mathematics

MDPI

Article

Temporal–Semantic Aligning and Reasoning Transformer for Audio-Visual Zero-Shot Learning

Kaiwen Zhang [1], Kunchen Zhao [1] and Yunong Tian [2,*]

[1] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; zkw1014@bjfu.edu.cn (K.Z.); zkc15155885690@bjfu.edu.cn (K.Z.)
[2] CAS Engineering Laboratory for Intelligent Industrial Vision Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: yunong.tian@ia.ac.cn

Abstract: Zero-shot learning (ZSL) enables models to recognize categories not encountered during training, which is crucial for categories with limited data. Existing methods overlook efficient temporal modeling in multimodal data. This paper proposes a Temporal–Semantic Aligning and Reasoning Transformer (TSART) for spatio-temporal modeling. TSART uses the pre-trained SeLaVi network to extract audio and visual features and explores the semantic information of these modalities through audio and visual encoders. It incorporates a temporal information reasoning module to enhance the capture of temporal features in audio, and a cross-modal reasoning module to effectively integrate audio and visual information, establishing a robust joint embedding representation. Our experimental results validate the effectiveness of this approach, demonstrating outstanding Generalized Zero-Shot Learning (GZSL) performance on the UCF101 Generalized Zero-Shot Learning (UCF-GZSL), VGGSound-GZSL, and ActivityNet-GZSL datasets, with notable improvements in the Harmonic Mean (HM) evaluation. These results indicate that TSART has great potential in handling complex spatio-temporal information and multimodal fusion.

Keywords: audio-visual zero-shot learning; transformer

MSC: 68T07

Citation: Zhang, K.; Zhao, K.; Tian, Y. Temporal–Semantic Aligning and Reasoning Transformer for Audio-Visual Zero-Shot Learning. *Mathematics* **2024**, *12*, 2200. https://doi.org/10.3390/math12142200

Academic Editors: Hamad Naeem, Hong Su, Amjad Alsirhani and Muhammad Shoaib Bhutta

Received: 20 May 2024
Revised: 29 June 2024
Accepted: 10 July 2024
Published: 13 July 2024

1. Introduction

With the growth of social media, audio-visual content has proliferated across various domains, making zero-shot learning for audio-visual tasks a crucial research direction in the field of deep learning. Utilizing lots of labeled data to train models is a common approach in traditional supervised learning. However, acquiring such annotated data is often costly and complex. In real-world applications, we frequently encounter scenarios where new classes emerge during the identification and classification training phases. Therefore, zero-shot learning for audio-visual tasks becomes crucial for handling large-scale and diverse multimodal information. From a theoretical standpoint, audio-visual zero-shot learning has propelled the development of multimodal data fusion and zero-shot learning fields, offering a richer understanding for machine learning. In practical applications, it aids in constructing more intelligent audio-visual processing systems capable of addressing challenges in new contexts and with new objects, such as applications in security surveillance systems, intelligent audio-visual search, and multimodal content understanding. Existing methods for audio-visual zero-shot learning introduce additional complexity, such as preprocessing audio-visual data, leading to increased computational demands. However, these approaches have not effectively addressed the challenge of efficiently modeling temporal information. To tackle this issue, previous methods can be broadly categorized into four types: feature learning [1–3], capturing motion information [4,5], decoupling scene and motion information [6,7], and multimodal data fusion [8,9]. The approach of

utilizing feature learning enhances classification and recognition capabilities by learning discriminative features. In the study by Huang et al. [1], two pre-text tasks are designed to supervise context and motion information separately. MCL [2] proposes leveraging optical flow information for temporal and spatial sampling of video blocks, and enhancing the representation of motion information in feature learning through aligning gradient and optical flow maps. To focus more on foreground features, Modist [3] proposes a method that pays closer attention to learning objectives and cross-modal learning objectives. Feature learning addresses how to enhance classification and recognition capabilities, yet it still faces challenges in dealing with uncertainties in motion information. Therefore, we propose a method for capturing motion information. MDFT [4] constructs a spiking neural network to capture motion information, resolving the challenge of capturing motion information by focusing on contextual semantic information and dynamic motion information.

However, in this approach, the dependence on background scenes is overlooked. Building on this, the study by Bhat et al. [6] references an end-to-end visual tracking architecture, predicting the appearance information of both the target and the background. Similarly, the study by Wang et al. [7] proposes a method to decouple scene and motion information, coupling them through positive and negative clips of videos, enhancing the network's sensitivity to time to reducing bias towards background scenes in video learning. Nevertheless, this method overlooks scenarios with multiple different modalities, for which a common strategy is multimodal data fusion. For instance, AVMST [9] mentions the use of a transformer module to fuse audio and visual data, and AVCA [8] employs a spiking neural network module to extract salient temporal information, a cross-attention module to fuse time and semantic information, and a transformer inference module to explore the relationships between fused features for multimodal learning. Despite these methods addressing a majority of the challenges in deep learning, efficiently modeling temporal information remains a critical challenge yet to be resolved. In this paper, we focus on how to efficiently model temporal information to further enhance the performance and robustness of video analysis.

Based on the action captured in the images of a person shooting, it is evident that objects undergo significant changes over time, illustrating that the extracted features exhibit temporal dependencies. For instance, as depicted in Figure 1, the position and bending degree of the arm and bow vary each second. In conjunction with this, Figure 2 shows that the intensity of the sound signal, corresponding to the video in Figure 1, varies over time across different frequencies. This variation in sound, generated by the release of the bowstring and the flight of the arrow, underscores the significant temporal changes in the audio signal strength. To capture these nuances, the spectrogram of the audio signal is computed with specific parameters: a window length of 1024, which balances the time and frequency resolution; a Hann window, a widely used type that helps reduce spectral leakage and enhances the smoothness of the spectrum; and an FFT size of 2048, which determines the frequency resolution of the spectrum and is typically set to the nearest power of two greater than or equal to the window length. Additionally, to enhance the visualization of the spectrogram and facilitate the representation of a wide dynamic range of intensity values, a logarithmic transformation $Sxx_{log} = -np.log10(Sxx + 1 \times 10^{-10})$ is applied. Understanding these parameters is crucial for the model to predict and comprehend changes at different time points for unseen categories. By considering temporal information, the model can gain a deeper understanding of the context and background, enabling more accurate inference on unknown categories. Therefore, efficient modeling of temporal information plays a crucial role in audio-visual zero-shot learning. This approach ensures that the model not only captures static features but also appreciates the dynamic aspects of the scenes it analyzes, leading to more robust and context-aware predictions.

This paper focuses on efficiently modeling temporal information for audio-visual zero-shot learning. We propose a novel model comprising four key components: audio encoder, visual encoder, temporal information reasoning, and cross-modal reasoning module.

