Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning

Wenrui Li
liwr618@163.com
Harbin Institute of Technology
Harbin, China

Xi-Le Zhao
xlzhao122003@163.com
University of Electronic Science and Technology of China
Sichuan, China

Zhengyu Ma*
mazhy@pcl.ac.cn
Peng Cheng Laboratory
Shenzhen, China

Xingtao Wang
xtwang@hit.edu.cn
Harbin Institute of Technology
Harbin, China

Xiaopeng Fan*
fxp@hit.edu.cn
Harbin Institute of Technology
Harbin, China

Yonghong Tian
yhtian@pku.edu.cn
Peking University
Beijing, China

ABSTRACT

Audio-visual zero-shot learning (ZSL) has attracted board attention, as it could classify video data from classes that are not observed during training. However, most of the existing methods are restricted to background scene bias and fewer motion details by employing a single-stream network to process scenes and motion information as a unified entity. In this paper, we address this challenge by proposing a novel dual-stream architecture **M**otion-**D**ecoupled **S**piking **T**ransformer (MDFT) to explicitly decouple the contextual semantic information and highly sparsity dynamic motion information. Specifically, The Recurrent Joint Learning Unit (RJLU) could extract contextual semantic information effectively and understand the environment in which actions occur by capturing joint knowledge between different modalities. By converting RGB images to events, our approach effectively captures motion information while mitigating the influence of background scene biases, leading to more accurate classification results. We utilize the inherent strengths of Spiking Neural Networks (SNNs) to process highly sparsity event data efficiently. Additionally, we introduce a Discrepancy Analysis Block (DAB) to model the audio motion features. To enhance the efficiency of SNNs in extracting dynamic temporal and motion information, we dynamically adjust the threshold of Leaky Integrate-and-Fire (LIF) neurons based on the statistical cues of global motion and contextual semantic information. Our experiments demonstrate the effectiveness of MDFT, which consistently outperforms state-of-the-art methods across mainstream benchmarks. Moreover, we find that motion information serves as a powerful regularization for video networks, where using it improves the accuracy of HM and ZSL by 19.1% and 38.4%, respectively.

*Corresponding Author: Xiaopeng Fan and Zhengyu Ma (Email: fxp@hit.edu.cn; mazhy@pcl.ac.cn)

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611759

CCS CONCEPTS

• **Information systems** → **Video search**.

KEYWORDS

audio-visual zero-shot learning, spiking neural network

ACM Reference Format:
Wenrui Li, Xi-Le Zhao, Zhengyu Ma, Xingtao Wang, Xiaopeng Fan, and Yonghong Tian. 2023. Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611759

1 INTRODUCTION

The audio-visual zero-shot learning (ZSL) task involves using joint audio and visual modalities to classify or recognize objects or scenes in the absence of labeled data. Supervised audio-visual methods require an adequate number of labeled training instances for each class. However, the classifiers learned through these methods have a limited ability to recognize previously unseen classes and can only classify instances belonging to the classes covered by the training instances. In real-world scenarios, it is impractical to train models on all possible data classes. To address this challenge, the Generalized Zero-Shot Learning (GZSL) setting has been proposed.

Recently, most audio-visual ZSL approaches have focused on aligning temporal and semantic information to obtain more robust audio-visual representations. Mazumder et al. [1] introduce a cross-modal decoder and a composite triplet loss to project data points into their learned embedding space. Mercea et al. [2] utilize cross-attention to integrate information from averaged audio and visual input features, resulting in a computationally lightweight approach. TCaF [3] emphasizes the importance of correlating audio and video modalities by preprocessing temporal information. However, existing approaches primarily focus on modeling and fusing temporal and semantic information more efficiently while neglecting two crucial aspects: **background scene bias** and **motion**.

The previous study [4] revealed that certain action categories in current mainstream video datasets are strongly associated with the background scenery where the action occurs. By focusing on contextual semantic alignment, models tend to degrade by only encoding fine-grained scene information. For instance, a trained model may incorrectly label a video as "reading book" based solely

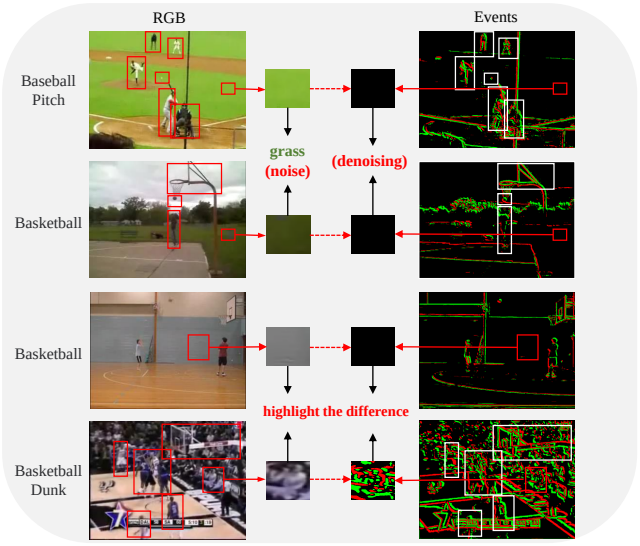

Figure 1: The advantages of converting RBG images to events. The events are only triggered when the background scene are significant changes. Therefore, transforming RBG images into events can mitigate background scene biases and highlight the differences between similar category videos.

on the presence of a library while ignoring the fact that the actual subject is a YouTuber dancing in this scene. This outcome contradicts the original objective of video representation learning and could introduce scene bias when the model is applied to different datasets, an issue that cannot be ignored.

However, decoupling motion information effectively from video input to optimize feature learning remains challenging. To deal with this, we employ an Event Generative Model (EGM) to convert RGB images into events, effectively removing background scene bias and capturing motion information. Specifically, as shown in Fig. 1, our EGM can easily eliminate background noise present in "Baseball Pitch" and "Basketball" images. Furthermore, our EGM captures and highlights critical information, such as the players and the motion trail of the object. Obviously, even for classes with similar semantic contexts, like "Basketball" and "Basketball Dunk," our approach effectively captures the differences in motion information, including spectator reactions, player movements, and ball positions. Event data is formed by the time and location information of event generation, exhibiting spatial and temporal sparsity. To process event data, we utilize Spiking Neural Networks (SNNs), which naturally encode temporal information and directly extract features from event data without additional preprocessing or dimensionality reduction. Moreover, the memorization characteristic of SNNs allows them to adapt well to subsequent attention operations.

