Enhancing Recognition of Stereotyped Movements in ASD Children through Action Pattern Mining and Multi-Channel Fusion

Baiqiao Zhang, Yanran Yuan, Wei Qin, Xiangxian Li, Weiying Liu, Wenxin Yao, Yulong Bian and Juan Liu

***Abstract*— Stereotyped movements play a crucial role in diagnosing Autism Spectrum Disorder (ASD). However, recognizing them poses challenges, due to limited data availability and the movements' specificity and varying duration. To support in-depth analysis of ASD children's movements, we constructed the ACSA653 dataset, comprising 653 videos across six classes of stereotyped movements. This dataset surpasses existing ones in both scale and category. To improve the recognition of stereotyped movements, we propose APMFNet, a model that integrates three modules: Visual Motion Learning (VML), Skeleton Relation Mining (SRM), and Multi-channel Fusion (MF). The VML module focuses on extracting spatial and motion information from RGB and optical-flow sequences. The SRM module effectively mines essential motion patterns associated with stereotyped movements through cross-modal graph. The MF module fuses multi-modal information through cross-modality attention to facilitate decision-making. Tested on ACSA653, APMFNet outperforms current state-of-the-art methods, suggesting its potential to identify stable patterns of stereotyped movements in children with ASD.**

***Index Terms*— Autism spectrum disorder, Human activity recognition, Multi-channel fusion, Stereotyped movement**

I. Introduction

Autism Spectrum Disorder (ASD) includes a range of neuro-developmental disorders characterized by challenges in social skills, communication abilities, and behavioral expression. Timely diagnosis of ASD is essential to mitigate the development of additional symptoms and minimize the impact on social interaction and behavior [1]. The two major symptoms of ASD are social interaction impairment and restricted and repetitive patterns of behavior, interests, or activities. Stereotyped movements stand as one of the primary indicators [2]. Therefore, the recognition of stereotyped movements becomes a key factor in diagnosing ASD. Traditional methods for diagnosing ASD-related stereotyped movements, such as scale assessments, observations, and video analyses, not only require involvement from medical professionals but also suffer from high labor intensity, time consumption, and subjectivity. Consequently, there is an urgent need to develop more efficient and accurate methods for recognizing stereotyped movements.

Data-driven approaches prove effectively in recognizing patterns of stereotyped movements [3]–[7], thereby alleviating the workload on experts and physicians. While wearable sensor-based methods have been employed for activity recognition [8]–[12], the use of inertial sensors may be intrusive, and children with Autism Spectrum Disorder (ASD) may experience discomfort or resistance, leading to reduced compliance. As a result, an increasing number of studies are focusing on video-based methods. These video-based approaches can be divided into three main categories: The first, visual-based methods [7], [13]–[16], focusing on extracting visual features related to ASD children from video frames, encompass RGB data and optical flow. This approach predicts types of stereotyped movements by parsing a sequence of visual information. Despite capturing rich spatiotemporal dimensions, its performance is constrained by data quality, lighting conditions, complex backgrounds, and noise . Therefore, there is need for further research to achieve accurate classification of stereotyped movements. Second, skeleton-based methods [3], [6], [17], extracts the positions of skeleton points of ASD children from video frames and predicts types of stereotyped movements by analyzing these skeleton point trajectories. This approach partially alleviates the impact of background noise to some extent but still suffers from the quality of visual information extraction. Additional studies are also required to enhance the recognition of stereotyped movements based on skeleton point motion patterns. The third category, vision-skeleton multi-modality fusion methods, combines the previous two approaches by integrating different modalities of data to enhance the performance of stereotyped movement recognition. This method not only inherits the richness of spatio-temporal information from vision-based approaches but

This work was partially supported by the Ministry of Education of Humanities and Social Science project (22C10422009), the National Natural Science Foundation of China (6227071327; 62007021), and the Young Scholars Program of Shandong University, Weihai (20820211005).

Baiqiao Zhang, Weiying Liu, Wenxin Yao, Yulong Bian and Juan Liu (Corresponding author) are with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China (emails: baiqiao@mail.sdu.edu.cn; 2453741601@qq.com; yaowenxin837@163.com; bianyulong@sdu.edu.cn; zzzliujuan@sdu.edu.cn).

Yanran Yuan and Wei Qin are with Jining No.1 People's Hospital, Jining 272011, China (emails: yuanyanran@163.com; qinwei1271541@163.com.

Xiangxian Li is with the School of Software, Shandong University, Jinan 250101, China (email: xiangxian_lee@mail.sdu.edu.cn).

Yulong Bian and Juan Liu (Corresponding author) are also associated with the Engineering Research Center of Digital Media Technology, Ministry of Education, Jinan 250101, China (emails: bianyulong@sdu.edu.cn; zzzliujuan@sdu.edu.cn).

also leverages the advantages of skeleton-based methods in minimizing background noise interference. However, current vision-skeleton multi-modality fusion methods have not been applied to the recognition of stereotyped movements in ASD. Therefore, video-based ASD stereotyped movements recognition still faces the following challenges: 1) the lack of data and detailed categories of ASD stereotypical movements limit the effects of algorithms, 2) the specificity and varying durations of stereotyped movements make it difficult to capture relevant patterns.

To address these challenges, we collected a novel video-based dataset for stereotyped movements in children with ASD, named ACSA653. With the advancement of mixed reality (MR) technology, diagnostic and rehabilitation approaches for ASD have evolved significantly [18]–[20], enabling the standardized collection of ASD behavioral data. The ACSA653 dataset was collected during MR cognitive training sessions and consists of 653 videos documenting six types of physician-annotated stereotyped movements. It aims to enhance the understanding of these movements in children with ASD, offering greater category diversity and scale compared to existing public datasets [3], [7], [16], [21]–[23].

Furthermore, we propose an action pattern mining and multi-modal fusion-based method for recognizing stereotyped movements, termed APMFNet. This network effectively constructs structural models of skeleton point information and ensures comprehensive fusion of features derived from both visual and skeletal information. APMFNet consists of three modules: **Visual Motion learning (VML)**, **Skeleton Relationship Mining (SRM)** and **Multi-Channel Fusion (MF)**, as illustrated in Figure 1. In the **VML** module, we employ RGB and optical flow data; by combining these two modalities, we extract both dense and sparse information from the visual channel, achieve the fusion of appearance and motion information. In the **SRM** module, we construct a cross-modality graph to represent the skeleton mapping, effectively capturing the spatial relationships and dynamic changes between joints and bones. We employ Adaptive Graph Convolutional Networks (AGCN) to accurately extract specific patterns of stereotyped movements. Finally, in the **MF** module, we used cross-modality attention to fuse the information from visual and skeletal modality. Through this approach, we effectively integrate the motion and contextual background information, thus improving recognition accuracy and robustness.

