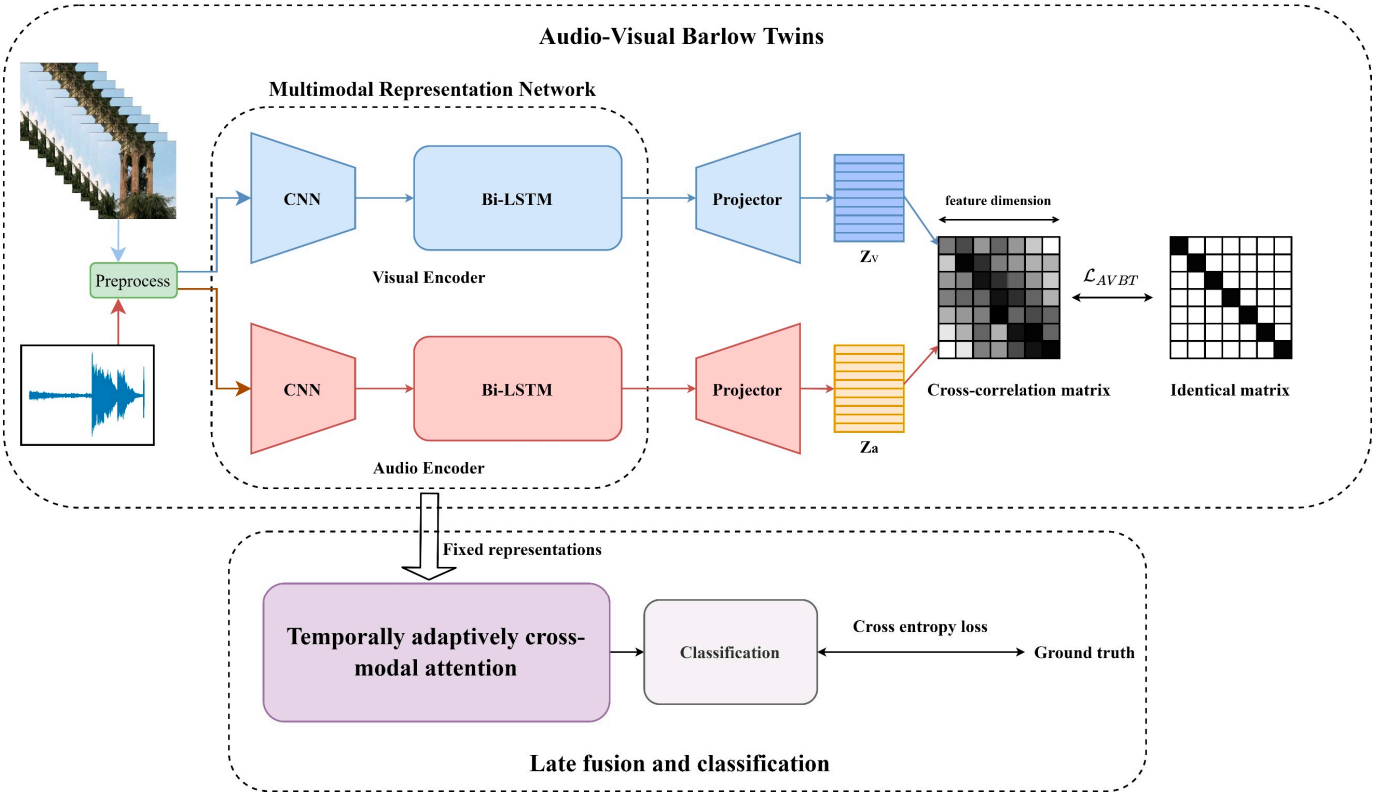

Figure 2. The pipeline of our proposal. The whole method is divided into two parts. We first train the audio-visual representation network using the proposed AV-BT. Then, the fixed representations produced by audio and visual encoders are sent to the proposed TACMA for late information fusion.

3.1. Task Statement

The meanings of the symbols used throughout this paper are provided in Table 1. Following [6], an AVE is defined as an event in a video clip that is both visible and audible. Performing the AVE localization task involves both predicting the temporal boundaries

and identifying the content information. Specifically, given a synchronous audio-visual video sequence $S = (S_a, S_v)$, where S_a and S_v denote the auditory portion and visual portion, respectively. S is continuously divided into T nonoverlapping segments of equal duration. As in [19,50], the annotation of the t-th segment $S^t = (S_a^t, S_v^t)$ is defined as $\boldsymbol{y}^t = \{y_c^t | y_c^t \in \{0, 1\}, \sum_c y_c^t = 1\} \in \mathbb{R}^C$, where C denotes the number of categories. If no event happens in a video segment, it will be labeled as *background*. Note that in our method, we directly classify the *background* as one of the categories and use no additional supervision to help the network distinguish it from events. Compared to an AVE, it is noticeably more challenging to identify a *background* because its audio and visual content are not correlated with each other.

Table 1. Main symbols used throughout the paper.

Symbols	Definition
S_a/S_v	Audio/visual portion of the video
$\boldsymbol{y}^t$	Annotation of the t-th video segment
$\boldsymbol{A}/\boldsymbol{V}$	Preprocessed audio/visual data
$\boldsymbol{f}_a/\boldsymbol{f}_v$	Encoded audio/visual feature
$\boldsymbol{Z}_a/\boldsymbol{Z}_v$	Projected audio/visual embedding
$\boldsymbol{\mathcal{C}}$	Cross-correlation matrix of $\boldsymbol{Z}_a/\boldsymbol{Z}_v$
$\boldsymbol{W}_a^1/\boldsymbol{W}_v^1$	Parameters for computing audio/visual query and key
$\boldsymbol{M}^{av}/\boldsymbol{M}^{va}$	Cross-modal attention matrix
$\boldsymbol{W}^{mv}/\boldsymbol{W}^{ma}$	Weight matrix for $\boldsymbol{M}^{av}/\boldsymbol{M}^{va}$
$\boldsymbol{W}^{av}/\boldsymbol{W}^{va}$	Weighted attention matrix
$\boldsymbol{W}_a^2/\boldsymbol{W}_v^2$	Parameters for computing audio/visual value
$\boldsymbol{A}^{fuse}/\boldsymbol{V}^{fuse}$	Attended audio/visual feature
$\boldsymbol{W}_a^3/\boldsymbol{W}_v^3$	Parameters for $\boldsymbol{A}^{fuse}/\boldsymbol{V}^{fuse}$
$\boldsymbol{f}_{a\leftarrow v}/\boldsymbol{f}_{v\leftarrow a}$	Fused audio/visual feature
$\boldsymbol{pred}$	Prediction of the input video

3.2. Self-Supervised Audio-Visual Representation

According to the above definition, the localization of an AVE requires a combination of information from both visual and audio signals. To extract useful information from the raw signal, we followed existing methods [6–13] to extract spatial features in video frames and audio spectrograms using two-dimensional CNN. Considering that the video is sequential, Bi-LSTMs are then used for the modeling of temporal relations.

Due to the scarcity of audio-visual video samples that can be used for training, the performance of training feature encoders from scratch is often unsatisfying, making it necessary to pretrain the CNN on large-scale datasets [14,15]. However, this pretraining also poses a problem: there is a huge heterogeneity gap that exists between the features extracted by the pretrained CNN and it naturally hinders information fusion. To overcome this problem, our solution is to adapt the CNN features from different modalities to the same semantic space. In practice, we accomplish the feature adaptation in a self-supervised manner by exploiting the relationship between vision and hearing as supervision. Specifically, there are two general cross-modal relationships in AVEs, namely, audio-visual *co-occurrence* and *correspondence*. On the one hand, *co-occurrence* refers to the fact that when an AVE happens in a video segment, its visual and auditory stimuli are provided simultaneously. For example, a baby's cry is accompanied by a sad facial expression. Audio-visual *correspondence*, on the other hand, means that the appearance of a particular object is associated with its characteristic sound. For example, when a bell appears in a scene, the neural network should match it to the ringing sound in the audio stream at that moment.