In this work, we aim to decouple contextual semantic information and dynamic motion information, which could alleviate the background scene bias and pay more attention to motion in video classification. Specifically, our model employs a Recurrent Joint Learning Unit (RJLU) to capture joint knowledge and connect information from different modalities for efficient joint learning. The RJLU enables the network to learn contextual semantic information and understand the environment in which actions occur. The combination of the EGM and SNNs facilitates the capture of motion and temporal information. Since extracting motion information from audio is challenging, we propose a Discrepancy Analysis Block (DAB) to model the discrepancy between transformed visual event data and audio features, defining them as audio motion features. We dynamically adjust the threshold of Leaky Integrate-and-Fire (LIF) neurons based on the statistical cues of global motion information and contextual semantic information to enhance the efficiency of SNNs in extracting dynamic temporal and motion information.

To sum up, the main contributions of this paper are as follows:

- We propose a dual-stream architecture MSFT to decouple the scenes and motion information. To the best of our knowledge, it is the first work that uses synthetic events with SNN for audio-visual ZSL.
- The EGM and SNN work together to effectively capture highly sparsity motion information while mitigating the influence of background scene biases. The dynamic adjustment of the threshold of LIF neurons further enhances the efficiency of SNNs in extracting temporal and motion information.
- The RJLU is proposed to extract contextual semantic information. The fusion block could integrate the scene and motion information for further inference.

Extensive experimental results validate that our proposed approach outperforms state-of-the-art methods. Moreover, our ablation study demonstrates the effectiveness of each key component of the proposed MDFT.

2 RELATED WORK

2.1 Audio-visual zero-shot learning

With the development of deep learning [5–8], several approaches in audio-visual ZSL aim to learn a joint embedding space that effectively captures the correlation between audio and visual features [9–22]. The Avgzslnet [1] architecture consists of two main components: a multi-modal embedding network and a label feature reconstruction network. Some methods utilize deep neural networks to extract features from both modalities and map them into a shared space, where the similarity can be measured [23]. A coordinated joint multimodal embedding approach is proposed to map both audio and visual features to shared space and incorporates a coordination mechanism to enhance cross-modal semantic alignment [24]. The multimodal embedding network is used to learn joint representations of audio and visual features, which are used to generate multimodal embeddings. The label feature reconstruction network reconstructs the label features of novel audio-visual classes. [25] presents a neural network to learn a mapping between audio and visual features extracted from video frames to align audio and video features for audio-visual speech recognition. Unlike previous methods, our approach simplifies this task by decoupling the scenes and motion information. The combination of EGM and SNNs can capture motion and temporal information and relieve background scene noise, leading to superior zero-shot classification performance.

2.2 Synthetic Events

Recent advancements in event camera technology have led to increased interest in the generative modeling of events [26, 27]. Event cameras capture temporal changes in pixel intensity as events, are generated by changes in the scene's illumination, motion, and texture. The spatiotemporal nature of events and their high temporal resolution makes them suitable for applications such as robotics, autonomous driving, and augmented/virtual reality [28]. Events are generated by applying a threshold to the image difference, resulting in the generation of positive or negative events depending on the intensity difference of the pixels. Some simulators produce high frame-rate images and generate events by interpolating the intensity signals through a tight coupling between the rendering engine and the event simulator. In addition to high frame-rate image rendering and linear interpolation of intensity signals for event generation, Rebecq et al. [29] proposed an adaptive sampling scheme based on the maximum displacement between frames. By selectively sampling frames only when necessary, their approach leads to improved accuracy for fast motion and reduced computation for slow motion scenarios.

2.3 Spiking neural network

Spiking Neural Networks (SNNs) are a type of artificial neural network inspired by the biological neurons and synapses of the brain. Some approaches [30–35] provide an insightful exploration of SNN training strategy, application scenarios, and biological similarities. SNNs use discrete-time and spike-based communication between neurons as opposed to conventional neural networks, which use continuous-valued activations and backpropagation. An SNN's neurons produce spike trains as their outputs, which show the neuron's activity over time. Each neuron in SNNs integrates incoming information from neighbouring neurons and produces a spike when the sum of the signals exceeds a threshold. These spikes' precise timing is essential because it corresponds to the timing of action potentials in real neurons. The transmission of spikes between neurons occurs across synapses, which have a weight associated with them that determines the strength of the connection. The correlation between pre- and post-synaptic spikes is typically used to train an SNN by changing the synaptic weights.

3 METHODOLOGY

The audio and visual features are denoted as $\boldsymbol{a}_i^x$ and $\boldsymbol{v}_i^x$ respectively, the textual labeled embedding of the corresponding ground-truth class i is denoted as $\boldsymbol{t}_i^x$. In the training phase, the seen classes training set with N samples can be written as $\boldsymbol{\mathcal{X}} = (\boldsymbol{a}_i^x, \boldsymbol{v}_i^x, \boldsymbol{t}_i^x)$. The MDFT is proposed to learn a projection function: $f(\boldsymbol{a}_i^y, \boldsymbol{v}_i^y) \mapsto \boldsymbol{g}_j^y$, where $\boldsymbol{g}_j^x$ is class-level textual embedding for class j. The unseen testing set $\boldsymbol{\mathcal{Y}} = (\boldsymbol{a}_i^y, \boldsymbol{v}_i^y, \boldsymbol{t}_i^y)$ can also be projected as $f(\boldsymbol{a}_i^y, \boldsymbol{v}_i^y) \mapsto \boldsymbol{g}_j^y$. The architecture of MDFT is shown in Fig. 2.