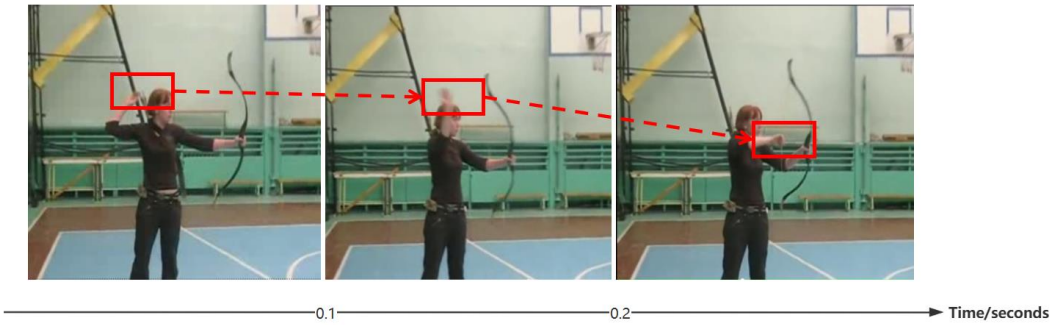

Figure 1. The changes in the position of the hand during shooting.

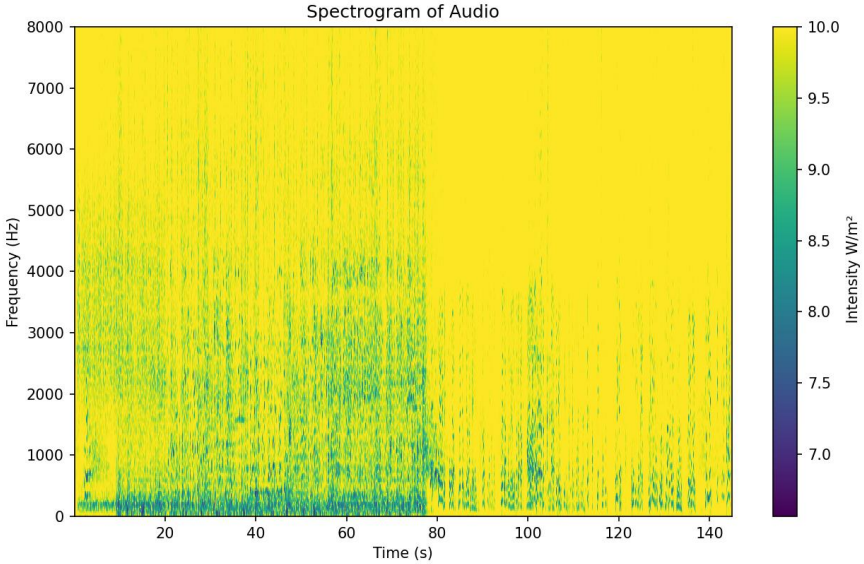

Figure 2. The intensity variation of sound signals at different frequencies in the audio.

Specifically, the audio encoder module is constructed based on a pre-trained feature extraction network. Its purpose is to extract rich audio features from the raw audio signal. Through a sequence of linear layers, batch normalization, ReLU activation functions, and dropout operations, the module further processes and refines audio data, uncovering deep semantic information within the audio signal.

Similar to the audio encoder, the visual encoder is established based on a pre-trained model, aimed at extracting features from visual data. The temporal information reasoning module is specifically designed to handle and leverage temporal information. It employs a multi-layer perceptron (MLP) to enhance the extraction capability of key temporal features within the audio modality. By incorporating layer normalization and residual connections, this module enables the model to comprehend and process dynamic audio-visual data that change over time.

The cross-modal reasoning module serves as the core of the model, combining audio and visual information to create a joint multimodal feature representation. Leveraging a cross-attention mechanism, this module fuses and strengthens semantic relationships between different modalities, enhancing the model's performance in integrating temporal and semantic features.

In summary, our paper aims to achieve the following primary objectives:

1. Introduce the temporal information reasoning module to efficiently model multimodal temporal information. By using an MLP, it strengthens the extraction of key temporal features within the audio modality, enhancing the efficiency of zero-shot learning.
2. The cross-modal reasoning module, employing a cross-attention mechanism, not only integrates information from different modalities but also reinforces semantic connec-

tions between them. This significantly boosts the model's capability in combining temporal and semantic features.

3. Extensive experiments validate the advanced efficiency of our method in modeling temporal information. Various ablation studies further prove the crucial value of each module in our approach.

The remainder of this paper is organized as follows. Section 2 provides a review of related work. Section 3 details the proposed methodology, including the design and functionality of the Temporal–Semantic Aligning and Reasoning Transformer (TSART). Section 4 presents the experimental setup and results, followed by a discussion of the findings. Section 5 concludes the paper and suggests directions for future research.

2. Related Work

With the progress of deep learning, there have been emerging methods in recent years dedicated to constructing a joint embedding space to effectively capture the correlation between audio and visual features, especially in the field of audio-visual zero-shot learning.

2.1. Multiple Learning

Multimodal learning has made significant strides in various fields, such as audio-visual learning. In the separation and localization of sound within videos [4,10–16], the study by Afouras et al. [17] offers a novel perspective and approach. Early on, it employed CNN architectures for detection and, through self-supervised learning, enhanced the accuracy and efficiency of object detection and sound localization, thereby furnishing valuable references for subsequent related research.

In the context of audio–video synchronization, the emphasis is on addressing the synchronization challenges, especially in uncontrolled ('wild') environments [18–22]. This includes investigating effective ways to synchronize audio and visual signals in settings with challenges such as noisy backgrounds, varying audio quality, and visual obstructions. In the realm of speech recognition and spoken-key-word-spotting [23–25], the emergence of attention mechanisms has led to the introduction of an attention-based visual keyword detection method [23]. This approach utilizes attention models to enhance the recognition accuracy of oral gestures in videos, effectively detecting key words in spoken language. By incorporating attention mechanisms and an end-to-end learning framework, this method brings a new research perspective and technological advancement to the field of visual keyword detection.

In the domain of utilizing visual information for audio synthesis [26–32], an innovative deep learning model has been introduced [26,33–35]. This model has the capability to comprehend and predict corresponding music or audio based on the observed video scenes. Unlike these methods, our model places a greater emphasis on aligning multimodal information, such as audio and visual. Closely associated with aligning multimodal information is the task of image-text retrieval. In previous research, the proposed Lightweight Transformer Alignment Network (LTAN) [32] enhanced the performance and efficiency of image-text retrieval by integrating lightweight transformers and an augmented pathway. Moreover, NSTRN is built upon a spiking neural network [36], effectively addressing the challenges of image-text retrieval in wireless communication environments. It achieves efficient binary encoding and transmission of features, reducing bandwidth requirements and enhancing retrieval accuracy. CKSTN [37] has introduced an efficient image-information retrieval network, leveraging shared knowledge and style embeddings to enhance cross-modal matching performance. Additionally, a novel dual-stream network, MDFT, is proposed to decouple contextual semantics and dynamic motion information, improving the accuracy of video classification in zero-shot learning. This approach utilizes a spiking neural network to handle sparse data and surpasses existing technologies on standard benchmarks. Inspired by the aforementioned papers, we employ an attention mechanism to integrate multimodal features and achieve efficient alignment.