Utilizing these two attributes of AVE means that the network should match the relevant semantic information carried by the synchronized audio and visual signal pairs.

3.2.1. Audio-Visual Barlow Twins

Based on the above consideration, we modified the pipeline in [23] and design an AV-BT that trains the audio-visual encoder in a self-supervised manner to narrow the heterogeneity gap between different modal data. Figure 2 illustrates the full AV-BT pipeline, with the three steps involved detailed below.

Preprocessing. For each video segment, we convert the raw audio into a log-Mel spectrogram and randomly sample one frame from the visual stream. The resulting spectrograms and extracted frames of the whole video are denoted as $\boldsymbol{A} \in \mathbb{R}^{T \times H_a \times W_a}$ and $\boldsymbol{V} \in \mathbb{R}^{T \times H_v \times W_v \times C}$, respectively. Since random data augmentations can boost the performance of self-supervised learning and help to fight against overfitting [23–28,51], we employ similar strategies in AV-BT. Specifically, cropping and Gaussian blurring are randomly applied to the extracted visual frames. To simplify notation, $\boldsymbol{V}$ is assumed to be augmented. In addition, the random frame sampling technique also plays a role in data augmentation, as it can provide slight variations in visual content during each training epoch and helps the feature encoder capture more information along the timeline when the visual content remains stable. An in-depth analysis of the random sampling technique is provided in the next subsection.

Feature encoding. $\boldsymbol{A}$ and $\boldsymbol{V}$ are then sent to a two-stream audio-visual encoder simultaneously to learn spatial-temporal features. The output features are denoted as $\boldsymbol{f}_a \in \mathbb{R}^{T \times d_l}$ and $\boldsymbol{f}_v \in \mathbb{R}^{T \times d_l}$, where d_l denotes the feature dimension.

Projecting. Eventually, two three-layer multi-layer perceptrons (MLP) are used as projectors to map the $\boldsymbol{f}_a$ and $\boldsymbol{f}_v$ into the same latent space, where we compute their cross-correlation matrix afterward to explore the audio-visual relationships. We assign d as the output dimension of these projectors, so the output embeddings are denoted as $\boldsymbol{Z}_a \in \mathbb{R}^{T \times d}$ and $\boldsymbol{Z}_v \in \mathbb{R}^{T \times d}$. To simplify notations, $\boldsymbol{Z}_a$ and $\boldsymbol{Z}_v$ are assumed to be normalized along the time dimension after projection. It is necessary to note that these two projectors share the same parameters so that they can force the encoders to learn joint representations carrying similar semantic information about the video.

To calculate the objective function of AV-BT, we first compute the cross-correlation matrix $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{d \times d}$ of the normalized embeddings $\boldsymbol{Z}_a$ and $\boldsymbol{Z}_v$ along the time dimension. Specifically, the element in $\boldsymbol{\mathcal{C}}$ is obtained by the following equation:

$$\mathcal{C}_{ij} \triangleq \frac{\sum_t Z_{at,i} Z_{vt,j}}{\sqrt{\sum_t \left(Z_{at,i}\right)^2} \sqrt{\sum_t \left(Z_{vt,j}\right)^2}}, \tag{1}$$

where t indexes the time dimension and i, j indexes the feature dimension of $\boldsymbol{Z}_a$ and $\boldsymbol{Z}_v$ individually. Then we optimize the network by minimizing the difference between $\boldsymbol{\mathcal{C}}$ and the identity matrix, and the objective function is written as:

$$\mathcal{L}_{AVBT} \triangleq \underbrace{\sum_i \left(1 - \mathcal{C}_{ii}\right)^2}_{\textit{invariance term}} + \underbrace{\lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}{}^2}_{\textit{redundancy reduction term}}, \tag{2}$$

where λ is the same hyperparameter in the original method [23] and is used to trade off the importance between the first term (*invariance term*) and the second term (*redundancy reduction term*). Note that Equation (1) is defined for each video, and the total loss of a training step is obtained based on the average $\boldsymbol{\mathcal{C}}$ over all videos in a batch. By minimizing $\mathcal{L}_{AVBT}$, the representation network is trained to bridge the semantic information between synchronized audio and visual data. Another key point in Equation (1) is that our approximative identity correlation matrix is computed on the video level rather than the segment level. This is because of the existence of the *background* segment where audio and visual signals are not semantically associated with each other. In-depth analysis of

how the objective function $\mathcal{L}_{AVBT}$ guides the network to encode informative audio-visual representations is discussed in the next subsection.

The pseudocode of our AV-BT is shown in Algorithm 1.

Algorithm 1 PyTorch-style pseudocode for Audio-Visual Barlow Twins.

```
# f: two-stream audio-visual encoder and projectors
# p: preprocess the video
# lambda: weight on the off-diagonal term
# N: batch size
# D: dimensionality of projected embeddings
# transpose: transpose the last two dimensions of a tensor
# bmm: matrix-matrix multiplication with batch dimension at first
# off_dia: off-diagonal elements of a matrix
# eye: identity matrix

for video in loader: # load a batch with N videos
    # extract frames randomly and converts raw audio
    spectrograms, frames = p (video)

    # encode feature and project
    z_a, z_v = f (spectrograms, frames) # [N, T, D]

    # normalize along the time dimension
    z_a_norm = (z_a − − z_a.mean (1) )/z_a.std (1) # [N, T, D]
    z_v_norm = (z_v − − z_v.mean (1) )/z_v.std (1) # [N, T, D]

    # compute cross-correlation matrix
    c = bmm (transpose (z_a_norm), z_v_norm) # [N, D, D]
    c = c.sum (0)/N # [D, D]

    # loss
    diff = (c − eye (D) ).pow (2) # [D, D]
    off_dia (diff) .mul_(lambda)
    loss = diff.sum ()

    # optimize
    loss.backward ()
    optimizer.step ()
```

3.2.2. Discussion on Audio-Visual Barlow Twins

Insight of the $\mathcal{L}_{AVBT}$. First, we need to briefly explain the original design in [23]. The original objective function was applied to the projected embeddings of two identical networks fed with different distorted versions of an image. The outputs were forced to be invariant to the distortions by the *invariance term* while the *redundancy reduction term* decorrelated different components of the embedding so that the outputs comprised as much information as possible. Although our objective function is very close in form to the original one, there are differences and connections in many aspects.

First, the information from both audio and visual modalities needs to be considered for judging an AVE. Therefore, our cross-correlation matrix is computed through different modalities in the same latent space, while the original method only needs to consider

one modality. Second, similarly to the different distorted versions of an image, vision and sound can be regarded as two heterogeneous expressions of the same event. Hence in our $\mathcal{L}_{AVBT}$, the *invariance term* makes the features extracted from audio and visual streams both carry semantic information about the underlying target. Note that such a "target" refers to an object that may potentially have both visual and aural attributes. For example, for a person who appears in the scene, the semantic representation of the appearance extracted by the visual encoder should be tightly bound to the representation of the speaking voice extracted by the audio encoder. This is like how humans naturally associate the sound of speech when they see a person talking. In contrast, a table does not make any characteristic sound, so it is not audibly perceptible despite its visual characteristics. Therefore, an object like a table is treated as visual background and the encoder naturally ignores it. Third, since ambient noise and visual background in the environment do not carry meaningful semantic information across modalities, the *redundancy reduction term* forces the representations of both modalities to contain as little information about irrelevant interference as possible. This is analogous to the idea of the original method.