3.1 Contextual semantic information modeling

Audio and visual encoder. The robust and discriminative audio and visual features are extracted using pre-trained SeLaVi [36]. After feature extraction, the audio and visual encoders which correspond to E_{aud} and E_{vis} are proposed to further explore the semantic information of different modalities. The outputs of the audio and visual encoder can be written as: $\boldsymbol{a}^t = E_{aud}(\boldsymbol{\mathcal{X}}_a)$ and $\boldsymbol{v}^t = E_{vis}(\boldsymbol{\mathcal{X}}_v)$, where $\boldsymbol{a}^t \in \mathbb{R}^{D_a * D_{emb}}$ and $\boldsymbol{v}^t \in \mathbb{R}^{D_a * D_{emb}}$, each encoder of different modalities consists of a sequence of two linear layers f_1^m and f_2^m for $m \in (\boldsymbol{a}_t, \boldsymbol{v}_t)$. f_1^m: $\mathbb{R}^{D_m * T_{in}} \rightarrow \mathbb{R}^{D_m * T_{hid}}$ and f_2^m: $\mathbb{R}^{D_m * T_{hid}} \rightarrow \mathbb{R}^{D_m * T_{emb}}$. Each of the linear layers is followed by batch normalization, ReLU activation function and dropout with dropout rate d_{enc}. The E_{aud} and E_{vis} could further explore the semantic information for each modality.

Recurrent joint learning unit. (RJLU) To perform audio-visual joint learning more efficiently and integrate the global temporal information into the extracted contextual semantic features, we propose the RJLU, as shown in the bottom-left area of Fig. 2. In this unit, we extract the global joint knowledge in audio and visual features and iteratively update it using a recurrent architecture. First, the audio features $\boldsymbol{a}^t$ and visual features $\boldsymbol{v}^t$ at time step t are going to be integrated with joint knowledge $\boldsymbol{h}^{t-1}$ at time step $t-1$, which can be defined as follows:

$$\boldsymbol{C}_a^t = \mathrm{CA}(\mathrm{AP}(\boldsymbol{a}^t), \mathrm{AP}(\boldsymbol{h}^t)), \boldsymbol{C}_v^t = \mathrm{CA}(\mathrm{AP}(\boldsymbol{v}^t), \mathrm{AP}(\boldsymbol{h}^t)), \quad (1)$$

where $\mathrm{AP}(\cdot)$ represents the average pooling function and $\mathrm{CA}(\cdot)$ denotes the cross attention function. Also, for efficient joint learning, we use joint knowledge as a buffer to connect audio features with visual features. The self-attention function is used to further inference the intrinsic connection between the connected features. Formally, the output of the recurrent joint learning unit is defined as:

$$\begin{aligned} \boldsymbol{C}_{ahv}^t &= \mathrm{CAT}(\boldsymbol{a}^t, \boldsymbol{h}^{t-1}, \boldsymbol{v}^t), \\ \boldsymbol{S}_{ahv}^t &= \mathrm{MLP}(\mathrm{LN}(\mathrm{SA}(\boldsymbol{C}_{ahv}^t))) + \mathrm{SA}(\boldsymbol{C}_{ahv}^t), \end{aligned} \quad (2)$$

where $\mathrm{CAT}(\cdot)$ represents the concatenation operation, $\mathrm{SA}(\cdot)$ represents the self attention function, $\mathrm{LN}(\cdot)$ represents the layer normalization and $\mathrm{MLP}(\cdot)$ represents the multi-layer perceptron. The audio-visual joint features we obtained for this time step will be used to update the previous joint knowledge. To extract robust joint knowledge, we use a hierarchical progressive update strategy, which connected $\boldsymbol{C}_a^t$, $\boldsymbol{C}_v^t$ and $\boldsymbol{S}_{ahv}^t$ in series step by step. Formally, the joint knowledge $\boldsymbol{h}^t$ at time step t is defined as:

$$\begin{aligned} \boldsymbol{C}_{av}^t &= \boldsymbol{C}_a^t + \boldsymbol{C}_v^t, \\ \boldsymbol{h}^t &= \boldsymbol{C}_{av}^t * \boldsymbol{h}^{t-1} + (1 - \boldsymbol{C}_{av}^t) * \boldsymbol{S}_{ahv}^t. \end{aligned} \quad (3)$$

3.2 Motion information modeling (MIM)

Events generate model (EGM). The event camera is a novel type of camera that deviates from the traditional imaging paradigm. Unlike conventional cameras that capture a sequence of images at fixed intervals, an event camera detects changes in brightness within the scene and produces an asynchronous and high-speed event stream. Each event signifies a modification in brightness, either an increment or decrement, and conveys both the precise timing of its occurrence and the corresponding location in the image where it was detected. An event is defined as:

$$e_k = (x_k, y_k, t_k, p_k) \quad (4)$$

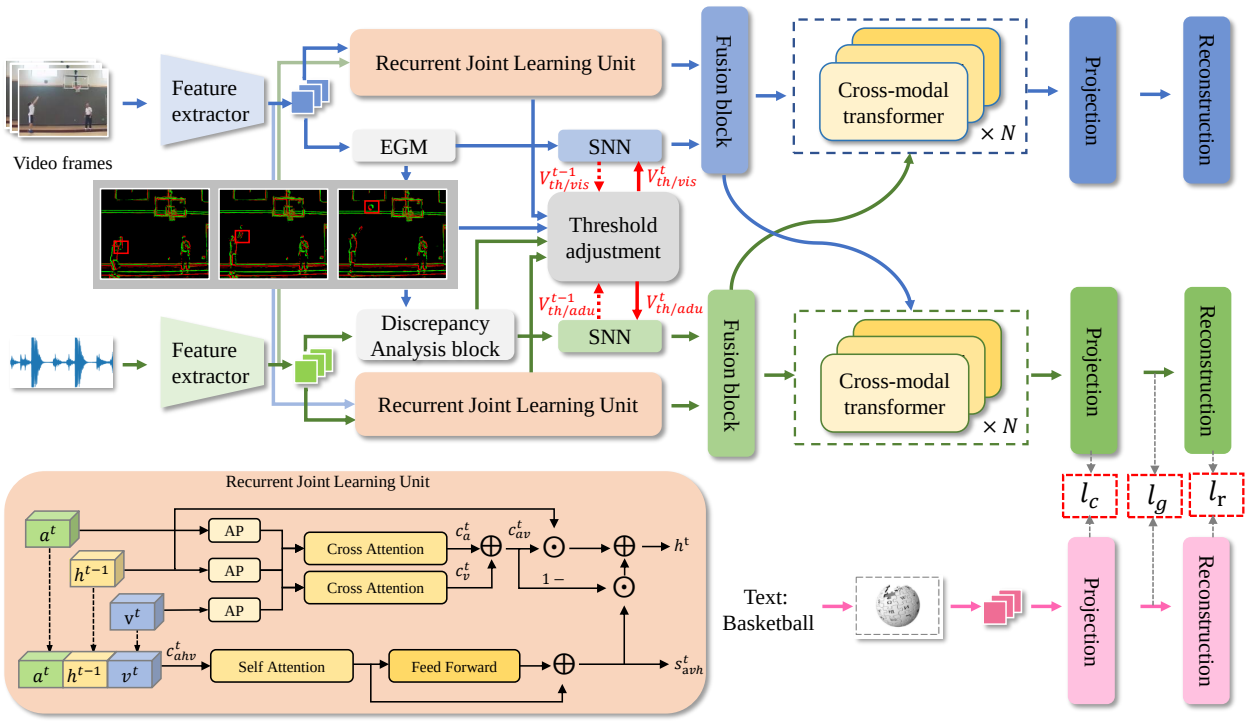

Figure 2: The MDFT architecture integrates visual, audio, and textual features (denoted by blue, green, and pink lines) employing a two-stream architecture for separate extraction of scene contextual semantic and motion information. The "threshold adjustment" block dynamically modulates SNN thresholds ($V^t_{th/vis}$ and $V^t_{th/adu}$) to effectively regulate the firing rates of neurons and mitigate potential noise. The EGM eliminates background scene bias and captures salient motion information, while the DAB ensures a comprehensive representation of audio features. The fusion block effectively fuses scene contextual semantic and motion features. The illustration of the recurrent joint learning unit is displayed in the left-bottom area.

where (x_k, y_k) is the pixel location when the event is triggered, t_k represents the timestamp, $p_k \in \{-1, 1\}$ represents the polarity of this event which demonstrates the direction of the changes.

In our event generation model, an event is triggered when there is a change in the magnitude $\Delta L(\boldsymbol{u}, t_k)$ of the logarithm of brightness at a given pixel $\boldsymbol{u}$ and time t_k that exceeds a predetermined threshold C, since the occurrence of the last event at the same pixel.

$$\Delta L(\boldsymbol{u}, t_k) = L(\boldsymbol{u}, t_k) - L(\boldsymbol{u}, t_k - \Delta t_k) \geq p_k C \tag{5}$$

where Δt_k represents the time since the previous triggered event. The generative event model for an ideal sensor is described by Eq. (1) and (2). The extracted visual motion features as $\boldsymbol{\mathcal{E}}_v = \{e_k\}_{k=1}^N$, where the dimension of $\boldsymbol{\mathcal{E}}_v$ is the same as $\boldsymbol{v}^t$, and all positions without triggered events are filled with 0.

Discrepancy analysis block (DAB). Audio features have inherent temporal characteristics, but defining their motion nature is challenging. Here, we propose to model the motion nature of audio features based on the visual event data, which we define as the difference in audio features. Formally, the discrepancy matrix is calculated as:

$$\boldsymbol{\mathcal{E}}_a = \boldsymbol{a}^t + (1 - exp(-\left\| \frac{\boldsymbol{\mathcal{E}}_v - \boldsymbol{a}^t}{\beta_a} \right\|_2^2)). \tag{6}$$

where β^a is the learnable factor. Through the discrepancy matrix, SNNs can extract both the motion and temporal characteristics of the audio features simultaneously. This allows our model to capture a more comprehensive representation of the audio features.

Dynamic threshold block. The dynamic spike threshold serves as a spontaneous regulatory mechanism that mitigates excessive excitation and prevents neuronal death. In this block, we adaptively adjust the spike threshold of LIF neurons based on scene contextual semantics and motion features. We introduce a metric $\boldsymbol{\varphi}$ to represent the global properties of contextual semantic features, along with an entropy value $\boldsymbol{\omega}$ to quantify the richness of motion features, which are used to define the spike threshold. Formally, the dynamic spike threshold in the visual pipeline is defined as follows:

$$\begin{aligned} V^t_{th} &= (\boldsymbol{\varphi} + \boldsymbol{\omega}) V^{t-1}_{th}, \\ \boldsymbol{\varphi} &= \text{Sigmoid}(\text{AP}(S^t_{ahv})), \\ \boldsymbol{\omega} &= -\mathcal{N}(S^t_{ahv}) log(\frac{1}{\mathcal{N}(S^t_{ahv})} + \boldsymbol{\mathcal{E}}^t_v). \end{aligned} \tag{7}$$

where V^t_{th} represents the spiking threshold at time t and $\mathcal{N}(\cdot)$ is normalization operation. In general, a high $\boldsymbol{\varphi}$ value signifies a substantial amount of contextual semantic information at a specific moment, resulting in a decreased firing rate of neurons. The entropy

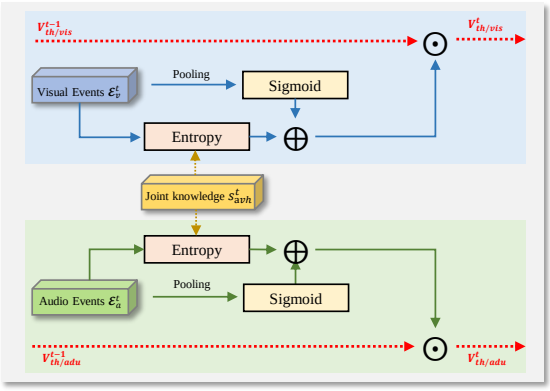

Figure 3: The architecture of dynamic threshold block.