Our comparison experiments on the ACSA653 dataset indicate that our proposed APMFNet outperforms existing models for recognition ASD stereotyped movements with an accuracy rate of 85.71% across six categories. In our ablation study, we validated the effectiveness of each module in the APMFNet framework. Additionally, using the Grad-CAM [24] visualization method, we demonstrated how the model focuses on key parts and identifies stable patterns of stereotyped movements in each modality. This not only enhances the model's interpretability but further confirms its efficacy.

Thus, the main contributions of this paper are as follows:

- We have constructed a dataset, called ACSA653, covering 653 videos, representing six different classes of stereotyped movements. This dataset has significant advantages in both volume and variety.
- We propose a multi-modal fusion framework for ASD stereotyped movements recognition through the effective construction of structured modeling of skeleton point information and the achievement of efficient fusion of visual and structural information.
- Through extensive experiments and case studies, we achieved the State-of-the-Art (SOTA) performance in ASD stereotyped movements recognition; moreover, through deep experiment analysis, the complementarity between skeleton point information and video information is explored, and a stable patterns of stereotyped movements is computed.

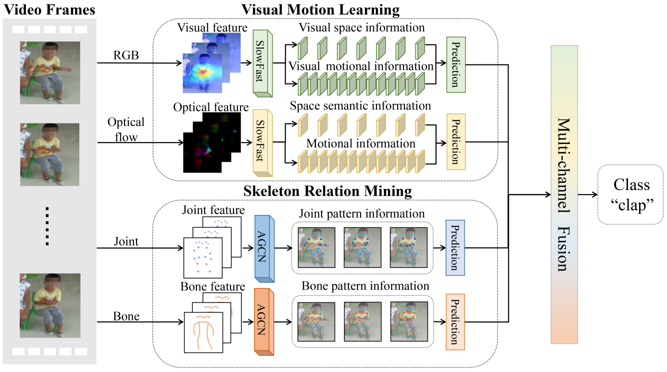


Fig. 1. The illustration of APMFNet, which extracts visual and skeleton information via two separate modules. Visual information encompasses spatial and motion data from RGB and optical flow, while skeleton information comprises joint and bone pattern information. APMFNet fuses the prediction from each modality to yield the final prediction.

II. Related Works

Traditional methods for recognizing ASD stereotyped movements are mainly based on observation and rating scales, yielding in time-consuming and subjectivity. Although sensors have also been deployed for recognition, their analytical capabilities are limited and may be intrusive. Therefore, visual methods are gradually introduced. Moreover, current work can mainly be divided into three categories: Video-based Stereotyped Movement Recognition, Skeleton-based Stereotyped Movement Recognition, and Multi-Modality Fusion.

A. Video-based Stereotyped Movement Recognition

Currently, there are few studies on ASD that employ RGB video data for tasks related to stereotyped movement recognition. However, methods designed for action recognition based in videos can be adapted effectively for detecting stereotyped movements in ASD. These approaches primarily depend on video data, preserving a more comprehensive range of information. The four primary types of video-based action recognition methods include handcrafted feature methods, static image feature aggregation methods, 3D convolutional methods, dual-stream methods and self-supervised learning methods.

1) Handcrafted Feature Methods: The handcrafted feature method focuses on local feature or holistic feature detectors and descriptors. These features, crucial in determining the final recognition rate, are designed empirically [25]. For example, Negin et al. [3] have collected a new public video dataset from social media channels to expand the dataset on stereotyped movements. They endeavored to extract motion features in videos of children with autism, employing descriptors such as HOF (Histogram of Oriented Optical Flow) and HOG (Histogram of Oriented Gradients). Subsequently, machine learning classifiers, such as MLP, GNB, and SVM were employed for classification. Moreover, Zhang et al. [26] combined handcrafted features with learned features from deep learning models. This involved integrating Discrete Wavelet Transform (DWT) with dense trajectory models and pre-trained dual-stream CNN-RNN models. Such integration serves to address the limitations of handcrafted features in action modeling. However, it is important to mention that this approach may not capture all the intricate patterns in video data, particularly concerning stereotyped movements in children with ASD, where the differences between classes are relatively small.

2) Static Image Feature Aggregation Methods: These approaches [27], [28] view a video as a sequence of RGB images and employ CNNs to extract static image features. These features are subsequently connected over time using Long Short-Term Memory (LSTM) networks. However, this method is contingent on high-quality datasets, as the quality of the dataset's quality directly impacts the accuracy of feature extraction. Moreover, this method tends to exhibit lower robustness with complex or changing environments.

3) 3D Convolutional Methods: These methods [29] can simultaneously extract spatiotemporal features from videos in horizontal, vertical, and temporal directions. This allows for a more natural integration of multi-scale features. The O-GAD network [7] consists of a 3D ConvNet time feature extractor and a temporal pyramid network; it is employed to detect ASD actions and distinguish among repetitive behaviors.

4) Dual-stream Methods: Inspired from the dual visual information processing channels in the human brain, dual-stream networks typically consist of two branches: one concentrates on static frames to capture appearance features, utilizing input from RGB image sequences, while the other takes in optical flow data to capture motion information. This dual approach facilitates the fusion and complementarity of appearance features and motion information [30]. However, scenarios involving stereotyped movements in children with ASD often present complexity. RGB data is susceptible to factors such as lighting, background, and occlusions [13]–[15], potentially diminishing the accuracy of recognition. Simultaneously, optical flow data, while adept at capturing motion, can often contain a considerable amount of noise [31]. Therefore, relying solely on single-modal information may prove insufficient for accurate recognition of stereotyped movements in children with ASD. The integration of multi-modal information becomes imperative for effective differentiation.

5) Self-supervised Learning Methods: With the advancement of video understanding, current supervised learning methods require increasingly large amounts of annotated data, which costs enormous human effort and time. Meanwhile, a vast amount of unlabelled video data is easily available on the internet. Therefore, some studies have focused on leveraging unlabelled data through self-supervised learning to enhance model generalization capabilities and reduce reliance on annotated data. Among these methods, contrastive learning [32] [33] [34] and masked image/video modeling [35] [36] [37] are the main research directions.