Examination of random frame sampling technique. It is worth noting that in Equation (1), the cross-correlation matrix is calculated along the time dimension for each sample in a batch (while it was computed along the batch dimension in [23]), which means that only the similarity of visual and audio embeddings from the same moment is measured. This is general to exploit audio-visual *co-occurrences* but inadequate for AVE representation owing to the possible contextual relevance between asynchronous audio and visual signals. For example, if the t-th segment and the $(t-1)^{th}$ segment contain the same AVE, obviously $\boldsymbol{Z}_{a,t}$ and $\boldsymbol{Z}_{v,t-1}$ are correlated. To fully describe our technique for compensating for this defect, an example is given as follows.

We first denote the t-th audio-visual segment pair as (S_a^t, S_v^t), and there are N frames in S_v^t, denoted as f_n^t where $n \in \{1, 2, \ldots, N\}$. Suppose that we extract f_1^t at random during a training epoch, such that the network is taught to match the information shared between f_1^t and S_a^t. Similarly, the network will learn to model pairs such as S_a^t, f_1^t and S_a^{t-1}, f_N^{t-1} by training with abundant epochs. Note that f_1^t and f_N^{t-1} are indeed two adjacent frames and only a minor difference exists between them (if visual content has not changed much in this period), such that the association within S_a^t, f_N^{t-1} can be broadcast through f_1^t. This technique allows the network to model the relation between adjacent segments to a certain extent. An intuitive example is demonstrated in Figure 3. Moreover, processing only one frame for each segment can help to fit the video memory of the GPU and this random sampling technique also helps to prevent overfitting.

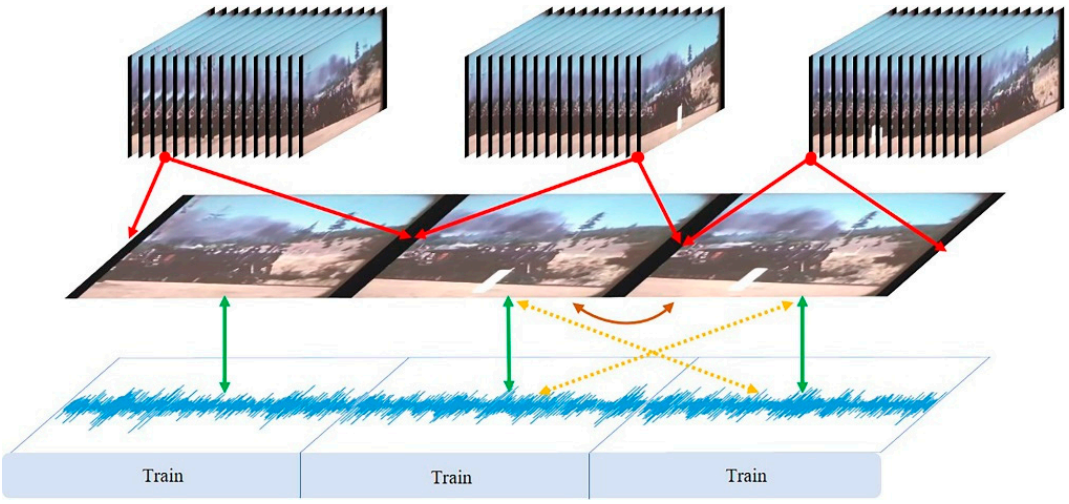

Figure 3. An illustration of the random sampling technique in AV-BT. Three continuous video segments segments that contain the "train horn" are depicted in this example. The random frame

sampling process is denoted as red arrows and the modeled cross-modal relations are denoted as green arrows. Since the frames picked from the last two segments are nearly identical, the audio-visual correlations across the last two segments can be broadcast through these frames, as indicated by the yellow arrows. *Best viewed in color*.

3.3. Temporally Adaptive Cross-Modal Attention

The attention mechanism was first proposed to deal with natural language processing tasks, and it is very capable of modeling the complex relationships within serialized tokens [52]. By considering features from each segment in the video as tokens, the attention module can also be modified to learn cross-modal relationships in AVE [8,10–13]. Furthermore, since task-related information is often contained in only some of the tokens, a threshold-based approach can better model the relationships between modalities sparsely by obtaining all the semantically similar feature pairs at all times [19]. However, unlike sentences, where words that are far apart may still be highly correlated, the changing state of events in the video is often continuous, namely, the audio (visual) signal at one moment is primarily correlated with its synchronized visual (audio) signal. Therefore, this paper designs a TACMA that makes one modality adaptively assign attention to useful information in another modality by considering their temporal difference. In such a way, the multimodal information is integrated along the timeline in a more fine-grained manner. The detailed structure of our adaptive attention is depicted in Figure 4.

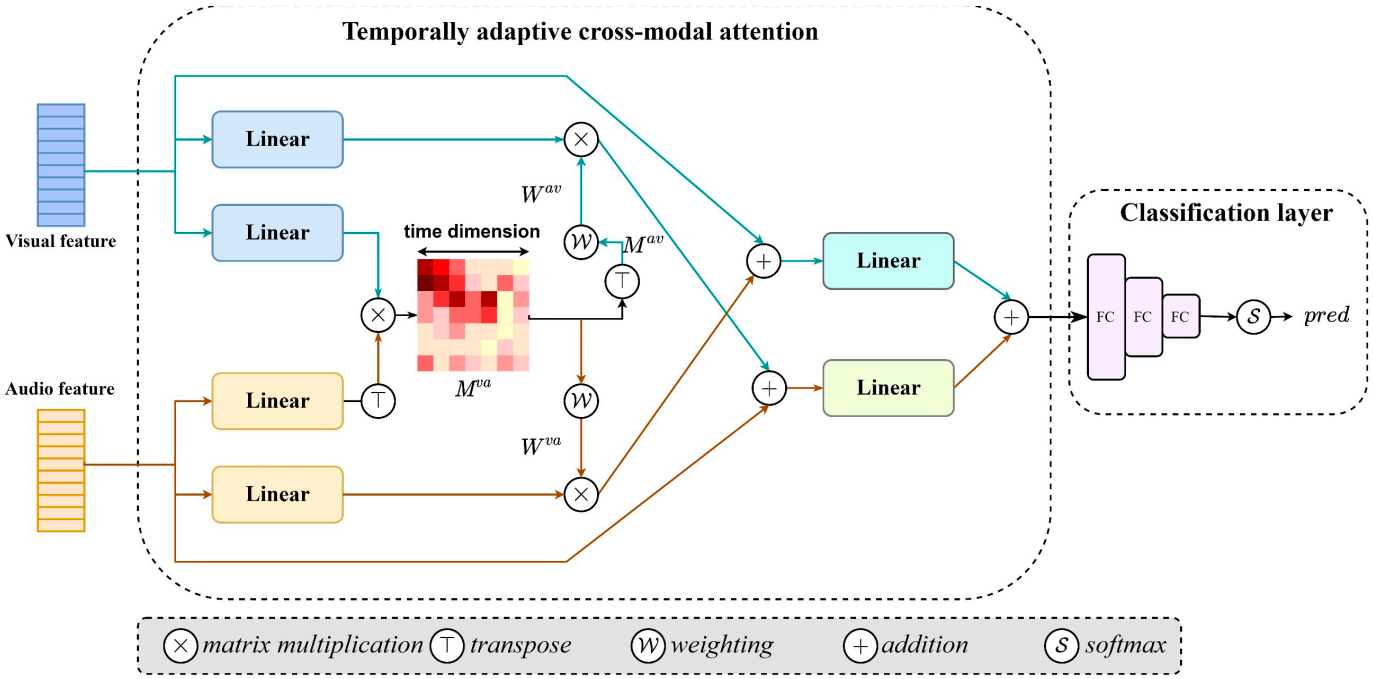

Figure 4. Structures of the proposed TACMA and classification layer.