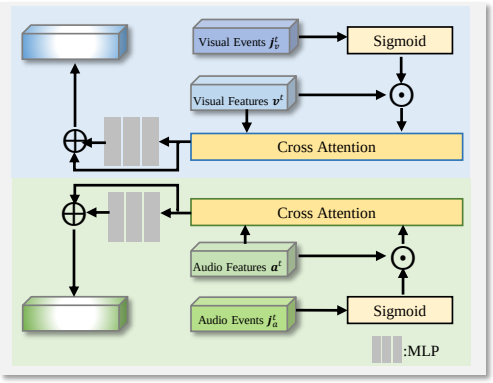

Figure 4: The architecture of the fusion block.

value ω conveys the richness of information embedded in motion features. A high entropy value indicates drastic changes within a scene (e.g., due to camera shake), necessitating an elevation in the spike threshold to filter out potential noise. On the other hand, a reduced threshold is preferred to boost sensitivity to information. This strategy effectively regulates the firing rate of neurons and adaptively adjusts the spike threshold based on different contexts, providing significant implications for enhancing the performance of audio-visual zero-shot learning.

Spiking neural network (SNN). SNNs are a type of neural network that utilize brief pulses of activity, known as spikes, to communicate information between neurons. This spike-based communication is similar to the event-driven format of event cameras, making SNNs a natural choice for processing event camera data.

Our SNN network consists of three linear SNN blocks, each comprising a linear layer followed by a LIF-based layer [37]. The inputs of the SNN layer are membrane potentials, obtained by converting spikes using a linear projection layer. Specifically, the membrane potential of a LIF neuron can be described by the following equation:

$$\tau_m \frac{dV(t)}{dt} = -V(t) + RI(t), \tag{8}$$

where τ_m is the membrane time constant, $V(t)$ is the membrane potential at time t, R is the membrane resistance, and $I(t)$ is the current input at time t. To compute the input of the i-th LIF neuron $I_i(t)$, we calculate the dot product between the output of the previous layer and the synaptic weight vector $\boldsymbol{w}_i$ and add the bias term b_i as:

$$I_i(t) = \boldsymbol{w}_i^T \boldsymbol{s}(t) + b_i. \tag{9}$$

When the membrane potential of the i-th LIF neuron reaches the threshold V_{th}, the neuron fires a spike and the membrane potential is reset to the resting potential V_{rest}.

Audio-visual input pairs of motion features are processed by the proposed SNN block in time order. To capture more comprehensive temporal and motion cues, each LIF neuron maintains its previous membrane potential instead of resetting the initial potential for each new input. At the end of the SNN block, the accumulation layer resets the membrane potential after processing all the information pairs. Our proposed SNN block demonstrates the powerful temporal and motion information representation ability, which is essential for audio-visual analysis and understanding.

3.3 Cross-modal reasoning module (CRM)

The CRM is proposed to efficiently integrate the temporal and semantic audio-visual joint features of different modalities. To complement the information between the audio and visual features, we adopt a residual connection between two layers, followed by layer normalization. The outputs of the audio attention fusion block can be written as follows:

$$\begin{aligned} \boldsymbol{R}_a &= \mathrm{CA}(\boldsymbol{a}^t, \boldsymbol{a}^t * \mathrm{Sigmoid}(\boldsymbol{j}_{\mathrm{a}}^{\mathrm{t}})), \\ \boldsymbol{P}_a &= \mathrm{MLP}(\mathrm{LN}(\boldsymbol{R}_a)) + \boldsymbol{R}_a, \end{aligned} \tag{10}$$

where $\boldsymbol{j}_a^t$ represents the output of the SNN in audio pipeline. The intrinsic relationship between the fused semantic, temporal features, and joint knowledge needs to be further explored by cross-modal transformers. The cross-modal transformer integrates the information from different modalities for feature inference and ultimately obtains a joint audio-visual representation. This block consists of a stack of standard transformer layers. The cross-modal transformer layer can be written as follow:

$$\begin{aligned} \boldsymbol{Z}_{av} &= \mathrm{MHCA}(\boldsymbol{P}_v, \boldsymbol{P}_a), \\ \boldsymbol{F}_{av} &= \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Z}_i)) + \boldsymbol{Z}_{av}, \end{aligned} \tag{11}$$

where $\mathrm{MHCA}(\cdot)$ represents the multi-head cross attention.

Our goal is to predict the final textually labeled class of the inputs. The projection and reconstruction layers are used to project audio-visual joint embeddings into the textual labeled embedding space and reconstruct the projected features to recover the original information, making features of different modalities comparable. The projection layer is a stack of two linear layers f_3^m and f_4^m, each of the linear layer are followed by batch normalization, a ReLU activation function and dropout with rate d_{proj}. $f_3^m : \mathbb{R}^{D_m * T_{emb}} \rightarrow \mathbb{R}^{D_m * T_{hid}}$ and $f_4^m : \mathbb{R}^{D_m * T_{hid}} \rightarrow \mathbb{R}^{D_m * T_{fin}}$. The final audio-visual joint feature embeddings can be obtained as:

$$\boldsymbol{O}_{av} = \mathrm{AV}_{proj}(\boldsymbol{F}_{av}), \tag{12}$$

where AV_{proj} represents the projection function. The final textual labeled embedding $\boldsymbol{O}_w$ is obtained by projecting j-th class label embedding $\boldsymbol{w}_j$ by the word projection layer W_{proj}. The architecture

of the W_{proj} is similar with AV_{proj} but different from dropout rate d_{wproj}.

3.4 Training strategy

We trained the MDFT model on a single Nvidia V100S GPU, following the procedure for extracting audio and visual embeddings per second as described in [2]. Specifically, we set $T_{in} = 512$, $T_{hid} = 512$, $T_{proj} = 64$, and $T_{fin} = 300$. The dropout rates for the VGGSound, UCF, and ActivityNet datasets are $d_{enc} = 0.20/0.25/0.10$, $d_{dec} = 0.25/0.20/0.15$, and $d_{wproj} = 0.1/0.1/0.1$, respectively. In the cross-modal transformer, we employ 8 heads with a dimension of 64 for each head. The training is optimized using the Adam optimizer, and the MDFT model is trained for 50 epochs with a learning rate of 0.0001. To learn more effective feature representations, our model is updated using the loss function $\mathcal{L}_{all}$, which comprises a joint triplet loss $\mathcal{L}_n$, a projection loss $\mathcal{L}_p$, and a reconstruction loss $\mathcal{L}_r$.