2.2. Zero-Shot Learning for Audio-Visual Tasks

Many image-based generative zero-shot learning methods have been suggested now. For instance, the study by Narayan et al. [38] introduces an approach that enhances the classification performance of zero-shot learning by combining latent embedding feed-back and discriminative feature representation. However, a limitation lies in the requirement of prior knowledge about unseen classes, and the model's generalization capability may still be constrained. In contrast, non-generative methods focus on learning the mapping from input features to class semantics, such as text category label embeddings, without the need for prior knowledge. Video-based zero-shot learning has also been extensively explored. For instance, the AVCA model [8] enhances the efficiency of the model framework by using pre-extracted audio and visual features as inputs. Differing from these methods, we focus on efficiently modeling temporal information, enhancing the feature extraction process, and improving the model's generalization capability. This allows for better performance on unseen classes or tasks.

In the realm of audio-visual zero-shot learning, a notable research contribution is the proposed MDFT architecture [4]. By decoupling semantic information and dynamic motion information, it significantly improves classification accuracy, particularly in biased background scenarios. A method employing cross-modal attention was introduced [8], utilizing text label embeddings for knowledge transfer. In the context of generalized audio-visual zero-shot learning, the introduction of training classes as distractors increases the difficulty while ensuring unseen testing classes are not part of supervised training data. The AVCA model has demonstrated excellent performance in generalized zero-shot learning tasks on the VGGSound, UCF, and ActivityNet datasets. As a novel self-supervised video representation learning approach, MoDist [3] focuses on explicitly extracting motion information. In contrast to previous methods that implicitly learn motion cues in RGB inputs, MoDist's learned representation pays more attention to foreground motion regions, exhibiting superiority in action recognition and detection tasks. Diverging from previous approaches, our emphasis lies in efficiently modeling temporal aspects, enhancing cross-modal alignment, aiding the model in better understanding the correlation be-tween audio and video. This facilitates improved comprehension of temporal data for new classes.

3. Methodology

The audio, visual, and embedding features for real labels are, respectively, represented as x_i^a, x_i^v, and x_i^t. During the model training phase, the training set for the seen class i can be expressed as the set X = (x_i^a,x_i^v,x_i^t), while the invisible class is represented as Y = (y_i^a,y_i^v,y_i^t). The Temporal–Semantic Aligning and Reasoning Transformer proposes learning a mapping function within the seen classes, denoted as $f(y_i^a, y_i^v) \rightarrow y^g$, where y_i^g represents the class-level text embedding for class j. Similarly, the test set for unseen classes Y can also be mapped as $f(y_i^a, y_i^v) \rightarrow y_i^g$. The framework of the Temporal–Semantic Aligning and Reasoning Transformer is illustrated in the diagram. As shown in Figure 3, the Temporal–Semantic Aligning and Reasoning Transformer (TSART) architecture seamlessly integrates audio, visual, and textual features, represented by blue, green, and orange lines, respectively. The system initially inputs raw audio and visual data. The audio encoder and visual encoder then extract distinctive features from their respective modalities, utilizing the pre-trained SeLaVi network. These features are further refined by feature extractors. A cross-modal reasoning module fuses and infers the interrelations between features of different modalities, enhancing the understanding of their mutual connections. The temporal information reasoning module and temporal transformer encoder analyze time-series data, capturing the dynamics of temporal changes. A projection layer maps the features into dimensions suitable for model processing. The model is ultimately applied to test tasks, such as the reconstruction of bowling movements, thereby validating its performance. This architecture effectively models temporal information in audio-visual content, demonstrating its capability to handle complex, multimodal data within a unified framework.

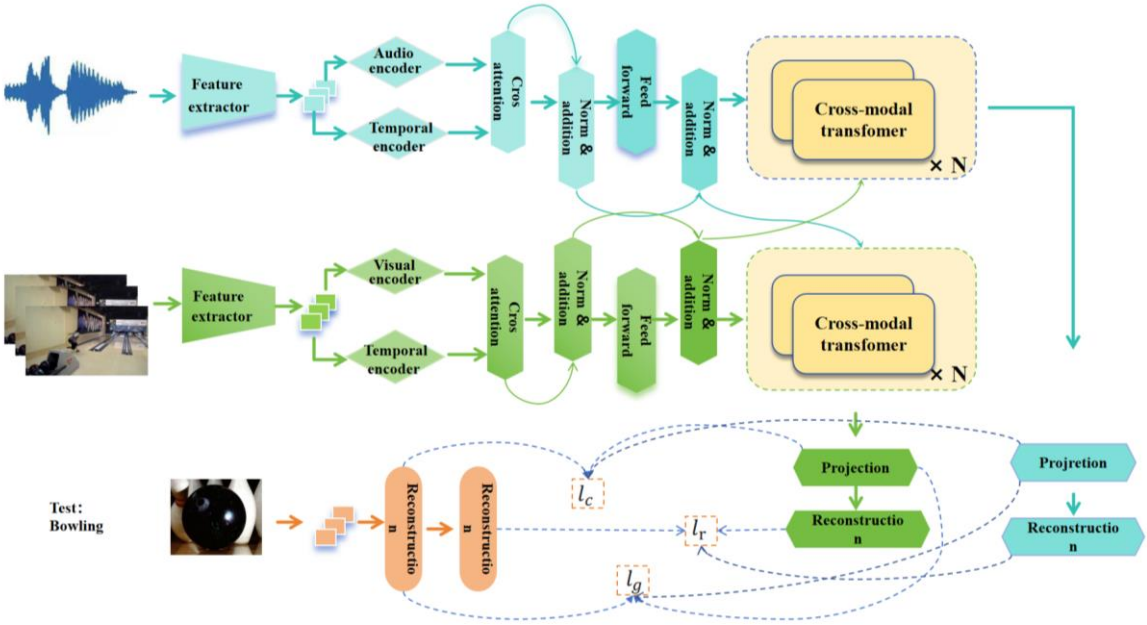

Figure 3. The Temporal–Semantic Aligning and Reasoning Transformer architecture includes an audio encoder, visual encoder, temporal information reasoning, and cross-modal reasoning module. It effectively models temporal information in audio-visual content.

3.1. Semantic–Temporal Relationship Reasoning

The robust and discriminative audio and visual features are extracted using pre-trained SeLaVi [39]. The distinctive feature of SeLaVi lies in its self-supervised learning approach, enabling it to learn from a substantial amount of unlabeled data. This capability allows SeLaVi to delve deeply into understanding the inherent patterns and subtle differences present in multimedia content. By leveraging SeLaVi's pre-trained network, we can make use of its extensive feature representation capabilities without the need to invest significant computational resources in training such a complex model from scratch. Moreover, SeLaVi's self-supervised characteristics ensure that the extracted features not only exhibit robustness to variations in the data but also possess sufficient distinctiveness. This distinctiveness enables the identification of subtle differences in audio and visual signals.