Dave et al. [32] proposed a temporal contrastive learning framework (TCLR), which introduces local-local and global-local temporal contrastive losses to significantly enhance video representation performance in downstream tasks such as action recognition. Tian et al. [33] introduced the CLSA contrastive learning framework, which synthesizes hard negative samples and employs a selective global aggregation module, achieving state-of-the-art performance in video rescaling tasks. Subsequently, Tian et al. [34] further proposed a hybrid coding framework for low-bitrate video understanding, combining traditional encoders with neural networks to enable efficient understanding and optimized encoding of low-bitrate videos.

He et al. [35] proposed Masked Autoencoders (MAE), which achieve efficient self-supervised visual representation learning by masking a large proportion of input images and reconstructing the masked parts. Building on this, Tong et al. [36] proposed VideoMAE, which utilizes an extremely high masking ratio to enhance the efficiency of video self-supervised pre-training. Finally, Tian et al. [37] introduced a non-semantic information suppression mask learning method, which uses masking and self-supervised learning to reduce redundant non-semantic information in videos, thereby improving video semantic compression efficiency and performance across multiple downstream tasks.

B. Skeleton-based Stereotyped Movement Recognition

Skeleton-based action recognition methods have demonstrated exceptional performance in the realm of action recognition, primarily owing to their stable, accurate motion information and resilience against lighting and background variations. These approaches commonly employ pose estimation algorithms to derive low-dimensional skeleton data from RGB videos. This data, characterized by its robustness to background noise, encapsulates essential spatial and temporal information necessary for effective action recognition. Numerous methods exist to infer the skeleton information of children with ASD from images. The obtained skeleton information is then fed into a neural network to predict the category of movements. Finally, these methods can be divided into handcrafted feature methods, 3D CNN methods, RNN-type methods, and graph convolution methods.

1) Handcrafted Feature methods: Features related to movements are extracted from the raw data in these methods. For instance, Negin et al. [3] used the AlphaPose algorithm [38] to extract skeleton joint information from RGB images. To do so, they applied HOG, HOF, and HOG-HOF combination, as well as SIFT, and SURF techniques to recognize the stereotyped movements. For example, Wang et al. [39] employed 3D joint position features and Local Occupancy Patterns (LOP) to describe human activities. However, the quality of the video affects the accuracy of pose estimation, thereby influencing stereotyped movements recognition.

2) 3D CNN Methods: 3D CNN methods are primarily designed to address challenges of feature extraction in both the spatial and temporal domains by applying deep convolutional structures. For instance, Kim et al. [40] redesigned the TCN model, introducing residual connections to improve both the model's explainability and recognition capability. This modification yielded excellent results, particularly on the NTU-RGB+D dataset. Moreover, Liu et al. [41] utilized a dual-stream 3D CNN to map skeleton joints into a 3D coordinate space. To do so, they encoded spatial and temporal information separately, contributing to enhance spatiotemporal features. In addition, Cao et al. [42] introduced a more effective and robust Joints-pooled 3D Deep-convolutional Descriptor (JDD). They proposed a two-stream bilinear model that can learn guidance from the body joints and capture the spatiotemporal features simultaneously. Despite this advancement, these methods have shortcomings in either spatial or temporal perception.

3) RNN-type Methods: RNN-type methods mainly employ recurrent structures to capture sequential features of action, addressing the challenges posed by dynamic changes and non-linear patterns in continuous skeleton frames [43], [44]. The Part-aware LSTM model [45] divides human actions into different body parts and establishes memory units within each part. This allows the network to independently learn the long-term patterns of each part, enhancing its ability to capture actions features. For instance, Zhang et al. [6] used the OpenPose [46] algorithm to generate the initial skeleton data from autistic children's videos. After eliminating skeleton data noise, stereotyped movements were identified through LSTM.

4) Graph Convolution Methods: Graph convolution methods capture the topological relationships of skeletons to alleviate the shortcomings of traditional methods in capturing skeleton spatial relationships. Moreover, ST-GCN [47] introduced joint correlations and effectively used skeleton topological information, showing good performance on datasets like NTU-RGB+D [45]. 2s-AGCN [48] employed second-order information (bone length and direction), which has more discriminative power and information content for the task of action recognition, thereby improving classification accuracy. However, for video action recognition, this method lacks the capacity to model the long-term temporal dependency of the entire video [49].

Considering the similarity of stereotyped movements among classes , models should make full use of skeleton information to achieve improvements. Although skeleton information can provide stable and accurate motion information, unaffected by lighting and background to a large degree, it lacks surface information, making it difficult to differentiate some similar actions, especially when the inter-class differences in stereotyped movements are small, leading to potential confusion.

C. Multi-modality Fusion in Action Recognition

Behavior can be described through various modalities such as RGB, depth, sound, optical flow, and skeleton information.Utilizing multi-modal information enables a more comprehensive and complete describe representation of the data source. Stereotyped movements are frequently influenced by various factors in the scene, such as interference from other objects or people [3]. Therefore, there is a need for more robust methods to recognize stereotyped movements in adversarial environments.

In video-based action recognition, dual-stream methods adopt multi-modality fusion, integrating predictions based on multiple modalities to enhance recognition accuracy. For instance, Simonyan et al. [30] employed two modalities for fusion: the spatial modality (appearance information from static video frames) and the temporal modality (optical flow information). Moreover, Feichtenhofer et al. [50] introduced a new ConvNet architecture using appearance and optical flow information for spatiotemporal fusion of video clips. Yet, Carreira et al. [51] fed feature vectors from the RGB and optical flow modalities into two separate I3D models, averaging their predictions for the final output.

A different approach [49] was trained with four types of modal information: single-channel RGB images, stacked RGB difference images, stacked optical flow fields, and stacked warped optical flow fields. For instance, Ali et al. [16] employed I3D models to fuse RGB and optical flow information in both early and late fusion stages, verifying that the late fusion performs better than the early one in recognizing ASD stereotyped movements. However, the improvement from these modal fusion methods is not significant.

Moreover, video and skeleton multi-modal fusion has been proven to be effective. For instance, Du et al. [39] enhanced action recognition accuracy by complementing skeleton information with optical flow data at the joint points. In addition, Khaire et al. [52] transformed RGB, depth, and skeleton data into Motion History Image (MHI), Depth Motion Maps (DMMs), and skeleton images, respectively. They trained these methods via five Convolutional Neural Networks (5-CNNs) and eventually used a Weighted Product Model (WPM) for decision-level fusion. Furthermore, Song et al. [53] fused joint data, RGB images, and optical flow data to design a skeleton index transformation layer to automatically extract visual features around key joints. This multi-modal information complemented each other and performed excellently on the NTU RGB+D dataset.