Joint triplet loss. The joint triplet loss could cluster the final audio-visual embeddings to make the results more reasonable. The joint triplet loss $\mathcal{L}_n$ can be written as:

$$\mathcal{L}_n = [\gamma + \boldsymbol{O}^+_{av} - \boldsymbol{O}^+_w]_+ + [\gamma + \boldsymbol{O}^-_{av} - \boldsymbol{O}^+_w]_+ \tag{13}$$

where γ is margin parameters that define the minimum separation between negative pairs of different modalities and truly matching audio-visual embeddings, $\boldsymbol{O}_w$ represents the textual embeddings, $\boldsymbol{O}^+_{av}$ and $\boldsymbol{O}^-_{av}$ correspond to positive and negative examples respectively, and $[x]_+ \equiv \max(x, 0)$.

Projection loss. The distance between the output joint embeddings of projection layer and corresponding textual labeled embedding is reduced by the projection loss, which can be written as:

$$\mathcal{L}_p = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{O}_{av} - \boldsymbol{O}_w), \tag{14}$$

where n is the number of samples.

Reconstruction loss. The reconstruction loss is proposed to ensure that maintain the original data distribution while projecting audio-visual features to shared embedding space. The architecture of reconstruction layer is the same as projection layer. The reconstruction loss $\mathcal{L}_r$ can be written as:

$$\mathcal{L}_r = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{O}^{rec}_{av} - \boldsymbol{O}_w), \tag{15}$$

where $\boldsymbol{O}^{rec}_{av}$ is the output of the reconstruction layer. The total loss is formulated as $\mathcal{L}_{all} = \mathcal{L}_n + \mathcal{L}_p + \mathcal{L}_r$.

4 EXPERIMENT

In this study, we conduct a comprehensive evaluation of our proposed model in both ZSL and GZSL scenarios. As suggested by [2, 43], we calculate the mean class accuracy for all models to quantify their effectiveness in classification tasks. For the ZSL evaluation, we specifically analyze the test samples from the subset of unseen test classes. As for GZSL evaluation, we assess the models on the entire test set, encompassing both seen (S) and unseen (U) classes. This allows us to calculate their harmonic mean HM = $\frac{2\text{US}}{\text{U+S}}$.

4.1 Dataset statistic

In this study, we utilize three benchmark datasets ActivityNet, VGGSound, and UCF101 to conduct our experiments and evaluate the proposed models.

ActivityNet: The ActivityNet dataset [23] is a mainstream video benchmark, consisting of 200 activity classes and approximately 27,801 video clips. It is designed for human activity understanding and covers a wide variety of complex activities, such as sports, hobbies, and daily routines. The diverse set of activities and temporal variations make it a valuable resource for evaluating video understanding models.

UCF101: The UCF101 dataset [44] comprises 13,320 video clips, spanning 101 distinct action categories. These categories encompass a broad range of human activities, such as sports, musical instruments, and body movements. The dataset is divided into 25 groups, with each group consisting of 4-7 action categories.

VGGSound: The VGGSound dataset [45] contains 212,894 video clips, covering 309 distinct audio categories. The categories include various acoustic events, such as musical instruments, human speech, animal sounds, and ambient noises.

4.2 Comparison with state-of-the-art

To validate the effectiveness of our model, we have compared it with the current state-of-the-art audio-visual (G)ZSL methods on three mainstream benchmark datasets. In this section, we will delve deeper into the details and discuss the differences between various methods and our model.

[2, 3] leverage cross-modal attention and textual label embeddings to learn multi-modal representations and transfer knowledge from seen to unseen classes. SJE [38] learns a compatibility function between image and class embeddings, assigning higher scores to match embeddings than mismatching ones. DEVISE [39] presents a deep visual-semantic embedding model that leverages labeled image data and semantic information from unlabeled text for scalable object recognition. Apn [40] proposes a novel zero-shot representation learning framework that jointly learns global and local features using class-level attributes to improve attribute-based knowledge transfer. VAEGAN [41] employs a conditional generative model that combines VAE and GANs strengths and leverages the marginal feature distribution of unlabeled images. Distinct from the above methods, our model emphasizes the motion information within videos, aiming to mitigate the negative impact of background scene biases on zero-shot learning tasks. Moreover, by leveraging the powerful time-series modeling capabilities of spiking neural networks, we can concurrently extract temporal and motion features, efficiently integrating them with contextual semantic information to yield more comprehensive audio-visual features.

Results comparison We demonstrate the superiority of the MDFT framework by comparing it with recent methods in Table 1. On the VGGSound-GZSL dataset, MDFT attains an HM of 8.72 and a ZSL performance of 7.13. In comparison with AVCA, the performance of HM and ZSL increased by 38.2 and 18.8 respectively. As for the UCFGZSL dataset, MDFT obtains a GZSL performance of 31.36, slightly lower than 31.72 of TCaF. We attribute this improvement to TCaF's preprocessing operation for the temporal alignment of input audio and visual features. However, MDFT outperforms TCaF

Table 1: The performance of our MDFT and state-of-the-art baselines for audio-visual (G)ZSL on three benchmark datasets.

Type	Model	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
		S	U	*HM*	*ZSL*	S	U	*HM*	*ZSL*	S	U	*HM*	*ZSL*
ZSL	SJE [38]	48.33	1.10	2.15	4.06	63.10	16.77	26.50	18.93	4.61	7.04	5.57	7.08
	DEVISE [39]	36.22	1.07	2.08	5.59	55.59	14.94	23.56	16.09	3.45	8.53	4.91	8.53
	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
Audio-visual ZSL	CJME [42]	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 [1]	18.05	3.48	5.83	5.28	52.52	10.90	18.05	13.65	8.93	5.04	6.44	5.40
	AVCA [2]	14.90	4.00	6.31	6.00	51.53	18.43	27.15	20.01	24.86	8.02	12.13	9.13
	TCaF [3]	9.64	5.91	7.33	6.06	58.60	21.74	**31.72**	24.81	18.70	7.50	10.71	7.91
	MDFT (ours)	16.14	5.97	**8.72**	**7.13**	48.79	23.11	31.36	**31.53**	18.32	10.55	**13.39**	**12.55**

Table 2: Ablation study of the superiority of SNN in processing event information.