3.2. Audio Encoder

After feature extraction, to further explore the semantic information of the audio modality, an audio encoder, denoted as A_{enc}, is proposed. The output of this encoder can be represented as $a^t = A_{enc}(x_a)$, where $a^t \in R^{kf}$, obtained through a pre-trained feature extractor. The audio encoder comprises two linear layers, f_1^a and f_2^a, specifically, f_1^a:$R^{kinput} \rightarrow R^{khidd}$ and f_2^a:$R^{khidd} \rightarrow R^{kf}$. After each linear layer, batch normalization, ReLU activation functions, and dropout with a rate of d_{enc} are applied. Through these processes, A_{enc} can effectively delve deeper into the semantic information of the audio data.

3.3. Visual Encoder

Similarly, a visual encoder, denoted as V_{enc}, is introduced to explore the semantic information of the visual modality. The output of the visual encoder can be expressed as $v^t = V_{enc}(x_v)$, where $v^t \in R^{kf}$, obtained through a pre-trained feature extractor. The visual encoder also consists of two linear layers, f_1^v and f_2^v, specifically, f_1^v:$R^{kinput} \rightarrow R^{khidd}$ and f_2^v:$R^{khidd} \rightarrow R^{kf}$. Like the audio encoder, after each linear layer, batch normalization, ReLU activation functions, and dropout with a rate of d_{enc} are applied. Through these processes, V_{enc} is capable of further extracting the semantic information from the visual data.

3.4. Temporal Information Reasoning

To address the effective capture and utilization of temporal information in zero-shot learning for audio-visual content, the growth of audio-visual content across various domains, driven by the development of social media, has become widespread. This surge has resulted in an increased demand for systems capable of handling unseen categories in audio-visual processing. These systems require not only the recognition of static features but also an understanding of how objects evolve and change over time. Efficient modeling of temporal information is crucial for comprehending the dynamics of entities and contributes to improving the accuracy and generalization capability of models in the recognition of new contexts and objects. As social media continues to evolve, the ability to handle the temporal dimension becomes increasingly vital in audio-visual zero-shot learning systems.

In the context of multimodal temporal information modeling, the processing flow $P_a = \mathrm{MLP}(\mathrm{LN}(R_a)) + R_a$ emphasizes the fine-grained extraction and enhancement of temporal features in the audio modality. In this module, we refine and strengthen the time-dependent audio features R_a by incorporating a self-attention mechanism.

Firstly, R_a, representing the time-dependent features of the audio modality, is fed into Layer Normalization (LN). This step serves not only to mitigate potential internal covariate shift during training but also, through normalization, ensures consistent scaling of features extracted at different time points within the model. This promotes stability and comparability in time-series data, contributing to the overall stability and interpretability of the model.

Subsequently, the temporally adjusted feature $\mathrm{LN}(R_a)$ is fed into a multi-layer perceptron (MLP). The MLP, with its multi-layer structure, not only captures complex nonlinear relationships within the audio signals, but also, due to the selective attention capability of each layer's weight allocation, enhances self-attention to critical features in the temporal sequence. This enables a better extraction of patterns and trends within the temporal information.

Ultimately, the output of the MLP is combined with the original temporal features R_a, forming a composite representation pa that includes both the original temporal information and the features enhanced through the deep neural network. This design of residual connection not only provides a self-correction mechanism for features, preserving the integrity of the original temporal characteristics, but also introduces deep and rich temporal dynamics extracted by the MLP through its self-attention learning process. Consequently, the final output P_a offers a comprehensive understanding of the audio features, which is crucial for processing and interpreting multimodal temporal information. The formula can be represented as follows:

$$P_a = \mathrm{MLP}(\mathrm{LN}(R_a)) + R_a \tag{1}$$

$$P_v = \mathrm{MLP}(\mathrm{LN}(R_v)) + R_v \tag{2}$$

In the framework of multimodal learning, this formula not only reflects the intuitive characteristics of audio signals but also integrates rich contextual information along the temporal dimension, providing downstream tasks with a more profound and comprehensive feature representation. Through this design, the model can effectively capture and utilize crucial information when faced with multimodal temporal data, thereby enhancing the system's perception and responsiveness to temporal changes.

3.5. Cross-Modal Reasoning Module

The cross-modal reasoning module (CRM) aims to integrate temporal and semantic features from different modalities efficiently, creating a unified feature representation for both audio and visual information. This is achieved by adding a residual connection between two layers and applying layer normalization to enhance complementarity and stability between features.

In the cross-modal reasoning module (CRM), the audio attention fusion block utilizes a cross-attention mechanism (CA) to integrate the audio feature a^t with the information from the previous time step, processed through a sigmoid activation function, forming a new audio feature representation $\mathrm{R_a}$. To further refine these fused features, $\mathrm{R_a}$ is passed through a multi-layer perceptron (MLP). This MLP initially undergoes Layer Normalization (LN) to eliminate potential scale differences between different layers. The features processed by the MLP are combined with the original audio features $\mathrm{R_a}$ to form the final audio feature output P_a. This step is accomplished through a residual connection, which not only preserves the integrity of the original features but also incorporates the depth-learning information from the MLP layer. In this way, the CRM captures subtle changes in the temporal dimension while retaining the original audio data features, resulting in a comprehensive audio feature representation P_a.

In summary, the processing flow and formulas pertaining to audio and video in the CRM module can be, respectively, represented as follows:

$$\mathrm{R_a} = \mathrm{CA}(a^t, \mathrm{Sigmoid}(a^{t-1})) \tag{3}$$

$$\mathrm{P_a} = \mathrm{MLP}(\mathrm{LN}(\mathrm{R_a})) + \mathrm{R_a} \tag{4}$$

$$\mathrm{R_v} = \mathrm{CA}(v^t, \mathrm{Sigmoid}(v^{t-1})) \tag{5}$$

$$P_v = \mathrm{MLP}(\mathrm{LN}(\mathrm{R_v})) + \mathrm{R_v} \tag{6}$$

This design ensures that CRM can effectively capture and leverage crucial information from cross-modal temporal data, providing robust support for modeling multimodal temporal information. Additionally, a standard transformer layer is introduced in this module, described as follows:

$$W_{av} = \mathrm{HMCA}(P_a, P_v) \tag{7}$$

$$F_{av} = \mathrm{MLP}(\mathrm{LN}(z_i)) + z_{av} \tag{8}$$

Our objective is to predict the category of text labels. To project joint audio-visual embeddings into the space of text label embeddings, we employ the reconstruction and projection technique. This involves utilizing reconstruction and projection features to restore the initial information, achieving a comparable effect on features across different modalities. The projections consist of two linear layers, namely f_3^a and f_4^a for audio and f_3^v and f_4^v for video. After each layer, normalization, ReLU activation function, and dropout with a ratio of d_{proj} are applied. Ultimately, the embedded audio and video features can be expressed as:

$$Pr_a = A_{proj}(F_a) \tag{9}$$

$$Pr_v = A_{proj}(F_a) \tag{10}$$

where A_{proj} represents the projection function, and the embeddings of text labels Pr_a and Pr_v are obtained by projecting the embedding of the k-th class label W_k through the projection layer $W_{project}$. The structure of $W_{project}$ is similar to that of Pr_a and Pr_v, with the distinction lying in the dropout ratio $d_{wproject}$ compared to them.