Given the fixed f_a and f_v output from feature encoders as input, the first step is to compute the attention matrices $\boldsymbol{M}^{av} \in \mathbb{R}^{T\times T}$ and $\boldsymbol{M}^{va} \in \mathbb{R}^{T\times T}$ by the scaled dot-product along the feature dimension:

$$\boldsymbol{M}^{av} = (f_a \boldsymbol{W}_a^1)(f_v \boldsymbol{W}_v^1)^T / \sqrt{d_l}, \quad \boldsymbol{M}^{va} = (\boldsymbol{M}^{av})^T, \tag{3}$$

where $\boldsymbol{W}_a^1$ and $\boldsymbol{W}_v^1 \in \mathbb{R}^{d_l \times d_h}$ are learnable weight matrices implemented as linear layers. Note that we adopt a similar strategy as in [19] that uses the same set of *queries* and *keys* (which are both the linear projections of input features) in the computation of the attention matrices. This can help to reduce computational costs and ensure efficiency. ReLU activation is applied on $\boldsymbol{M}^{av}$ and $\boldsymbol{M}^{va}$ afterward to filter out negatively correlated audio and

visual feature pairs. Then, the softmax function is performed along each row of $\boldsymbol{M}^{av}$ and $\boldsymbol{M}^{va}$ to rescale the values between 0 and 1. The larger the elements in $\boldsymbol{M}^{av}$ and $\boldsymbol{M}^{va}$ are, the more event-related information that is shared between their corresponding audio and visual features.

To weaken the connections between features that share no event-related information or are temporally asynchronous, we weigh the scores of all the feature pairs in the attention matrices. Particularly, given $M^{av}_{i,j}$ and $M^{va}_{i,j}$ as the elements in the i-th row and the j-th column of the attention matrices, where i, j also index the time dimension of audio and visual features, respectively, their corresponding weights $W^{mv}_{i,j}$ and $W^{ma}_{i,j}$ are

$$\begin{cases} W^{mv}_{ij} \triangleq exp\Big(-\theta|i-j|/M^{av}_{i,j}\Big), \\ W^{ma}_{ij} \triangleq exp\Big(-\theta|i-j|/M^{va}_{i,j}\Big), \end{cases} \tag{4}$$

where θ is a hyperparameter and the magnitudes of $W^{mv}_{i,j}$ and $W^{ma}_{i,j}$ are adjusted by the ratio of θ to the corresponding element in $\boldsymbol{M}^{av}$ and $\boldsymbol{M}^{va}$. Hence, the overall weighting method is expressed as the element-wise product, written as:

$$\begin{cases} \boldsymbol{W}^{av} = \boldsymbol{W}^{mv} \odot \boldsymbol{M}^{av}, \\ \boldsymbol{W}^{va} = \boldsymbol{W}^{ma} \odot \boldsymbol{M}^{va}, \end{cases} \tag{5}$$

where $\boldsymbol{W}^{mv} \in \mathbb{R}^{T\times T}$ and $\boldsymbol{W}^{ma} \in \mathbb{R}^{T\times T}$ are the weighting matrices, and $\boldsymbol{W}^{av} \in \mathbb{R}^{T\times T}$ and $\boldsymbol{W}^{va} \in \mathbb{R}^{T\times T}$ are the resulting attention matrices. With the above equations, we generally weaken those connections between asynchronous audio and visual signals, but also retain the consideration of the semantic similarity between features. That is to say when an event in the video continues for a long time, i.e., when the features of different modalities carry similar information over a long period, we do not weaken these connections very substantially (even if some of the audio and visual features are far apart). In contrast, if adjacent features of different modalities do not share useful information about the event, then the connections between them are still significantly weakened.

Once the attention matrices are adjusted, $\boldsymbol{W}^{av}$ and $\boldsymbol{W}^{va}$ are used to update the linear projected audio and visual features (which are also known as the *value* in the attention mechanism) by matrix multiplication, and features from each modality are summed together to further reduce their differences in amplitude and direction. The vectors that carry fused information are updated as

$$\begin{cases} \boldsymbol{V}^{fuse} = \boldsymbol{W}^{va}\Big(\boldsymbol{f}_v \boldsymbol{W}^2_v\Big) + \boldsymbol{f}_a, \\ \boldsymbol{A}^{fuse} = \boldsymbol{W}^{av}\Big(\boldsymbol{f}_a \boldsymbol{W}^2_a\Big) + \boldsymbol{f}_v, \end{cases} \tag{6}$$

where $\boldsymbol{W}^2_a$, $\boldsymbol{W}^2_v \in \mathbb{R}^{d_l\times d_l}$ and $\boldsymbol{V}^{fuse}$, $\boldsymbol{A}^{fuse} \in \mathbb{R}^{T\times d_l}$. Before being sent to the classification layer, we linear transform $\boldsymbol{V}^{fuse}$ and $\boldsymbol{A}^{fuse}$ followed by layer normalization, written as:

$$\begin{cases} \boldsymbol{f}_{v\leftarrow a} = layernorm\Big(\boldsymbol{V}^{fuse}\boldsymbol{W}^3_v\Big), \\ \boldsymbol{f}_{a\leftarrow v} = layernorm\Big(\boldsymbol{A}^{fuse}\boldsymbol{W}^3_a\Big) \end{cases} \tag{7}$$

where $\boldsymbol{W}^3_a$, $\boldsymbol{W}^3_v \in \mathbb{R}^{d_l\times d_e}$ and $\boldsymbol{f}_{v\leftarrow a}$, $\boldsymbol{f}_{a\leftarrow v} \in \mathbb{R}^{T\times d_e}$. Compared to AV-BT, TACMA can further improve the overall performance especially when the visual content changes rapidly since it constructs all pair relationships by attention matrices.

3.4. Classification and Optimization

After the information from both modalities is further fused by TMCMA, we simply add $f_{v\leftarrow a}$ and $f_{a\leftarrow v}$ together and the final prediction is made by a three-layer MLP, written as follows:

$$\boldsymbol{pred} = MLP(f_{v\leftarrow a} + f_{a\leftarrow v}) \tag{8}$$

where $\boldsymbol{pred} \in \mathbb{R}^{T\times C}$. Since the ground truth is denoted as $\boldsymbol{Y}^{gt} = \left[\boldsymbol{y}^0, \boldsymbol{y}^1, \ldots, \boldsymbol{y}^{T-1}\right] \in \mathbb{R}^{T\times C}$, we minimize the cross entropy (CE) loss between $\boldsymbol{pred}$ and $\boldsymbol{Y}^{gt}$ to optimize our adaptive attention network during training, written as

$$\mathcal{L}_{CE} = \frac{1}{TC}\sum_{t=1}^{T}\sum_{c=1}^{C} \boldsymbol{Y}_{t,c}^{gt} log(\boldsymbol{pred}_{t,c}) \tag{9}$$

where t and c index the time dimension and category, respectively. Compared to the threshold-based attention in [19], our method does not require additional loss items to supervise the challenging *background* segment but only uses the basic CE loss.

4. Experiments

4.1. Experiment Setup

Dataset. For the AVE localization task, we conduct our experiments on the AVE dataset which is introduced in [6] and consists of 4143 videos, covering 28 event categories (e.g., human activities, instrument playing, vehicle sounds, etc.). All videos are temporally labeled with audio-visual boundaries on the segment level. We follow the default split where the train/validation/test sets are 3339/402/402 clips, respectively. All clips are divided into 10 segments, with each segment lasting for one second. Whether training AV-BT or TACMA, both networks only have access to the training set.

Evaluation. We follow the routing in [7,10,11,13,19,50] and consider global segment-wise classification accuracy as the evaluation metric. Specifically, for evaluating TACMA, the predictions made by the classification layer are used for evaluation. To evaluate the audio-visual feature learned by AV-BT, we train another three-layer MLP to directly produce predictions based on the fixed feature without any other late fusion modules. All of our results are reported on the test set.

Architecture. During preprocessing before feature extraction, we sample 160 RGB frames at a resolution of 256 $\times$ 256 from each 10 s video. Then for each 1 s long segment that contains 16 frames, we randomly pick one frame for each training epoch. For raw audio, we sample it at 16,000 Hz with an equal duration of 10 s for each video. After that, a short-time Fourier transform (STFT) is performed and the signal in the frequency domain is then transformed into Mel-scale filter banks with 64 bins. The resulting spectrogram is 96 $\times$ 64 for every segment. For the random augmentations applied to the frames, we use the same settings as in [23].