Because the differences between classes of stereotyped movements in children with ASD are small, more modalities are required for differentiation. Therefore, in the task of detecting stereotyped movements, video and skeleton multi-modal fusion can reduce data uncertainty and noise interference, thereby enhancing classification stability and robustness.

III. Problem Formulation

For video-based ASD stereotyped movements recognition, current approaches mainly focus on employing video frames, optical flow images, or skeleton-based methods. However, the method proposed in this paper is designed to effectively mine the pattern of the stereotyped movements in ASD children, yielding in performing multi-channel fusion.

A. RGB-based and Optical Flow-based Methods

Given a set of action videos of ASD children, $V_{\text{RGB}} = \{v_1, v_2, \ldots, v_N\}$ where $v_i \in \mathbb{R}^{T\times H\times W\times 3}$ representing a

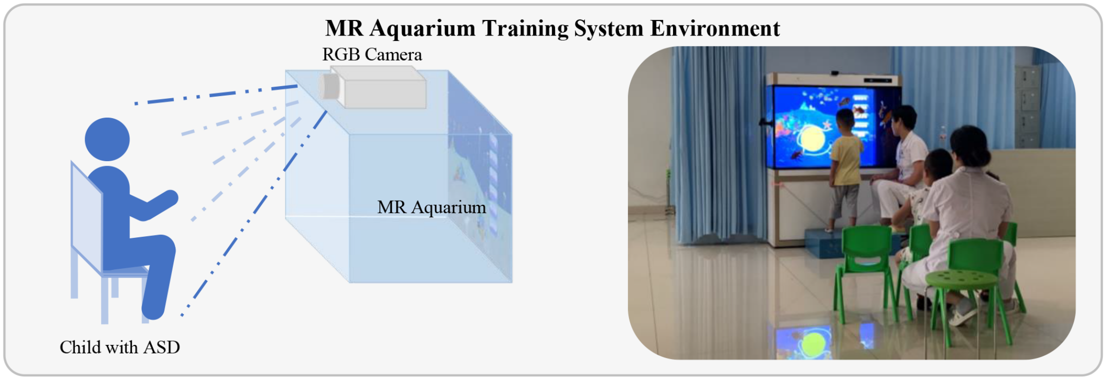


Fig. 2. Setup and implementation of the MR Aquarium Training System used for collecting video data of children with ASD.

video, T represents the video duration, H and W represent the height and width of each frame, respectively, and the number "3" represents the channel count. Setting the labels $L = \{l_1, l_2, \ldots, l_N\}$ and their corresponding relationships (v, l) where $v \in V_{\text{RGB}}$ and $l \in L$, a deep model, Φ, is utilized to transform the frame sequence of each video v_i into a fixed-length feature vector $f_i = \Phi(v_i)$. These feature vectors are then input into a classifier F to determine the relationship between feature vectors and action categories, such that $F(f) = l$ where $l \in L$. The optical flow information extracted from the RGB video stream is represented as $O_{\text{optical}} = \{o_1, o_2, \ldots, o_N\}$, where $o_i \in \mathbb{R}^{T\times H\times W\times 2}$ signifies the optical flow data within a video. Each piece of optical flow information comprises T frames of size $H \times W$, with each frame containing two channels representing the horizontal and vertical movements. After serializing the optical flow information, the optical flow frame sequence is transformed into a feature vector $f_i = \Phi(o_i)$, which is then fed into the classifier F to determine the action categories, resulting in $F(f) = l$ where $l \in L$.

B. Skeleton-based Methods

Skeleton-based methods derive joint streams $J = \{j_1, j_2, \ldots, j_N\}$ from RGB video frames, where $j_i \in \mathbb{R}^{T\times N\times X\times Y}$. Similarly, bone streams $B = \{b_1, b_2, \ldots, b_N\}$ are obtained, with $b_i \in \mathbb{R}^{T\times N\times X\times Y}$. Here, T denotes time steps, and N represents the number of skeletons and joints at each time step, while X and Y correspond to the horizontal and vertical coordinates in the 2D space. For joint data j_i, each joint point $j_{i,m}$ (where $m = 1, 2, \ldots, N$) represents the m^{th} joint in the i^{th} data and can be viewed as a node v_m in the graph $G = (V, E)$, where V denotes the set of nodes representing all joints, and E is a set of edges defined by physical connections or other criteria between joints.

The graph G is then fed into the GCN for feature extraction, producing features f_b and f_j from the skeleton and node data, respectively. These features are independently processed by two separate softmax layers, yielding P_{joint} and P_{bone}, which are subsequently combined to produce the final skeletal prediction P_{skeleton}.

C. Vision-Skeleton Multi-modality Fusion Methods

In the vision-skeleton fusion methods, the RGB video v and optical flow o streams are processed through the VML module to learn both visual spatial and motion information, resulting in the extraction of visual feature f_{RGB} and f_{Flow}.

The SRM module processes both joint and bone streams to extract spatiotemporal features from the skeleton data. A cross-modal spatiotemporal graph $G = (V, E)$ is constructed for a skeleton sequence with N joints and N bones over T frames. This graph contains $2N$ nodes, where each node represents either a joint or a bone, allowing the integration of both motion types. The vertex set $V = \{v_{ti}^{\text{Joint}}, v_{ti}^{\text{Bone}} \mid t = 1, \ldots, T, i = 1, \ldots, N\}$ represents each joint and bone across frames.

The edge set E comprises two types of connections: E_S, which defines intra-frame connections based on the natural links within the skeleton structure (joint-to-joint, bone-to-bone, joint-to-bone), and E_F, which captures inter-frame temporal dependencies by linking the same joint or bone across consecutive frames. The graph G is fed into GCN to extract skeleton movement pattern f_{Skeleton}.

To effectively fuse features from different modalities, we apply a cross-modality attention mechanism to the extracted f_{RGB}, f_{Flow} and f_{Skeleton}. Each modality's feature is projected into a shared feature space and then fused using cross-attention to produce a unified representation $\mathbf{O}$. This fused representation is then passed through a classification layer to obtain the final prediction P.

IV. Dataset

A. Construction of the ACSA653 Dataset

1) The Environment of Dataset Acquisition: The data acquisition environment for the ACSA653 dataset is the MR aquarium training environment [19] (e.g., Figure 2). In this environment, children with ASD can participate in ASD training courses and interact with virtual creatures through multi-touch for cognitive training.