3.6. Training Strategy

During the training process of the TSART model, to expedite the convergence of our model, we employed a composite loss function $\mathcal{L}_{fal}$ for updates, consisting of three components: triplet loss $\mathcal{L}_t$, projection loss $\mathcal{L}_{pro}$, and reconstruction loss $\mathcal{L}_{re}$. The amalgamation of these triplet losses allows us to more accurately cluster the final audio and visual embeddings, thereby enhancing the coherence and reliability of the model output.

The design of the joint triplet loss $\mathcal{L}_t$ aims to tightly cluster the final audio-visual embeddings for more coherent results. This loss function achieves this objective by ensuring that negative samples between different modalities have a minimal margin with the audio-

visual embeddings that correspond to true matches. This minimum margin is defined by the margin parameter Z, and the formula is as follows:

$$\mathcal{L}_t = \left[\mathrm{Z} + Pr_a^+ + Pr_v^+ - \left(Pr_a^- + Pr_v^-\right)\right] + \left[\mathrm{Z} + Pr_a^- + Pr_v^- - \left(Pr_a^+ + Pr_v^+\right)\right] \quad (11)$$

Pr_a^+, Pr_v^+, Pr_a^-, and Pr_v^- represent positive and negative samples. The projection loss $\mathcal{L}_{pro}$ aims to reduce the distance between the joint embeddings from the projection layer output and the corresponding text label embeddings. This optimization helps the model generate embedded representations that are closer to the text labels.

$$\mathcal{L}_{pro} = \sum_{\mathrm{x}=1}^{\mathrm{i}} \frac{Pr_a + Pr_v^- - Pr_w}{\mathrm{x}} \quad (12)$$

The reconstruction loss $\mathcal{L}_{re}$ is introduced to preserve the original data distribution while projecting audio-visual features into a shared embedding space. This helps the model learn the ability to reconstruct the original data during the reconstruction phase. The overall composite loss function is:

$$\mathcal{L}_{fal} = \mathcal{L}_t + \mathcal{L}_{pro} + \mathcal{L}_{re} \quad (13)$$

By combining these three losses to guide the model training, we observed that when applying our designed composite loss function, the model demonstrated outstanding Generalized Zero-Shot Learning (GZSL) performance on the UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL datasets. Moreover, it significantly outperformed in the Harmonic Mean (HM) evaluation.

4. Experiment

In our research, we conducted a comprehensive evaluation of the proposed model, specifically focusing on zero-shot learning (ZSL) and generalized zero-shot learning (GZSL). We carried out extensive experiments to ensure that our model assessments were both thorough and comprehensive.

We trained the TSART model on a single NVIDIA GeForce RTX 4080 Laptop GPU, strictly following the procedures described in reference [8] to extract audio and visual embeddings per second. During this process, we set *kinput* to 512, *khidd* to 512, and *kf* to 64. For the UCF, VGG-Sound, and ActivityNet datasets, dropout rates were set as d_{enc} = **0.5, 0.2, 0.3**, d_{dec} = **0, 0.1, 0**, and d_{wproj} = **0.1, 0.2, 0.2**, respectively. In our cross-modal transformer architecture, we designed a structure with eight attention heads, each having a dimension of 64. The training process utilized the Adam optimizer, and the entire TSART model underwent 50 training epochs, with a learning rate set to 0.001. This rigorous training regimen ensured that the model learned robust audio and visual embeddings, facilitating its performance across diverse datasets and tasks.

In evaluating ZSL, we not only focused on the model's overall recognition ability for the visible categories but also conducted a detailed analysis of its performance across various unseen categories. This approach ensured that the model exhibited balanced generalization across different dimensions. In addition, to evaluate the model's capability in capturing subtle differences among unseen categories, we introduced more complex classification scenarios designed to simulate real-world category distribution and variability.

In the evaluation of GZSL, our approach not only addresses the model's recognition capability for unseen categories but also assesses its performance when both seen and unseen categories coexist. This implies that our evaluation was not conducted in a simplified environment but rather in a more complex setting that closely resembles real-world applications. In this environment, the model is required to make precise discriminations between seen and unseen categories. We employed the Harmonic Mean (HM) as the primary evaluation metric. The Harmonic Mean is a balanced measure of both seen and unseen category performance, calculated using the formula HM = 2US/(U + S), where (U) represents the accuracy of unseen categories, and (S) represents the accuracy of seen categories.

This evaluation framework ensures that we can fully understand the model's performance, especially in complex scenarios that may occur in the real world. By evaluating the model's ability to recognize both seen and unseen categories, we can better understand and improve the model's generalization and practicality for future applications.

4.1. Data Statistic

The UCF101 dataset stands as a cornerstone in action recognition, housing 13,320 video segments spread across 101 action categories. These categories cover a broad spectrum of human activities, ranging from sports and musical performances to various physical movements. The dataset is meticulously organized into 25 groups, each containing videos from four to seven action categories. This structured layout of UCF101 serves as a pivotal benchmark for evaluating the performance of video understanding models, particularly in the realm of action recognition. Derived from UCF101, the UCF101 Generalized Zero-Shot Learning (UCF-GZSL) dataset is utilized to delve into video action recognition within the framework of Generalized Zero-Shot Learning.

ActivityNet is another indispensable dataset in video understanding, with a dedicated focus on human activity recognition. Comprising approximately 27,801 video segments categorized into 200 different activity classes, ActivityNet aims to capture a wide array of complex activities, from sports to daily actions. It provides a comprehensive benchmark for evaluating video understanding and activity recognition algorithms. The diverse activity types and temporal dynamics present in ActivityNet pose a robust challenge for models aimed at interpreting and predicting human behavior from video data. ActivityNet-GZSL, a derivative of ActivityNet, is employed to explore activity recognition within the context of Generalized Zero-Shot Learning.

The VGG Sound Generalized Zero-Shot Learning (VGG Sound-GZSL) dataset is built upon the VGG Sound dataset and is utilized for studying sound event recognition within the Generalized Zero-Shot Learning paradigm. Additionally, the VGGSound dataset, encompassing 212,894 video segments categorized into 309 distinct audio classes, serves as a crucial resource in audio event detection and classification research. It covers various acoustic events such as instrument sounds, human speech, animal sounds, and environmental noises. VGGSound is invaluable for tasks requiring an understanding of audio backgrounds in video segments. These three datasets, UCF101, ActivityNet, and VGGSound, play pivotal roles as benchmarks in audio-visual analysis. They significantly contribute to the development and evaluation of models designed to comprehend and classify a wide range of human activities and sounds.