Considering the size of the AVE dataset is too small to train the feature encoder from scratch, ResNet-50 [53] (removing the final fully connected layer and retaining the 512-dimensional representation) pretrained on ImageNet [14], and VGGish [54] (outputting 128-dimensional representation) pretrained on AudioSet [15], are employed as CNN backbones for extracting spatial features. In AV-BT, the projector has three linear layers, and the output dimensions of each are 2048/4096/4096. For TACMA, the parameter θ in Equation (4) is set to 0.03 to achieve the best performance.

Optimization. For training AV-BT, we use the LARS optimizer [55] and train 100 epochs with a batch size of 128. The learning rates of weights and biases in the representation network are the same in [23] except that we multiply them by 2e-4. We reduce the learning rate by a factor of 10 using a cosine decay schedule after using 10 epochs of warm-up period [56]. In Equation (2), λ is set to 5e-3. For training TACMA, we use an initial learning rate of 1e-4 and multiply it by 0.98 after every 10 epochs, and the batch size is 128 using the Adam [57] optimizer.

4.2. Results

4.2.1. Evaluating AV-BT

AVEL in [6] adopts an audio-guided visual attention module (AGVA) which interacts between the audio and the visual encoder. It uses audio information to help the visual encoder focus on the sounding target to better filter out interference caused by the visual background. Apart from the AGVA module, the structures of our encoders are quite similar to those of AVEL. Particularly, the CNN backbone used for audio feature extraction is identical while the backbone for visual features has comparable capabilities. Even if AVEL is trained with full labels in an end-to-end manner, we still think it is reasonable to compare with it since the following research [12,13,19] directly adopted it to extract features.

Comparison with existing methods. Analogous to the protocols in self-supervised image representation learning [23–28], we directly train a classifier on top of the fixed representations produced by the encoders trained with AV-BT. Unlike using a single linear layer as the prediction head for single-modal representations, sequential audio-visual representations are much more complex, so we adopt a three-layer MLP (which has the same structure as the classification layer in Figure 4) to make predictions. The results are reported in Table 2, where **our AV-BT significantly outperforms the supervised method in both unimodal and multimodal cases**. Specifically, in the case where the CNN backbone used for audio feature extraction is identical, our method surpasses AVEL by 4% when using only audio data. This improvement in the unimodal case suggests that by maximizing the similarity of different modal features in the same latent space, features from each modality are enhanced to be more discriminative. An intuitive explanation is that by reducing the differences between different modal representations, both the visual and auditory properties of an object can be considered more collaboratively. For example, mandolins and guitars may be very close in appearance, but they are very different in sound. Combining both the visual and audio characteristics of the mandolin allows it to be better distinguished from the guitar. Moreover, the results also show that the improved separability of the unimodal feature also further enhances the performance in the multimodal case.

Table 2. Evaluation of AV-BT. Our results are in **bold**. Note that we use no extra module for fusion at this point, but simply add audio and visual features together to make classifications. For comparison, the result of AVEL in [6] is obtained by end-to-end fully supervised training and they used the same strategy of addition before prediction.

Method	Accuracy (%)
Audio only [6]	59.5
Visual only [6]	55.3
AVEL [6]	71.3
Ours (Audio only)	**63.5**
Ours (Visual only)	**61.0**
Ours (Audio-Visual)	**75.6**

4.2.2. Evaluating TACMA

PSP in [19] is one of the state-of-the-art solutions for AVE localization. It adopts a threshold-based attention style module to aggregate information from all the similar audio and visual features. This module is highly reliant on applying additional supervision for the *background* to force the network to learn appropriate features for thresholding.

Comparison with state-of-the-art. By training the proposed TACMA based on fixed representations produced by AV-BT, we decoupled multimodal data representation and late fusion in the AVE localization task. From this perspective, it increases the difficulty of adapting from the source domain to the target domain while the solutions we compare are all combined late fusion networks with feature encoders, and are trained in an end-to-end manner supervised with full labels. Nevertheless, **our overall method still achieves state-of-the-art performance** on the AVE localization task, as shown in Table 3. This result suggests that channeling attention according to both the time difference and seman-

tic similarity between features encoded from AV-BT can further improve the quality of fused features.

Table 3. Comparing the proposed method to state-of-the-art methods. Our result is in **bold**.

Method	Accuracy (%)
CMAN [12]	73.3
AVRB [10]	74.8
AVIN [11]	75.2
MPN [8]	75.2
JCAN [13]	76.2
PSP [19]	76.6 *
AVT [50]	76.8
AV-BT+TACMA (Ours)	**77.2**

* For a fair comparison, the result of PSP is obtained by training without audio-visual pair similarity loss, which applies additional supervision for differentiating the *background*.

Detailed comparison with AV-BT. In addition to comparing with the latest methods, we would also like to know the detailed performance improvement of TACMA compared to AVBT. Hence, we show the confusion matrices of our solutions in Figure 5. From Figure 5a we can observe that most of the categories can be well represented by AV-BT. However, there are still many misclassifications between events and the *background*. After further fusing information with TACMA, the misclassification of events as the *background* improves for almost all categories, indicating that TACMA can grasp more event-related information since it dynamically considers all connections between different modal features to integrate useful cues, as shown in the bottom row in Figure 5b. In addition, for certain categories with very few samples, TACMA also improves the accuracy based on AV-BT. For example, index 11: "Truck", index 12: "Shofar", index 13: "Motorcycle", index 22: "Cat" and index 23: "Horse".

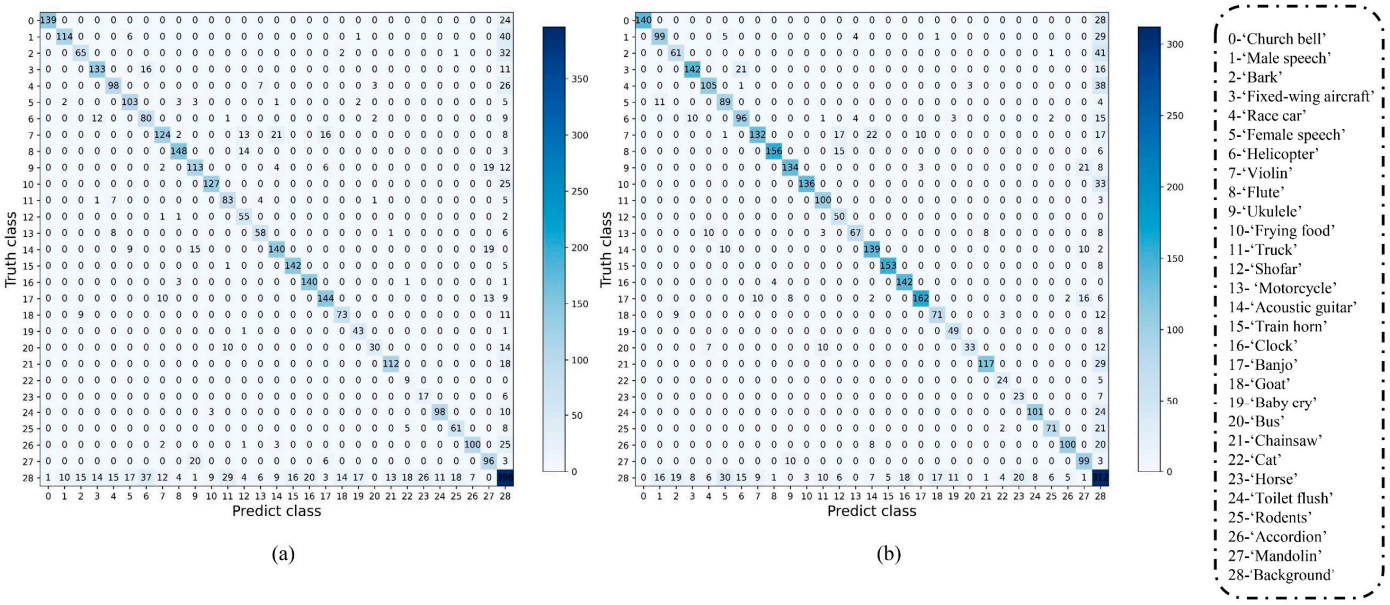

Figure 5. The confusion matrices of the classification results based on the representations of AV-BT and the whole proposed method are shown in (**a**) and (**b**), respectively. The classes corresponding to the numbers in the axes of the confusion matrix are shown in the dashed boxes on the right.