A camera is used to capture RGB video data directly above the aquarium. Throughout the training, it points at the performance of the child records as he is facing directly the camera and spends most of his time sitting down. During the training, a therapist will be standing behind the child. As a result, a part of the video will show the therapist and the child together, with the therapist's body partially visible. Notably, several patient subjects are present during the training sessions.

2) Data Collection and Quality Control: After five months of acquisition, a total of 1.78 TB of video data with a total

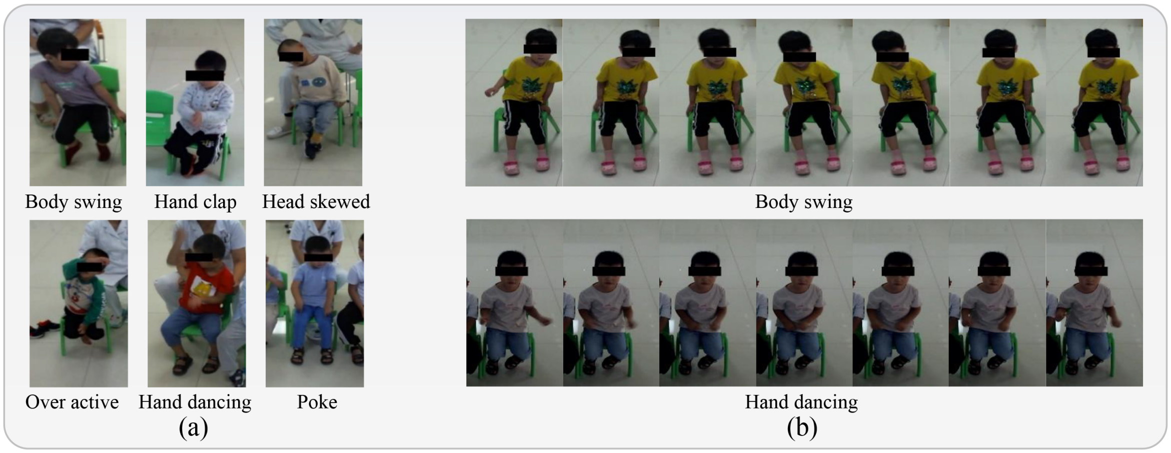


Fig. 3. Frames of ASD Children's Stereotyped Movements Videos where (a) represents the single static RGB frames of six classes of stereotyped movements and (b) denotes the sequential multiple static RGB frames of six classes of stereotyped movements.

length of about 360 hours, a width of 1600 pixels, and a height of 896 pixels is acquired. Moreover, data is collected from six children with ASD who are trained in the MR aquarium training environment.

It is important to mention that collecting unguided behavior data from ASD children in a natural MR rehabilitation training setting ensures more authentic, representative, and low-bias data, thereby enhancing the understanding of the behavioral characteristics of children with ASD and the effectiveness of intervention strategies.

3) Video Editing and Expert-Guided Data Annotation: Five doctors were involved in the data annotation process as it was achieved manually. The doctors were asked to label the stereotyped movements of all six children with ASD in the videos, recording the start and end times and the category of the stereotyped movements.

Consequently, the videos were edited and cropped based on the doctors' annotations. The process involved two main steps:

1) Cropping the videos to ensure that the annotated subjects were centered and fully visible. To achieve a uniform format, the cropping dimensions were set to 400 pixels in width and 800 pixels in height.
2) Editing the videos according to the recorded start and end times of the stereotyped movements. The videos acquired from this editing and cropping process were then categorized and saved according to the annotated types of stereotyped movements.

Ultimately, 653 video clips of stereotyped movements were acquired in this study; some frames of the videos are displayed in Figure 3. These video clips are labeled based on the following six categories of stereotyped movements: body swing, hand clap, head skewed, over active, hand dancing, and poke. As a result, 122 videos related to body swing, 45 videos related to hand clap, 86 videos related to head skewed, 218 videos related to over active, 95 videos related to hand dancing, and 87 videos related to poke were identified.

4) Comparison with Other Datasets: We conducted a multidimensional comparison between the proposed ACSA653 dataset and existing datasets. Unlike the SSBD dataset, which relies on publicly available YouTube videos, the ACSA653 dataset was collected during the MR training rehabilitation process of children with ASD. This process was imperceptible to the children, ensuring the natural authenticity of the data and minimizing potential biases. Compared to other datasets, ACSA653 has advantages in terms of the number of samples, action classes, and modalities (Referring to Table I). This expanded sample size not only enhances the dataset's richness but also facilitates a more in-depth exploration of ASD stereotyped movements.

B. Extract Skeleton Information

Extracting skeleton features through human pose estimation is an important step in bone behavior recognition. In a study conducted by Duan et al. [54], they have found that a 2D pose derived from a lightweight backbone network outperforms any 3D human pose source in action recognition tasks. As the cropped image still contains parts of the body of other objects, except for the labeled object, it is probably not suitable to use the Single-Person Pose Estimation (**SPPE**) method. While trying and visualizing the Single-Person Pose Estimation (**SPPE**) and Multi-Person Pose Estimation (**MPPE**) results, as shown in Figure 4(a), it is evident that single-person pose estimation is difficult to extract accurate skeleton information when applied on ACSA653. Moreover, as **SPPE** methods focus on single person, their effectiveness decrease when there is more than one person in the video. However, **MPPE** can extract the multi-person accurate skeleton information in such situation. Therefore, 2D **MPPE** method was applied to extract the skeleton information throughout this work.

To extract skeleton information from a single ASD child in the video, two initial pre-processing steps are required: pose estimation and human objects tracking. The full details are provided in sections IV-B.1 and IV-B.2

1) Pose Estimation: To achieve accurate and high-quality human pose estimation, this study utilizes AlphaPose [38], which employs the FastPose algorithm known for its strong performance on the MSCOCO dataset [55]. When applied to video data, AlphaPose performs frame-by-frame pose estimation, extracting 17 key skeletal points. These key points, as labeled in the model trained on the MSCOCO dataset, include: 0 - nose, 1 - left eye, 2 - right eye, 3 - left ear, 4 - right ear, 5 - left shoulder, 6 - right shoulder, 7 - left elbow, 8 - right

TABLE I
COMPARISON BETWEEN ACSA DATASET AND SOME OF THE OTHER DATASETS FOR ASD STEREOTYPED MOVEMENTS RECOGNITION.