4.2. Comparison with State-of-the-Art

To demonstrate the effectiveness of our model, we have compared it with the state-of-the-art audio-visual generalized zero-shot learning (GZSL) methods on three major benchmark datasets. In this section, we will delve into a detailed exploration and discussion of the differences between various methods and our model. SJE [39] introduced a novel learning strategy that effectively distinguishes matching embeddings from non-matching embeddings by learning a compatibility function between image and category embeddings. This strategy assigns higher scores to matching embeddings.

This approach enables the model to more accurately identify and classify visual objects, especially in scenarios with multiple categories. Apn [40] introduced an innovative zero-shot representation learning framework that not only co-learns global and local features but also leverages class-level attributes to enhance attribute-based knowledge transfer.

This approach allows the model to recognize categories it has not directly encountered, which is particularly useful in practical applications. VAEGAN [41] employs a conditional generative model that combines the advantages of variational autoencoders and generative adversarial networks. This enables the model to better understand and generate images that have not been seen during the training process, especially those with marginal feature distributions.

In contrast to the aforementioned methods, our work emphasizes the efficient modeling of multimodal temporal information [42]. By integrating time-series modeling techniques, our model can not only handle static image data but also comprehend and analyze the temporal dynamics in videos, as well as temporal patterns in audio signals.

This allows our model to capture richer contextual information, providing deeper insights into understanding complex audio-visual scenes. In this way, our model exhibits significant improvements in generalized zero-shot learning tasks, particularly when handling unlabeled and diverse real-time data streams, offering more accurate and robust recognition capabilities.

Results comparison: By comparing with recent mainstream methods in Table 1, we demonstrate the outstanding performance of the temporal–semantic aligning and reasoning transformer. In the case of the VGGSound-GZSL dataset, our model achieves a score of 10.45 in the seen domain (S), which is lower than SJE's 48.33. However, in the unseen domain (U), our model outperforms SJE with a score of 3.43 compared to SJE's 1.10. Moreover, in the harmonic mean (HM), our model achieves a score of 5.16, indicating a 3.01 improvement over SJE's result. This suggests that our framework exhibits a stronger ability to transfer knowledge from seen to unseen classes. However, due to the improvement potential in our external noise handling module for input audio, our framework does not outperform SJE in ZSL results for this dataset. On the ActivityNet-GZSL dataset, while SJME focuses more on handling highly sparse event data, our model emphasizes efficient modeling of temporal information in multimodal data. As a result, we achieve a harmonic mean (HM) of 8.12 and a Zero-Shot Learning (ZSL) score of 7.65, surpassing CJME's harmonic mean of 5.12 and ZLS of 5.84. On the UCF-GZSL dataset, our model achieves superior performance, with an HM score of 21.11, surpassing APN's 18.05. In terms of ZSL, our model outperforms AVGZSLNet with a score of 22.86 compared to 13.65, exhibiting a significant improvement of 9.21.

Table 1. We assessed the performance of our Temporal–Semantic Aligning and Reasoning Transformer alongside state-of-the-art baseline methods in audio-visual (G)ZSL across three benchmark datasets.

Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S	**U**	**HM**	**ZSL**	**S**	**U**	**HM**	**ZSL**	**S**	**U**	**HM**	**ZSL**
SJE [39]	48.33	1.10	2.15	4.06	63.10	16.77	26.50	18.93	4.61	7.04	5.57	7.08
APN [40]	7.48	3.88	5.11	4.49	28.46	16.16	20.61	16.44	9.84	5.76	7.27	6.34
VAEGAN [41]	12.77	0.95	1.77	1.91	17.29	8.47	11.37	11.11	4.36	2.14	2.87	2.40
CJME [43]	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 [44]	18.05	3.48	5.83	5.28	52.52	10.90	18.05	13.65	8.93	5.04	6.44	5.40
TSART	10.45	3.43	5.16	4.03	20.96	21.27	21.11	**22.86**	8.99	7.41	**8.12**	**7.65**

The bold numbers in Table 1 indicate where the TSART model significantly outperforms other models.

Our primary evaluation metrics are the harmonic mean (HM) and Zero-Shot Learning (ZSL). Overall, these results confirm that our model has stronger domain adaptation and knowledge transfer capabilities.

4.3. Ablation Study

In the ablation study conducted on the UCF-GZSL task, as shown in Table 2, different dropout configurations had a significant impact on the model's performance metrics, including seen classes (S), unseen classes (U), harmonic mean (HM), and Zero-Shot Learning (ZSL). Notably, setting dropout rates for the reconstruction (rdec), encoding (renc), and projection (rproj) parts to 0.2, 0.3, and 0.5, respectively, resulted in the model achieving the highest performance on unseen classes (U) with a score of 18.57, indicating a strong domain adaptation ability. However, this configuration did not exhibit the best performance on harmonic mean (HM) and Zero-Shot Learning (ZSL), suggesting a tendency to overfit seen classes. On the other hand, configuring drop-out rates for rdec, renc, and rproj as 0.5, 0.3,

and 0.2, respectively, resulted in the model showing optimal performance on seen classes (S) and harmonic mean (HM), with scores of 24.93 and 19.89, demonstrating the model's efficient domain adaptation and knowledge transfer capabilities.

Table 2. Ablation study on the advantage of dropout in processing event information.

Dropout			UCF-GZSL			
rdec	**renc**	**rproj**	**S**	**U**	**HM**	**ZSL**
0.5	0.3	0.2	24.93	16.54	19.89	21.61
0.2	0.3	0.5	17.4	18.57	17.97	19.36
0.3	0.2	0.5	8.66	21.39	12.33	21.56
0.5	0.2	0.3	20.96	21.27	21.11	22.86

This analysis highlights that finely adjusting dropout rates can effectively control the model's recognition abilities for both seen and unseen classes, optimizing the model's adaptability to different data distributions while preserving its generalization ability.

The effectiveness of temporal–semantic aligning and reasoning transformer components. In the evaluation of UCF-GZSL, as shown in Table 3, different variants of the model exhibit varying degrees of accuracy differences. Specifically, when the audio encoder is removed (W/O A_{TEM}), the recognition accuracy for seen classes (S) is 6.73, and for unseen classes (U) it is 13.56, with a harmonic mean (HM) of 8.99 and a Zero-Shot Learning (ZSL) accuracy of 13.71. In contrast, when removing the visual encoder (W/O V_{TEM}), the model achieves S and U accuracies of 21.07 and 12.67, respectively, with an improved HM of 15.82, and a slightly increased ZSL accuracy of 13.77. Our complete model (OURS) outperforms in all metrics, particularly showing significant improvements in U and ZSL, reaching 21.27 and 22.86, while achieving an HM of 21.11. This indicates that the complete model possesses stronger capabilities in generalizing between seen and unseen categories. This result highlights the importance of the audio encoder and visual encoder for the overall performance of the model, especially in enhancing the model's ability to recognize unseen categories.