4.3. Ablations

In this section, we delve into the designs of our method and explore their influences. It is necessary to clarify that we employ our results reported in the previous subsection as the baseline of the ablation study.

4.3.1. Architecture of AV-BT

Weight sharing. The two projectors share the same parameters by default in AV-BT. One may wonder how heterogenous audio and visual features can be mapped to the same latent space by a single MLP. For this problem, we cut off the weight sharing between two projectors to deeply explore our AV-BT, denoted as "w/o sharing projector" in Table 4.

Table 4. Ablations on the architecture of the representation network. The best performance is in **bold**. Note that when applying an interaction module such as AVGA, the feature extraction network is no longer able to encode features under single modal data input.

Modification	Audio-Visual	Audio Only	Visual Only
Baseline	**75.6**	**63.5**	**61.0**
w/AGVA	71.5	-	-
w/o sharing projector	72.5	63.0	60.4

Attention. Sound source localization on the visual portion is one of the optional demands in the AVE localization task. Most existing approaches ambiguously localize the event-related sounding object by a heatmap generated in the attention block as a byproduct. Although the AVE dataset does not have bounding boxes as annotations (so the results of localization cannot be quantified for comparison) and localizing sound sources is not in the scope of this paper, we still wonder if we could improve feature quality by integrating attention modules such as AGVA. We denote this setting as "w/AGVA" in Table 4.

Analysis. As discussed above, the architecture of AV-BT is adjusted in two ways and Table 4 recapitulates these modifications tested along with their results. When disabling the weight sharing between projectors, the performance dropped in the unimodal case. In addition, the quality of audio-visual joint representation is significantly worse. We argue that the point here is not about the method of mapping heterogeneous data. Using the same MLP for projection forces the encoders to learn representations that carry as similar semantic information as they possible can. Another phenomenon that can be observed from Table 4 is that spatial attention modules such as AGVA, which are successful in other methods, conversely fail in our case. We argue that because our representation network is trained in a self-supervised way, the attention block is not guided by event labels, thereby the network is uncertain about which sounding objects should be focused on. Especially when encountering a background segment, AGVA may further corrupt the representations since it may attend to an interfering area. Multi-head co-attention may be a potential solution to the visual localization problem because it allows the construction of complex many-to-many audio-visual relationships spatially, and we will investigate such a design in the future.

4.3.2. Data Augmentations in AV-BT

Data augmentations. Popular self-supervised methods such as [23,26,27] are competitive with supervised learning partly thanks to their random data augmentations. Random cropping, horizontal flipping, color jittering, Gaussian blurring, solarizing, and converting to grayscale are among the most used methods. We also study their influences in our self-supervised method. Multiple representative experimental results are presented in Table 5.

Analysis. Compared to unimodal self-supervised learning cases, our results in Table 5 suggest that most of the commonly used augmentations (except color jittering) do not make a huge difference in our case. Different from a single image, AVE has a stricter definition that requires the target to be both seen and heard. Adding strong interruptions on data may break the agreed rules. An extreme case in Table 5 is that when applying color jittering, performance is badly damaged. We argue that altering the exposure and contrast ratio of a visual frame too much would make the target very likely to be undetectable in a noisy environment. For example, a black church bell ringing in almost complete darkness in an attic. Before the experiment, we suspect that cropping too many regions might result in

losing the target in a frame. However, as the results shown in the first two rows of Table 5 indicate, a larger cropping ratio helps to train stronger encoders. After delving deeper into the dataset, we found that certain kinds of events were still defined as visible when only a portion of the sounding object was observed, such as the nose of an airliner. Ultimately, the best result is achieved by randomly adopting cropping and Gaussian blurring.

Table 5. Ablations on adding random data augmentations in our AV-BT. The best result is in **bold.** Note that * denotes that the crop scale is adjusted to be relatively smaller. The original intention for this was to prevent cropping out the entire sounding target, resulting in breaking the agreed rule of an AVE, i.e., visible and audible at the same time.

Crop	Flip	Solarize	Grayscale	Jitter	Blur	Accuracy (%)
✓ *	✓	✓	✓	✓	✓	74.0
✓	✓	✓	✓	✓	✓	74.5
✓	✘	✓	✓	✓	✓	74.2
✓	✘	✘	✓	✓	✓	74.3
✓	✘	✘	✘	✓	✓	74.4
✓	✘	✘	✘	✘	✓	**75.6**
✓	✘	✘	✘	✘	✘	74.3
✓	✓	✘	✘	✘	✓	74.7
✓	✓	✓	✓	✘	✓	74.5
✘	✘	✘	✘	✘	✘	74.4

4.3.3. Ablation on θ in TACMA

In this section, we further study how the proposed weighting operation influences the performance of TACMA by gradually changing the hyperparameter θ in Equation (4). The corresponding results are shown in Figure 6. The overall localization accuracy reaches a high level when θ is set in the range from 0.01 to 0.06, and peak performance is achieved when θ = 0.03, indicating that the proposed weighting method contributes a lot to modeling cross-modal relations by discriminating those less-relevant audio-visual connections that are unrelated to the event or temporally asynchronous. In addition, when θ = 0, the proposed weighting method is completely disabled and the whole design degrades to standard cross-modal attention which has a very limited contribution to the localization performance by only improving by 0.4% compared to the result in Table 2. This shows that such a cross-modal attention module, without considering the time difference and semantic similarity, is not very capable of fusing audio and visual information.

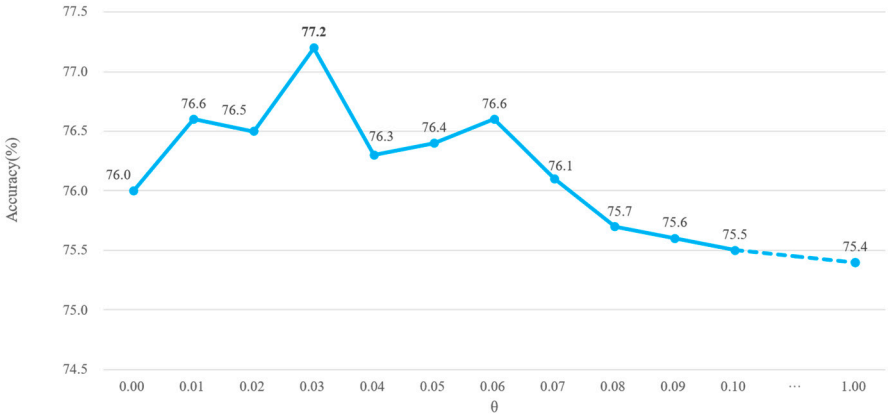

Figure 6. Illustration of the effect of the hyperparameter θ in TACMA. The best result is in **bold**.

The curvature of Equation (4) is controlled by θ and the similarity of the feature pair together so that θ is playing a similar role of thresholding that filters out those feature pairs with smaller similarity scores. This is close to the idea in PSP [19], while our weighting operation works in conjunction with the similarity determined by the feature itself instead

of counting on an external loss item to guide it. This makes our attention module more adaptive, and our results verify the effectiveness.