Model	UCF-GZSL			
	S	U	*HM*	*ZSL*
MLP	52.28	14.98	23.29	13.65
EGM+MLP	46.52	16.34	24.19	18.93
SNN	45.29	19.46	27.23	26.65
MDFT(EGM+SNN)	48.79	23.11	**31.36**	**31.53**

Table 3: Ablation study on UCF-GZSL.

Model	UCF-GZSL			
	S	U	*HM*	*ZSL*
W/o MIM	50.64	17.74	26.34	22.78
W/o RJLN	38.79	20.11	26.44	23.14
W/o CRM	44.72	21.92	29.44	27.67
MDFT	48.79	23.11	**31.36**	**31.53**

for ZSL with a performance of 29.34 compared to 24.81. On the ActivityNet-GZSL, MDFT outperforms AVGZSLNet in GZSL performance with 13.39 compared to 6.44. For ZSL, MDFT reaches a score of 11.94, considerably higher than CJME's 5.84. In conclusion, the MDFT framework consistently exhibits outstanding performance across multiple datasets, particularly in the ZSL test setting.

4.3 Ablation study

The superiority of SNN in processing event information. We present the superiority of SNN in processing event information in Table 2. In our MIM branch, we adopt a combination of EGM and SNN. We replace the three layers SNN with three layers MLP as a comparison, and perform various combinations of the modules. We conduct experiments using only SNN and MLP, and the combination of EGM and MLP, denoted as "SNN," "MLP," and "EGM+MLP," respectively. It is evident that EGM+SNN achieves the best GZSL and ZSL performance. Only using MLP performs best on the seen classes but has a significant gap with our model on unseen classes and ZSL. We believe this is due to background scene bias, which makes it difficult for the model to focus on the fine-grained and crucial motion clues that occur within the scene. It is worth noting that when using the combination of EGM and MLP, the performance in GZSL and ZSL is even worse than using only SNN. This demonstrates that traditional MLP is not suitable for highly sparsity event information and motivated us to use SNN to process event information, which is effective.

The effectiveness of MSFT components. We evaluate the impact of different components of our model in Table 3. The model without the recurrent joint learning unit, motion information modeling, and cross-modal reasoning module are denoted as "W/o RJLN," "W/o MIM," and "W/o CRM," respectively. The ablation study confirms that each component of our model is effective. When any component is removed, the performance of the MDFT declines. MIM is the most critical component. The transformation of images into events helps alleviate the noise caused by background scene biases. The integration of robust temporal features extracted from the SNN helps capture the relationships between different modalities, ultimately improving zero-shot classification performance. It is worth mentioning that when the MIM is removed, the performance on unseen classes significantly decreases, but there is a slight increase in the seen class. This is due to the inconsistency in the data distribution between the SNN and RJLN outputs. In our future work, we would investigate how to efficiently and rationally fuse the features of both. The recurrent joint learning unit leverages the dynamic joint knowledge among different modalities to fully extract and integrate temporal and semantic features for efficient audio-visual joint learning. The cross-modal transformer further explores the relationships between the fused features.

The impact of different loss items. We investigated the effects of the various loss functions in Table 4. Through a comprehensive comparison of experimental outcomes, it was observed that utilizing the complete loss function yielded the best HM and ZSL performance across the UCF-GZSL, VGGSound-GZSL, and

Table 4: Ablation study of different loss items.

Loss	UCF-GZSL			
	S	U	*HM*	*ZSL*
W/o $\mathcal{L}_n$+$\mathcal{L}_r$	28.71	8.52	13.14	10.11
W/o $\mathcal{L}_n$+$\mathcal{L}_c$	34.17	10.14	15.64	13.19
W/o $\mathcal{L}_n$	40.77	16.79	23.78	16.98
W/o $\mathcal{L}_r$	42.78	16.96	24.29	19.82
W/o $\mathcal{L}_p$	43.17	19.27	26.65	23.11
MDFT	48.79	23.11	**31.36**	**29.34**

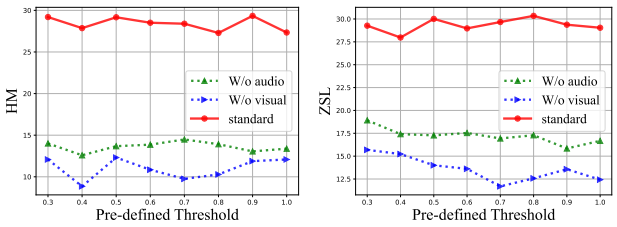

Figure 5: Ablation study on different pre-defined thresholds with unimodal and multi-modal inputs in UCF-GZSL dataset.

ActivityNet-GZSL. Specifically, for the UCF-GZSL, excluding the loss function $\mathcal{L}_n$ during the model training process had the most pronounced impact on both the HM and ZSL results, which are 23.91 and 16.98, respectively. However, employing the complete loss function $\mathcal{L}_{all}$ led to 31.36 and 29.34 for HM and ZSL performance, respectively. This experiment verifies the indispensable role of each loss function within the model training process. Consequently, the adoption of the complete loss function in the training of the MDFT model could ensure enhanced GZSL and ZSL performance.

The effectiveness of dynamic threshold. To prove the benefits of employing an adaptive spiking threshold along with the integration of both audio and visual inputs, we present the outcomes of training our model using multimodal and unimodal inputs with various fixed spiking thresholds in Fig. 5. It is evident that the performance of the model trained with distinct fixed spiking thresholds exhibits considerable fluctuations, suggesting that our model is highly sensitive to the neurons' spiking thresholds. The MDFT using dynamic threshold demonstrates the superiority of all other methods that use fixed spiking thresholds. Generally, models trained exclusively with visual inputs yield better results than those trained with audio inputs. The MDFT simultaneously trains with both visual and audio modalities, demonstrating superior performance compared to unimodal inputs. While visual features play a crucial role in audio-visual ZSL, audio features can also contribute to the optimization of visual information.