Table 3. Analysis of ablation study on UCF-GZSL dataset.

Model	**UCF-GZSL**			
	S	**U**	**HM**	**ZSL**
W/O A_{TEM}	6.37	13.56	8.99	13.71
W/O V_{TEM}	21.07	12.67	15.82	13.77
TSART	20.96	21.27	21.11	22.86

Setting the learning rate is crucial in training machine learning models because it determines the size of the step that the model takes while searching for the optimal solution [45,46]. If the learning rate is too small, the model may take a long time to converge to the optimal solution, or it may get stuck in a suboptimal solution [47,48]. On the other hand, if the learning rate is too large, the model may overshoot the optimal solution and fail to converge. Therefore, we evaluate the learning rate for ActivityNet-GZSL, as shown in Table 4. We observe that at a learning rate of 0.0005, the model achieves a ZSL score of 7.65, identical to that at a learning rate of 0.01. However, its HM score of 4.01 is significantly lower compared to the HM score of 8.12 achieved at a learning rate of 0.001. Similarly, at a learning rate of 0.01, both the HM and ZSL scores are much lower than those obtained at a learning rate of 0.001. Furthermore, at a learning rate of 0.001, the model achieves the best results on seen classes (S) and unseen classes (U), with values of 8.99 and 7.41, respectively, indicating its high domain adaptation and transfer capabilities in this setting.

Table 4. Ablation study on the impact of learning rate on processing event information.

Learning Rate	ActivityNet-GZSL			
	S	**U**	**HM**	**ZSL**
0.0005	5.14	3.29	4.01	7.65
0.01	4.28	4.20	4.24	4.50
0.001	8.99	7.41	**8.12**	**7.65**

The bold numbers in Table 4 represent the best results achieved on HM and ZSL.

Finally, we evaluated the impact of using different loss functions during the training of TSART on the performance of GZSL and ZSL, as shown in Table 5. Our analysis indicates that the complete loss function $\mathcal{L}_{fal}$ produces the strongest GZSL results (HM) on the UCF-GZSL, VGGSound-GZSL, and ActivityNet-GZSL datasets. This significant improvement demonstrates the crucial importance of our proposed complete loss function $\mathcal{L}_{fal}$ for achieving strong overall performance on all three datasets. On the VGGSound-GZSL dataset, the model omitting $\mathcal{L}_t$ slightly outperforms the complete loss function $\mathcal{L}_{fal}$ in ZSL, with 4.84 compared to 4.03. However, the complete loss function $\mathcal{L}_{fal}$ demonstrates superior performance in GZSL, with a Harmonic Mean (HM) of 5.16, outperforming other approaches. On the UCF-GZSL dataset, using the complete loss function $\mathcal{L}_{fal}$ significantly improves GZSL performance, with HM of 21.11 and ZSL of 22.86. On the ActivityNet-GZSL dataset, the ZSL performance of $\mathcal{L}_{fal} - \mathcal{L}_{pro}$ is superior to that of the complete loss function $\mathcal{L}_{fal}$, with 8.02; however, using the complete loss function $\mathcal{L}_{fal}$ yields better results in ZSL, with 8.12. Our ablation study confirms that when trained using our proposed complete loss function, strong overall performance can be achieved on all three datasets.

Table 5. Comparative analysis of TSART training using our full loss function $\mathcal{L}_{fal}$ versus removing individual components ($\mathcal{L}_t$, $\mathcal{L}_{pro}$, $\mathcal{L}_{re}$) on the GZSL and ZSL performance across the VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL datasets.

Model	VGGSound-GZSL		UCF-GZSL		ActivityNet-GZSL	
	HM	**ZSL**	**HM**	**ZSL**	**HM**	**ZSL**
$\mathcal{L}_{fal} - \mathcal{L}_t$	5.06	**4.84**	18.51	19.17	7.39	7.53
$\mathcal{L}_{fal} - \mathcal{L}_{pro}$	4.87	4.31	17.88	17.51	8.02	**8.02**
$\mathcal{L}_{fal} - \mathcal{L}_{re}$	4.24	4.43	21.03	17.01	7.14	7.94
$\mathcal{L}_{fal}$	**5.16**	4.03	**21.11**	**22.86**	**8.12**	7.65

The bold numbers in Table 5 are the best results on HM and ZSL among different models.

5. Conclusions

In general, we have explored how to model temporal information efficiently in the domain of zero-shot learning (ZSL) for audio-visual data.

Through a detailed analysis of the growing demand for audio-visual content in the evolution of social media, we are confident that effective temporal modeling is crucial for handling unseen categories.

To extract robust and discriminative audio and visual features, we employed the pre-trained self-supervised network SeLaVi. This network has the ability to learn from unlabeled data, allowing it to comprehend intrinsic patterns and subtle differences in multimedia content.

By adopting this approach, we circumvented the need for substantial computational resources required for training complex models from scratch. To efficiently capture and leverage temporal information, we devised a temporal information inference pipeline, emphasizing the nuanced extraction and enhancement of temporal features in the audio modality.

We utilized a self-attention mechanism to handle and analyze the temporally changing audio features Ra. Through layer normalization, a multi-layer perceptron (MLP), and resid-

ual connections, we created a composite representation for the audio signal, incorporating both the raw temporal information and enhanced features.

We utilized a cross-attention mechanism and a multi-layer perceptron (MLP) to reinforce the complementarity and stability between different modality features.

Finally, to project the joint audio-visual embeddings into the space of text label embeddings and restore the initial information, we designed reconstruction and projection functions. These functions consist of linear layers optimized through normalization, ReLU activation functions, and dropout. Our experiments thoroughly demonstrate the superiority of the model in efficiently modeling multimodal temporal information.

Author Contributions: Conceptualization, K.Z. (Kaiwen Zhang), K.Z. (Kunchen Zhao) and Y.T.; Methodology, K.Z. (Kaiwen Zhang) and Y.T.; Software, K.Z. (Kaiwen Zhang), K.Z. (Kunchen Zhao) and Y.T.; Validation, K.Z. (Kunchen Zhao) and Y.T.; Writing—original draft, K.Z. (Kaiwen Zhang) and K.Z. (Kunchen Zhao); Writing—review & editing, K.Z. (Kaiwen Zhang) and K.Z. (Kunchen Zhao); Supervision, Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 62206275.

Data Availability Statement: No new data were created or analyzed in this study.