Dataset	Video	Classes	Data Modalities	Year
SSBD [21]	75	3	RGB	2013
M. Jazouli et.al. [22]	50	5	RGB	2016
ASD40h [7]	30	5	RGB	2019
ESBD [3]	141	4	RGB	2021
Activis [16]	388	5	RGB	2021
Updated SSBD [23]	61	3	RGB	2023
ACSA653	**653**	**6**	**RGB+Skeleton**	2023

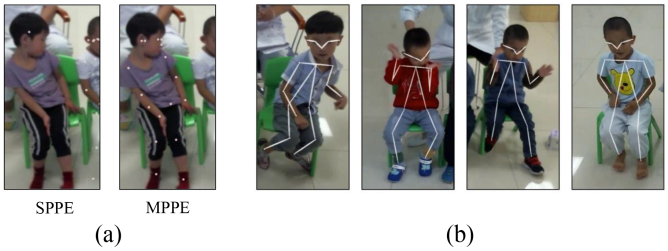


Fig. 4. (a) Visualization of Pose Estimation Methods. SPPE means single-person pose estimation method, and MPPE means multi-person pose estimation. (b) Visualization of pose estimation by AlphaPose.

elbow, 9 - left wrist, 10 - right wrist, 11 - left hip, 12 - right hip, 13 - left knee, 14 - right knee, 15 - left ankle, and 16 - right ankle. The results of the pose estimation are illustrated in Figure 4(b).

2) Human Objects Tracking: When implementing a multi-person pose estimation method, a video may yield predictions for multiple individuals. To address this issue, PoseFlow [56] is employed for continuous human objects tracking with the video. This step ensures that only data corresponding to labeled objects is retained, allowing the results of multi-person pose estimation to align with the respective object. Subsequently, the pose estimation results for labeled objects are saved in accordance with their assigned IDs.

V. METHOD

As depicted in Figure 5, APMFNet contains three modules:

1) The **Visual Motion Learning** (VML) module extracts spatial and motion information from RGB and optical-flow sequences.
2) The **Skeleton Relationship Mining** (SRM) module extracts skeleton pattern information from the joint and bone stream, and constructs them into a cross-modal graph. SRM effectively captures key motion patterns related to stereotyped movements.
3) Finally, the **Multi-channel Fusion** (MF) module fuses the visual information with the skeleton information to ultimately predict stereotyped movement recognition result P.

A. Spatial and Temporal Information Extraction in Visual Motion Learning

The Visual Motion Learning Module (VML) aims to extract visual and spatial features from RGB frames, providing a comprehensive scene description that captures both dynamic and static information. Considering the nuances of stereotyped movements, it is crucial to obtain both spatial details and temporal precision to understand the exact location and level of change in the child's body. To achieve this, SlowFast [57] is employed, offering a dual-stream model that integrates: (1) a slow pathway with high spatial detail but low temporal resolution, (2) a fast pathway with high temporal resolution but less spatial detail, and (3) a convolutional layer that fuses information from the fast pathway into the slow pathway.

Identifying the complexity of real-world motion patterns, capturing these patterns can't be accomplished just by applying RGB frames; therefore, the VML module uses the TV-L1 algorithm [51] to extract optical flow information from sequences, thereby representing dynamic motion. For each type of inputs (RGB frames V, Optical flow O), VML extract features from both Slow and Fast pathways: F_{xs}, F_{xf} ,the two features are concated as $F_x = [F_{xs}, F_{xf}]$ where x can be either RGB (denoted as RGB) or optical flow (denoted as $flow$).

Each set of features feeds to a Fully Connected (FC) layer and use *softmax* to obtain probability scores for each class.

$$P_{RGB} = \text{softmax}(FC(F_{RGB})) \tag{1}$$

$$P_{flow} = \text{softmax}(FC(F_{flow})) \tag{2}$$

The cross-entropy loss is applied to each prediction set for loss calculation. The loss functions for RGB and optical flow inputs are defined as follows:

$$\mathcal{L}_{\text{RGB}} = -\sum_i y_i \log(P_{\text{RGB},i}) \tag{3}$$

$$\mathcal{L}_{\text{flow}} = -\sum_i y_i \log(P_{\text{flow},i}) \tag{4}$$

where y_i represents the ground truth for class i, and $P_{\text{RGB},i}$ and $P_{\text{flow},i}$ denote the predicted probabilities for the RGB and optical flow pathways, respectively.

The VML module fuses static visual information from the RGB frame sequence with dynamic scene variations captured by the optical flow sequence. This not only improves the understanding of visual features in stereotyped movements of ASD children but also offers supplementary context for the SRM module.