4.4 Qualitative results

To demonstrate the importance of capturing motion information, we present qualitative results with and without the MIM branch visualization in Fig. 6. Taking the "Bowling" class in the seen class retrieval as an example, when we ignore motion information, the

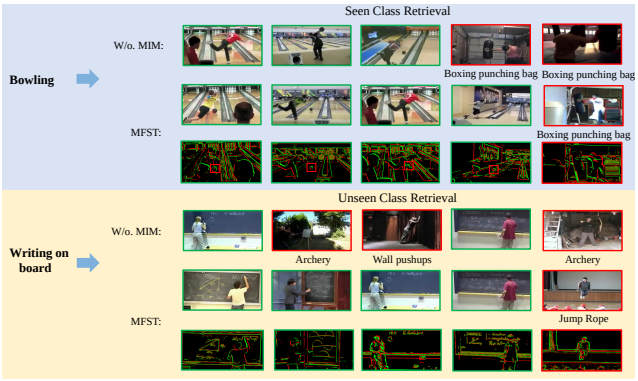

Figure 6: Qualitative comparison results of ablation study.

model's ability to capture fine-grained motion actions is limited. However, with the proposed MIM, the model can better identify and highlight the position of the bowling ball, improving the classification performance. Similarly, in the retrieval of unseen classes in the bottom part of Fig. 6, having MIM can better transfer the knowledge of seen classes in training to unseen classes in testing. By converting images into event sequences, background scene biases can be intuitively eliminated, resulting in more accurate classification results.

5 CONCLUSION

In conclusion, we have presented a novel Motion-Decoupled Spiking Transformer for Audio-Visual ZSL. Our model aims to mitigate the influence of background scene biases on video classification by focusing more on dynamic motion information. We propose a dual-stream architecture to decouple contextual semantic information and dynamic information. The RJLU is designed to capture joint knowledge and connect information from different modalities. The RJLU enables the network to learn contextual semantic information and understand the environment in which actions occur. The SNN could naturally encode temporal information and directly extract features from event data. As extracting motion information from audio is challenging, we propose a DAB to model the discrepancy between transformed visual event data and audio features, defining them as audio motion features. We dynamically adjust the threshold of LIF neurons based on the statistical cues of global motion information and contextual semantic information to improve the efficiency of SNNs in extracting dynamic temporal and motion information. Our experiments demonstrate the effectiveness of the proposed dual-stream architecture and The superiority of SNN in processing event information.

6 ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (2021YFF0900500, 2020YFA0714001) and the National Natural Science Foundation of China (NSFC) under grants U22B2035, 62272128, 62206141, 62236009, 62131005, 12171072, 61825101 and 62027804. Open Research Fund Program of Data Recovery Key Laboratory of Sichuan Province (Grant No. DRN2302).

REFERENCES

[1] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[2] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Otniel-Bogdan Mercea, Thomas Hummel, A Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. *European Conference on Computer Vision (ECCV)*, 2022.

[4] Jinpeng Wang, Yuting Gao, Ke Li, Jianguo Hu, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[5] Zhuangzhuang Chen, Jin Zhang, Zhuonan Lai, Jie Chen, Zun Liu, and Jianqiang Li. Geometry-aware guided loss for deep crack recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Zhuangzhuang Chen, Jin Zhang, Pan Wang, Jie Chen, and Jianqiang Li. When active learning meets implicit semantic data augmentation. In *European Conference on Computer Vision (ECCV)*.

[7] Jianqiang Li, Zhuang-Zhuang Chen, Luxiang Huang, Min Fang, Bing Li, Xianghua Fu, Huihui Wang, and Qingguo Zhao. Automatic classification of fetal heart rate based on convolutional neural network. *IEEE Internet of Things Journal*, 2019.

[8] Jianqiang Li, Zhuangzhuang Chen, Jie Chen, and Qiuzhen Lin. Diversity-sensitive generative adversarial network for terrain mapping under limited human intervention. *IEEE Transactions on Cybernetics*, 2021.

[9] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision (ECCV)*, 2020.

[10] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4281–4289, 2018.

[12] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[13] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[14] Yizhe Zhu, Jianwen Xie, Bingchen Liu, and Ahmed Elgammal. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[15] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[16] Alina Roitberg, Manuel Martinez, Monica Haurilet, and Rainer Stiefelhagen. Towards a fair evaluation of zero-shot action recognition using external data. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[17] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Shreyank N Gowda, Laura Sevilla-Lara, Kiyoon Kim, Frank Keller, and Marcus Rohrbach. A new split for evaluating true zero-shot action recognition. In *Pattern Recognition: 43rd DAGM German Conference (DAGM GCPR)*, 2022.

[19] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019.

[20] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning (ICML)*, 2015.

[21] Huang Xie, Okko Räsänen, and Tuomas Virtanen. Zero-shot audio classification with factored linear and nonlinear acoustic-semantic projections. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

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

[23] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[25] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2018.

[26] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[28] Jacques Kaiser, J. Camilo Vasquez Tieck, Christian Hubschneider, Peter Wolf, Michael Weber, Michael Hoff, Alexander Friedrich, Konrad Wojtasik, Arne Roennau, Ralf Kohlhaas, Rüdiger Dillmann, and J. Marius Zöllner. Towards a framework for end-to-end control of a simulated vehicle with spiking neural networks. In *IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, 2016.

[29] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Proceedings of The 2nd Conference on Robot Learning*, Proceedings of Machine Learning Research (PMLR), 2018.

[30] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[31] Yanqi Chen, Zhaofei Yu, Wei Fang, Zhengyu Ma, Tiejun Huang, and Yonghong Tian. State transition of dendritic spines improves learning of sparse spiking neural networks. In *International Conference on Machine Learning (ICML)*, 2022.

[32] Yanqi Chen, Zhaofei Yu, Wei Fang, Tiejun Huang, and Yonghong Tian. Pruning of deep spiking neural networks through gradient rewiring. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021.

[33] Liwei Huang, Zhengyu Ma, Liutao Yu, Huihui Zhou, and Yonghong Tian. Deep spiking neural networks with high representation similarity model visual pathways of macaque and mouse. *arXiv preprint arXiv:2303.06060*, 2023.

[34] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[35] Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Xiaopeng Fan, and Yonghong Tian. Neuron-based spiking transmission and reasoning network for robust image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2022.

[36] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[37] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience*, 2016.

[38] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[40] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[41] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[43] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[45] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.