4.4. Visualization

Feature visualization. To verify that our methods can learn discriminative audio-visual features, we visualize several classes of the features learned by AV-BT and TACMA with the high-dimensional data visualization tool, t-SNE [58], as shown in Figure 7. Specifically, we selected features from categories covering diverse targets, such as people, animals, vehicles, and environmental objects, to demonstrate the adaptability of our approach to a variety of situations. Despite its relatively small sample numbers for a certain class, such as index 23: "Horse", good within-class cohesion is still achieved in both audio and visual features. In addition, the discrimination between classes is also greatly enhanced after further information fusion by TACMA, and this is particularly remarkable for the visual stream.

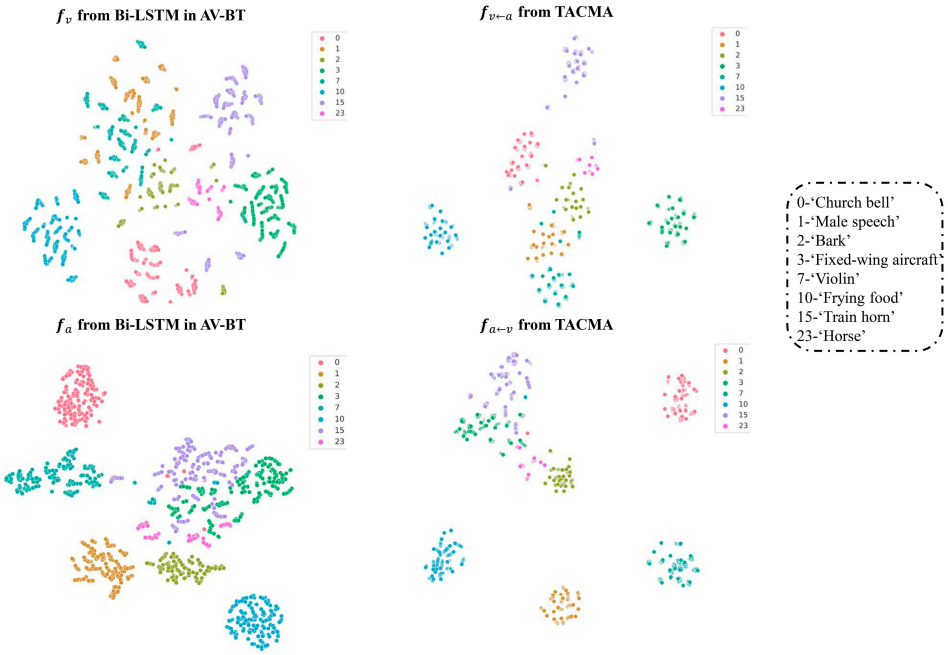

Figure 7. Illustration of the feature learning of our methods. The classes corresponding to the numbers in the legends are shown in the dashed boxes on the right. *Best viewed in color*.

Qualitative results. We evaluated our whole method with several interesting samples from the AVE dataset to verify the design. Predictions made by our networks are presented in Figure 8. Through these representative samples, we can intuitively gain a better understanding of our method, especially how TACMA coordinates with AV-BT for superior performance.

First, we observed that both of our proposed methods can detect various AVEs in the real world, including human speaking, vehicle sounds, environmental object activities, etc., as shown in in Figure 8b–d. Even if some samples are exceptionally challenging, for example, the visual quality of (e) in Figure 8 is very poor and the sound of the helicopter is almost completely distorted by extreme wind noise, our method still holds up to a certain extent. Generally, based on AV-BT, more accurate predictions can be made after fusion by TACMA.

Second, our networks only represent the data from a semantic perspective. This conclusion can be obtained from a typical case in Figure 8a, where there is a ringing across the whole timeline while only the first appearing bell is emitting sound. When the camera's view is turned to the second bell, both AV-BT and TACMA still treat the sounding objects as visible (since a bell is still visible even though it's not the one ringing at this

moment). We argue that adopting mono audio signals for feature encoding makes it difficult to solve this kind of problem because of the lack of information for tracking sounding targets spatially.

Last, we designed TACMA to integrate more useful information from the audio-visual pairs that are of various time intervals, and Figure 8f proves that our design meets such expectations to a certain extent. In that example, a man is trimming branches on a tree with a chainsaw and the video is recorded from a first-person perspective. Since the visual content is very unstable and the chainsaw is only captured in very few frames, it is very difficult for AV-BT to deal with such rapidly changing visual content (resulting in predicting all segments as *background*). After further aggregating useful cues with TACMA, the network can grasp more event-related information.

Figure 8. Qualitative results of our methods on the AVE dataset are shown in (**a**–**f**). For each sample, the annotations of the data are given in the first row. The purple box on the audio wave diagram indicates that the sound emitted by the target is audible, while the orange box on the frame represents the sounding object (person) that is visible at this moment. Predictions directly made by features from AV-BT are shown in the fourth row, and the results of TACMA are shown in the last row. 'bg' denotes *background*, and 'heli.' in (**e**) represents helicopter. *Best viewed in color*.

5. Conclusions

In this paper, we proposed a self-supervised audio-visual representation method and temporally adaptive cross-modal attention for AVE localization. When representing video data, the audio and visual features are encoded in the same space with their association explored in a self-supervised manner. When modeling cross-modal relationships, attention is channeled adaptively by considering both the time difference and semantic similarity between features. Our proposed framework performs better than previous event localization methods on the public AVE dataset and the results show that: (1) more discriminative features can be encoded by narrowing the heterogeneity gap between modalities with our

self-supervised method. (2) Our strategy for adaptively assigning attention by considering the temporal asynchrony of event-related features can effectively fuse the information across modalities.

6. Future Direction

Although our proposed framework achieves competitive AVE localization accuracy and the effectiveness of our design is verified by experiments, there is still room for further improvement. Specifically, the visualization result shows that both AV-BT and TACMA focus on representing the multimodal data and cross-modal relationships from a semantic perspective, which makes our network lack the ability to track the sound source spatially. This limits the event localization performance, especially in the face of multiple objects that may emit sound. A potential solution such as multi-head co-attention which can construct complex audio-visual relations spatially can be studied in the future. Moreover, as the task of AVE localization provides an intuitive form of multimodal video analysis, we will also investigate whether our proposed self-supervised method with improved encoder structures can be transferred to more tasks such as video parsing and object segmentation.

Author Contributions: Conceptualization, Y.R. and H.T.; methodology, Y.R.; software, Y.R.; data curation, Y.R.; writing—original draft preparation, Y.R. and H.T.; writing—review and editing, Y.R., H.T. and G.W.; supervision, B.L. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A publicly available dataset was analyzed in this study. These data can be found here: https://github.com/YapengTian/AVE-ECCV18 (accessed on 27 March 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, S.; Xu, K.; Jiang, X.; Sun, T. Pyramid Spatial-Temporal Graph Transformer for Skeleton-Based Action Recognition. *Appl. Sci.* **2022**, *12*, 9229. [CrossRef]
2. Gowda, S.N.; Rohrbach, M.; Sevilla-Lara, L.; Assoc Advancement Artificial, I. SMART Frame Selection for Action Recognition. In Proceedings of the 35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence, Electr Network, Virtually, 2–9 February 2021; pp. 1451–1459.
3. Park, S.K.; Chung, J.H.; Pae, D.S.; Lim, M.T. Binary Dense SIFT Flow Based Position-Information Added Two-Stream CNN for Pedestrian Action Recognition. *Appl. Sci.* **2022**, *12*, 10445. [CrossRef]
4. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
5. Koutini, K.; Eghbal-zadeh, H.; Dorfer, M.; Widmer, G. The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification. In Proceedings of the 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019.
6. Tian, Y.P.; Shi, J.; Li, B.C.; Duan, Z.Y.; Xu, C.L. Audio-Visual Event Localization in Unconstrained Videos. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 252–268.
7. Yu, W.; Linchao, Z.; Yan, Y.; Yi, Y. Dual Attention Matching for Audio-Visual Event Localization. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6291–6299.
8. Yu, J.; Cheng, Y.; Feng, R. Mpn: Multimodal parallel network for audio-visual event localization. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
9. Lin, Y.B.; Li, Y.J.; Wang, Y.C.F. Dual-Modality Seq2seq Network for Audio-Visual Event Localization. In Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2002–2006.
10. Ramaswamy, J.; Das, S.; Soc, I.C. See the Sound, Hear the Pixels. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 2959–2968.