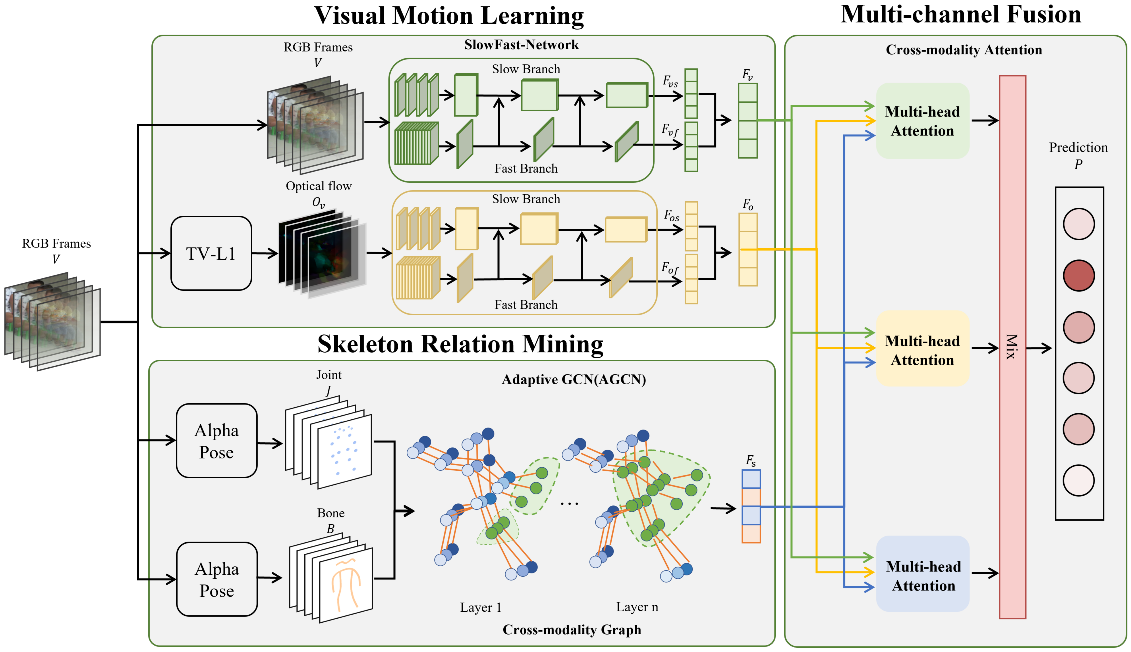


Fig. 5. The schematic diagram of APMFNet: The optical flow $\boldsymbol{O_v}$, joint $\boldsymbol{J}$, and bone $\boldsymbol{B}$ data is extracted from the RGB frames $\boldsymbol{V}$. The Visual module fuses $\boldsymbol{V}$ and $\boldsymbol{O_v}$ modalities, extracting spatial and motion information by SlowFast network; the SRM module uses $\boldsymbol{J}$ and $\boldsymbol{B}$ inputs for key motional pattern mining through cross-modality graph; $\boldsymbol{F_v}$, $\boldsymbol{F_o}$ and $\boldsymbol{F_s}$ are fused in the MF module through cross-modality attention to produce the final predition $\boldsymbol{P}$.

B. Stereotyped Movements Modeling in Skeleton Relationship Mining

The Skeleton Relationship Mining (SRM) module aims to build a relationship graph of human joints J and bones$B$ from skeleton data, effectively mining the patterns of stereotyped movements. This module uses a cross-modal graph to fuse the joint and bone modality for predicting action categories. Joint data provides the positional coordinates of each joint, while Bone data captures the connections between joints, including bone length and direction.

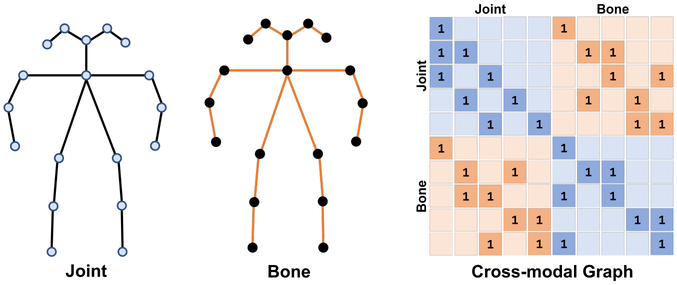
The SRM employs AlphaPose [38] to estimate joint positions and constructs the skeleton graph using joint connections from the MSCOCO dataset [55]. Bone data B is calculated based on the joint sequence J, using connected key points to derive bone vectors, such as $b_{1,2} = (x_2 - x_1, y_2 - y_1)$.

To model the relationships adaptively, the SRM uses three Adaptive Graph Convolutional Network (AGCN) layers with output channels of 64, 128, and 256, respectively. The AGCN layer allows the network to learn associations flexibly during training, thus enhancing the model's ability to understand the complexity of human movement patterns.

To leverage the relationships between the joint and bone modalities, we employed a graph structure to link these two modalities (see in Fig. 6). The adjacency matrix $A_{i,j}$ of the cross-modal graph is defined as:

$$A_{i,j} = \begin{cases} 1 & \text{if } D_{i,j} \text{ and } i < n, j < n \\ 1 & \text{if } i < n, j \geq n \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Where $D_{i,j}$ indicates that there is a relation between the two nodes (joint-joint, joint-bone, or bone-bone) in the dependency tree of the human skeleton mapping.



	Joint					Bone				
Joint	1					1				
	1	1					1	1		
	1		1					1		1
		1		1			1		1	
			1		1				1	1
Bone	1					1				
		1		1			1	1		
		1	1			1		1		
				1	1				1	1
			1		1	1				1

Fig. 6. Cross-modal graph linking joint and bone modalities for human motion representation.

The SRM module take both the joint J and bone sequence B as input noted f_{in}. Suppose the input skeleton information sequence data f_{in} has a size of $C \times T \times N$, where N is the number of key points, T represents the length of time, and C denotes the number of channels for the key points. In ST-GCN, the graph convolution operation is formulated as follows:

$$f_{\text{out}} = \sum_{k}^{K_v} W_k (f_{in} A_k) \odot M_k \tag{6}$$

where K_v denotes the size of the spatial dimensional convolution kernel, usually set to 3, A_k determines whether there is a connection, $\Lambda_k^{ii} = \sum_j \bar{A}_k^{ij} + \alpha$ represents the degree matrix, $\bar{A}_k$ is the $N \times N$ adjacency matrix, and element $\bar{A}_k^{ij}$ denotes the connection between vertices v_i and v_j, Moreover, W_k indicates the weight vector of size $C_{\text{out}} \times C_{\text{in}} \times 1 \times 1$, M_k represents the attention mask, determining the strength of the connection, and, finally,$\odot$ denotes the dot product operation.

In Eq. 6, the graph topological information is determined by the adjacency matrix A_k and the attention mask M_k. To make the graph topological information adaptive, Shi et al. [48] proposed the following equation:

$$f_{\text{out}} = \sum_{k}^{K_v} W_k f_{in}(A_k + B_k + C_k) \tag{7}$$

where A_k is consistent with that in Eq. 6 and represents the adjacency matrix information, B_k starts with the same initial value as A_k, but it is updated as a trainable parameter during the training process, and C_k determines the presence and strength of the connection between any both points, similar to the attention mechanisms. The dot product of the vectors is commonly used to calculate the similarity between Query and Key, the *softmax* function is used to normalize this similarity.

$$Query_k = W_{\theta k} f_{in} \tag{8}$$

$$Key_k = W_{\phi k} f_{in} \tag{9}$$

$$C_k = \text{softmax}(Query_k^T Key^k) \tag{10}$$

In Eq. 7, we define the operation of adaptive graph convolution. The Joint and Bone streams undergo three adaptive graph convolution operations, resulting in feature F_s. F_s is then passed through a fully connected layer and activated using the *softmax* function to get the predicted outcomes P_s defined as follows:

$$P_s = \text{softmax}(FC(F_s)) \tag{11}$$

Moreover, the cross-entropy loss is applied to prediction set for loss calculation. Therefore, the loss functions for the joint and bone inputs are defined as follows:

$$\mathcal{L}_{skeleton} = -\sum_i y_i \log(P_{s,i}) \tag{12}$$

where y_i is the ground truth label for the i-th class, and $P_{s,i}$ represents the predicted probabilities for the skeleton streams.

Finally, the predictions from the two streams, P_j and P_b, are added to derive the fused prediction $P_{skeleton}$, consequently used to predict the label of stereotyped movements.