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

References

1. Huang, L.; Liu, Y.; Wang, B.; Pan, P.; Xu, Y.; Jin, R. Self-supervised video representation learning by context and motion decoupling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13886–13895.
2. Li, R.; Zhang, Y.; Qiu, Z.; Yao, T.; Liu, D.; Mei, T. Motion-focused contrastive learning of video representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2105–2114.
3. Xiao, F.; Tighe, J.; Modolo, D. Modist: Motion distillation for self-supervised video representation learning. *arXiv* **2021**, arXiv:2106.09703.
4. Li, W.; Zhao, X.L.; Ma, Z.; Wang, X.; Fan, X.; Tian, Y. Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 3994–4002.
5. Torfi, A.; Iranmanesh, S.M.; Nasrabadi, N.; Dawson, J. 3D convolutional neural networks for cross audio-visual matching recognition. *IEEE Access* **2017**, *5*, 22081–22091. [CrossRef]
6. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6182–6191.
7. Wang, J.; Gao, Y.; Li, K.; Hu, J.; Jiang, X.; Guo, X.; Sun, X. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 10129–10137.
8. Mercea, O.B.; Riesch, L.; Koepke, A.; Akata, Z. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 10553–10563.
9. Li, W.; Ma, Z.; Deng, L.J.; Man, H.; Fan, X. Modality-Fusion Spiking Transformer Network for Audio-Visual Zero-Shot Learning. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 426–431.
10. Arandjelovic, R.; Zisserman, A. Objects that sound. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 435–451.
11. Gao, R.; Grauman, K. Co-separating sounds of visual objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3879–3888.
12. Owens, A.; Efros, A.A. Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 631–648.
13. Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; Lin, W. Multiple sound sources localization from coarse to fine. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XX 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 292–308.
14. Tian, Y.; Shi, J.; Li, B.; Duan, Z.; Xu, C. Audio-visual event localization in unconstrained videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 247–263.

15. Afouras, T.; Owens, A.; Chung, J.S.; Zisserman, A. Self-supervised learning of audio-visual objects from video. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVIII 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 208–224.
16. Tzinis, E.; Wisdom, S.; Jansen, A.; Hershey, S.; Remez, T.; Ellis, D.P.; Hershey, J.R. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv* **2020**, arXiv:2011.01143.
17. Triantafyllos, A.; Yuki, M.A.; Fagan, F.; Vedaldi, A.; Metze, F. Self-supervised object detection from audio-visual correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10575–10586.
18. Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Audio-visual synchronisation in the wild. *arXiv* **2021**, arXiv:2112.04432.
19. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In Proceedings of the Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part II 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2017.
20. Ebenezer, J.P.; Wu, Y.; Wei, H.; Sethuraman, S.; Liu, Z. Detection of audio-video synchronization errors via event detection. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4345–4349.
21. Khosravan, N.; Ardeshir, S.; Puri, R. On Attention Modules for Audio-Visual Synchronization. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019.
22. Akhtar, Z.; Falk, T.H. Audio-visual multimedia quality assessment: A comprehensive survey. *IEEE Access* **2017**, *5*, 21090–21117. [CrossRef]
23. Prajwal, K.R.; Momeni, L.; Afouras, T.; Zisserman, A. Visual keyword spotting with attention. *arXiv* **2021**, arXiv:2110.15957.
24. Momeni, L.; Afouras, T.; Stafylakis, T.; Albanie, S.; Zisserman, A. Seeing wake words: Audio-visual keyword spotting. *arXiv* **2020**, arXiv:2009.01225.
25. Rehman, A.U.; Ullah, H.S.; Farooq, H.; Khan, M.S.; Mahmood, T.; Khan, H.O.A. Multi-modal anomaly detection by using audio and visual cues. *IEEE Access* **2021**, *9*, 30587–30603. [CrossRef]
26. Gan, C.; Huang, D.; Chen, P.; Tenenbaum, J.B.; Torralba, A. Foley music: Learning to generate music from videos. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XI 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 758–775.
27. Koepke, A.S.; Wiles, O.; Moses, Y.; Zisserman, A. Sight to sound: An end-to-end approach for visual piano transcription. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1838–1842.
28. Goldstein, S.; Moses, Y. Guitar Music Transcription from Silent Video. In Proceedings of the BMVC, Newcastle, UK, 3–6 September 2018.
29. Koepke, A.; Wiles, O.; Zisserman, A. Visual pitch estimation. In Proceedings of the Sound and Music Computing Conference, Malaga, Spain, 28–31 May 2019; Society for Sound and Music Computing: Oslo, Norway, 2019.
30. Narasimhan, M.; Ginosar, S.; Owens, A.; Efros, A.A.; Darrell, T. Strumming to the beat: Audio-conditioned contrastive video textures. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 3761–3770.
31. Su, K.; Liu, X.; Shlizerman, E. Multi-instrumentalist net: Unsupervised generation of music from body movements. *arXiv* **2020**, arXiv:2012.03478.
32. Li, W.; Fan, X. Image-text alignment and retrieval using light-weight transformer. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 4758–4762.
33. Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Localizing visual sounds the hard way. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16867–16876.
34. Mandalapu, H.; PN, A.R.; Ramachandra, R.; Rao, K.S.; Mitra, P.; Prasanna, S.M.; Busch, C. Audio-visual biometric recognition and presentation attack detection: A comprehensive survey. *IEEE Access* **2021**, *9*, 37431–37455. [CrossRef]
35. 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]
36. Li, W.; Ma, Z.; Deng, L.J.; Fan, X.; Tian, Y. Neuron-based spiking transmission and reasoning network for robust image-text retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 3516–3528. [CrossRef]
37. Li, W.; Ma, Z.; Shi, J.; Fan, X. The style transformer with common knowledge optimization for image-text retrieval. *IEEE Signal Process. Lett.* **2023**, *30*, 1197–1201. [CrossRef]
38. Narayan, S.; Gupta, A.; Khan, F.S.; Snoek, C.G.; Shao, L. Latent embedding feedback and discriminative features for zero-shot classification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXII 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 479–495.
39. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.A.; Mikolov, T. Devise: A deep visual-semantic embedding model. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2121–2129.

40. Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21969–21980.
41. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. f-vaegan-d2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10275–10284.
42. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4660–4671.
43. Parida, K.; Matiyali, N.; Guha, T.; Sharma, G. 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, Snowmass, CO, USA, 1–5 March 2020; pp. 3251–3260.
44. Mazumder, P.; Singh, P.; Parida, K.K.; Namboodiri, V.P. 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, Waikoloa, HI, USA, 3–8 January 2021; pp. 3090–3099.
45. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2927–2936.
46. Kawoosa, A.I.; Prashar, D.; Faheem, M.; Jha, N.; Khan, A.A. Using machine learning ensemble method for detection of energy theft in smart meters. *IET Gener. Transm. Distrib.* **2023**, *17*, 4794–4809. [CrossRef]
47. Faheem, M.; Al-Khasawneh, M.A. Multilayer cyberattacks identification and classification using machine learning in internet of blockchain (IoBC)-based energy networks. *Data Brief* **2024**, *54*, 110461. [CrossRef] [PubMed]
48. Abubakar, M.; Nagra, A.A.; Faheem, M.; Mudassar, M.; Sohail, M. High-Precision Identification of Power Quality Disturbances Based on Discrete Orthogonal S-Transforms and Compressed Neural Network Methods. *IEEE Access* **2023**, *11*, 85571–85588. [CrossRef]