11. Ramaswamy, J. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 4372–4376.
12. Xuan, H.Y.; Zhang, Z.Y.; Chen, S.; Yang, J.; Yan, Y.; Assoc Advancement Artificial, I. Cross-Modal Attention Network for Temporal Inconsistent Audio-Visual Event Localization. In Proceedings of the 34th AAAI Conference on Artificial Intelligence/32nd Innovative Applications of Artificial Intelligence Conference/10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 279–286.
13. Duan, B.; Tang, H.; Wang, W.; Zong, Z.L.; Yang, G.W.; Yan, Y. Audio-Visual Event Localization via Recursive Fusion by Joint Co-Attention. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Electr Network, Virtual, 5–9 January 2021; pp. 4012–4021.
14. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE-Computer-Society Conference on Computer Vision and Pattern Recognition Workshops, Miami Beach, FL, USA, 20–25 June 2009; pp. 248–255.
15. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
16. Smith, L.; Gasser, M. The development of embodied cognition: Six lessons from babies. *Artif. Life* **2005**, *11*, 13–29. [CrossRef] [PubMed]
17. Schwartz, J.-L.; Berthommier, F.; Savariaux, C. Audio-visual scene analysis: Evidence for a" very-early" integration process in audio-visual speech perception. In Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002.
18. Omata, K.; Mogi, K. Fusion and combination in audio-visual integration. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2008**, *464*, 319–340. [CrossRef]
19. Zhou, J.X.; Zheng, L.; Zhong, Y.R.; Hao, S.J.; Wang, M.; Ieee Comp, S.O.C. Positive Sample Propagation along the Audio-Visual Event Line. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Nashville, TN, USA, 19–25 June 2021; pp. 8432–8440.
20. Sekuler, R. Sound alters visual motion perception. *Nature* **1997**, *385*, 308. [CrossRef]
21. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef]
22. Owens, A.; Efros, A.A. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *Computer Vision—ECCV 2018, Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018*; Proceedings: Lecture Notes in Computer Science (LNCS 11210); Springer: Cham, Swizerland, 2018; pp. 639–658.
23. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, Online, 18–24 July 2021; pp. 12310–12320.
24. Kaiming, H.; Haoqi, F.; Yuxin, W.; Saining, X.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9726–9735.
25. Grill, J.-B.; Strub, F.; Altche, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent a new approach to self-supervised learning. In Proceedings of the 34th Conference on Neural Information Processing Systems, NeurIPS 2020, Virtual Online, 6–12 December 2020.
26. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15750–15758.
27. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning (ICML), Electr Network, Vienna, Austria, 13–18 July 2020.
28. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Condens. Matter Phys.* **2020**, *33*, 9912–9924.
29. Ying, C.; Ruize, W.; Zhihao, P.; Rui, F.; Yuejie, Z. Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning. In Proceedings of the MM '20: Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3884–3892.
30. Sarkar, P.; Etemad, A. Self-Supervised Audio-Visual Representation Learning with Relaxed Cross-Modal Temporal Synchronicity. *arXiv* **2021**, arXiv:2111.05329.
31. Patrick, M.; Asano, Y.M.; Kuznetsova, P.; Fong, R.; Henriques, J.F.; Zweig, G.; Vedaldi, A. On compositions of transformations in contrastive self-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9577–9587.
32. Yang, K.; Russell, B.; Salamon, J. Telling Left From Right: Learning Spatial Correspondence of Sight and Sound. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9929–9938.
33. Terbouche, H.; Schoneveld, L.; Benson, O.; Othmani, A. Comparing Learning Methodologies for Self-Supervised Audio-Visual Representation Learning. *IEEE Access* **2022**, *10*, 41622–41638. [CrossRef]

34. Feng, Z.S.; Tu, M.; Xia, R.; Wang, Y.X.; Krishnamurthy, A. Self-Supervised Audio-Visual Representation Learning for in-the-wild Videos. In Proceedings of the 8th IEEE International Conference on Big Data (Big Data), Electr Network, Atlanta, GA, USA, 10–13 December 2020; pp. 5671–5672.
35. Arandjelovic, R.; Zisserman, A. Look, listen and learn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 609–617.
36. Parekh, S.; Essid, S.; Ozerov, A.; Duong, N.Q.; Pérez, P.; Richard, G. Weakly supervised representation learning for unsynchronized audio-visual events. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2518–2519.
37. Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; Morency, L.-P. Deep multimodal fusion for persuasiveness prediction. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 284–288.
38. Wang, H.; Meghawat, A.; Morency, L.-P.; Xing, E.P. Select-additive learning: Improving generalization in multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 949–954.
39. Pérez-Rúa, J.-M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; Jurie, F. Mfas: Multimodal fusion architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6966–6975.
40. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
41. Tay, Y.; Dehghani, M.; Aribandi, V.; Gupta, J.; Pham, P.M.; Qin, Z.; Bahri, D.; Juan, D.-C.; Metzler, D. Omninet: Omnidirectional representations from transformers. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10193–10202.
42. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. *Adv. Condens. Matter Phys.* **2016**, *29*, 289–297.
43. Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 18166–18176.
44. Sharma, S.; Mittal, R.; Goyal, N. An Assessment of Machine Learning and Deep Learning Techniques with Applications. *ECS Trans.* **2022**, *107*, 8979–8988. [CrossRef]
45. Popli, R.; Sethi, M.; Kansal, I.; Garg, A.; Goyal, N. Machine learning based security solutions in MANETs: State of the art approaches. *J. Phys. Conf. Ser.* **2021**, *1950*, 012070. [CrossRef]
46. Popli, R.; Kansal, I.; Garg, A.; Goyal, N.; Garg, K. Classification and recognition of online hand-written alphabets using Machine Learning Methods. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *2021*, 012111–012119. [CrossRef]
47. Gautam, V.; Trivedi, N.K.; Singh, A.; Mohamed, H.G.; Noya, I.D.; Kaur, P.; Goyal, N. A Transfer Learning-Based Artificial Intelligence Model for Leaf Disease Assessment. *Sustainability* **2022**, *14*, 19. [CrossRef]
48. Verma, V.; Gupta, D.; Gupta, S.; Uppal, M.; Anand, D.; Ortega-Mansilla, A.; Alharithi, F.S.; Almotiri, J.; Goyal, N. A Deep Learning-Based Intelligent Garbage Detection System Using an Unmanned Aerial Vehicle. *Symmetry* **2022**, *14*, 15. [CrossRef]
49. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
50. Yan-Bo, L.; Wang, Y.C.F. Audiovisual Transformer with Instance Attention for Audio-Visual Event Localization. In *Computer Vision—ACCV 2020, Proceedings of the 15th Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020*; Revised Selected Papers; Lecture Notes in Computer Science (LNCS 12647); Springer: Cham, Swizerland, 2021; pp. 274–290.
51. Redlich, A.N. Redundancy Reduction as a Strategy for Unsupervised Learning. *Neural Comput.* **1993**, *5*, 289–304. [CrossRef]
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Condens. Matter Phys.* **2017**, *30*, 6000–6010.
53. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
54. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN Architectures for Large-Scale Audio Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
55. You, Y.; Gitman, I.; Ginsburg, B. Large batch training of convolutional networks. *arXiv* **2017**, arXiv:1708.03888.
56. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
57. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
58. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn Res.* **2008**, *9*, 2579–2605.