Information acquired from joints mainly provides the positions of various parts of the human body, whereas bones describe the connections with these joints. This module extracts both skeleton joint and bone pattern information, capturing motion features comprehensively. Compared to the visual information, the information extracted from skeleton joints is more robust; moreover, the interrelations between joint and bone provide key information for mining the pattern of ASD stereotyped movements in complex background conditions.

C. Multi-Channel Fusion

Visual information primarily derived from RGB frames and optical flow data, provides detailed context and scene specifics. However, this information can be compromised by background noise or may struggle to accurately identify specific actions in complex scenarios. Conversely, skeleton information offers structural and action-related insights about the human body but may lack sufficient context to explain the reasons or intentions behind these actions.

Different modalities (M) provide various perspectives and information about the same object. Skeleton information captures the structure and movements of the human body, while visual information adds contextual background. To effectively integrate this information, in the MF module, we map the features of each modality into a shared feature space and then use a cross-attention mechanism to fuse these features. For modality M_i, its input feature is represented as: $\mathbf{X}_{M_i} \in \mathbb{R}^{B \times d_{M_i}}$, where B is the batch size and d_{M_i} is the feature dimension.

The features of each modality are mapped to a unified dimension d: $\mathbf{H}_{M_i} = \mathbf{X}_{M_i}\mathbf{P}_{M_i} \in \mathbb{R}^{B \times d}$, where $\mathbf{P}_{M_i} \in \mathbb{R}^{d_{M_i} \times d}$ is the projection matrix.

For each modality M_i, the query vector $\mathbf{Q}_{M_i}$ is computed, and key and value vectors are prepared for other modalities:

$$\mathbf{Q}_{M_i} = \mathbf{H}_{M_i}\mathbf{W}_q^{M_i} \in \mathbb{R}^{B \times d} \tag{13}$$

$$\mathbf{K}_{M_j} = \mathbf{H}_{M_j}\mathbf{W}_k^{M_j} \in \mathbb{R}^{B \times d} \tag{14}$$

$$\mathbf{V}_{M_j} = \mathbf{H}_{M_j}\mathbf{W}_v^{M_j} \in \mathbb{R}^{B \times d} \tag{15}$$

The query, key, and value vectors are split into h heads. For each head $k = 1, \ldots, h$:

$$\mathbf{A}_{M_i \to M_j}^{(k)} = \text{softmax}\left(\frac{\mathbf{Q}_{M_i}^{(k)}(\mathbf{K}_{M_j}^{(k)})^\top}{\sqrt{d_k}}\right)\mathbf{V}_{M_j}^{(k)} \tag{16}$$

The outputs of all heads are concatenated:

$$\mathbf{A}_{M_i \to M_j} = \text{Concat}\left(\mathbf{A}_{M_i \to M_j}^{(1)}, \ldots, \mathbf{A}_{M_i \to M_j}^{(h)}\right) \tag{17}$$

The outputs from all modalities are collected:

$$\mathbf{A}_{\text{all}} = \text{Concat}\left(\mathbf{A}_{M_i \to M_j} \;\middle|\; M_i, M_j \in M,\ i \neq j\right) \tag{18}$$

The concatenated cross-attention outputs are passed through a linear layer to obtain the fused feature representation:

$$\mathbf{O} = \mathbf{A}_{\text{all}}\mathbf{W}_{\text{out}} \in \mathbb{R}^{B \times d} \tag{19}$$

Moreover, to address class imbalance, the focal loss [58] is applied to the prediction set for loss calculation. Therefore, the loss functions for the joint and bone inputs are redefined using focal loss as follows:

$$\mathcal{L}_{\text{focal}} = -\sum_i \alpha_{y_i} y_i \left(1 - P_i\right)^\gamma \log\left(P_i\right) \tag{20}$$

where y_i is the ground truth label for the i-th class, P_i represents the predicted probability for the i-th class. γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted. α_{y_i} is the weighting factor for class y_i, used to address class imbalance.

This cross-modality attention enables each modality to focus on others, effectively capturing complementary information and improving the model's robustness across various scenarios.

D. Training Strategy

We employed a late fusion strategy by independently training each modality and then using cross-modality attention to fuse the extracted features, as shown in Algorithm 1.

Algorithm 1 APMFNet Optimization

```
Input:
    $V = \{V^{(i)} | i = 1, \dots, N\}$: RGB videos
    $O = \{O^{(i)} | i = 1, \dots, N\}$: Optical flow videos
    $J = \{J^{(i)} | i = 1, \dots, N\}$: Skeleton joints
    $B = \{B^{(i)} | i = 1, \dots, N\}$: Skeleton bones
Procedure:
1: Train $G_v$ with RGB videos $V$ using $\mathcal{L}_{RGB}$ as constraint.
2: Train $G_o$ with optical flow videos $O$ using $\mathcal{L}_{flow}$ as constraint.
3: Train $G_s$ with skeleton joints $J$ and bones $B$ using $\mathcal{L}_{skeleton}$ as constraint.
4: Train $G_{Att}$ with the output features $F_v$, $F_o$ and $F_s$ using $\mathcal{L}_{focal}$ as constraint.
Output:
    Trained modules $G_v$, $G_o$, $G_s$, and $G_{Att}$ in APMFNet.
```

VI. Experiment

A. Dataset

The stereotyped movements dataset ACSA653, presented in details in Section IV-A, is used in the experiments, and the details of this dataset can be found above. The dataset is acquired in a rehabilitation MR training environment for children with ASD. To sum up, a large number of videos were was acquired from the frontal side of the children, and a total of 653 video clips of uniform specifications and high quality of stereotyped movements in children with ASD were obtained by manual cropping and editing. This involves the use of six types of stereotyped movements. To conclude, Table II describes the overview of the dataset.

Using the Scikit-learn package, the dataset is randomly divided into proportions of 70% for the training set and 30% for the test set. Acknowledging the problem of an uneven distribution of sample labels in the dataset, SMOTE oversampling [59] is applied to the training set in this research to generate training samples with increased data volume and a balanced distribution.

B. Details of Implementation

Model building is implemented using PyTorch [60], a model training is performed on a GTX2080 SUPER graphics card, the applied optimization strategy for training is Adam [61], the learning rate decay strategy is an exponential, the loss function is represented by the cross-entropy function, and the Dropout ratio is set to 0.5. In this study, the Optuna framework [62] is applied to optimize the learning rate and the batch size in training automatically.

To select the model of the video stream, ResNet-50 was used for 3D convolution in both slow path and fast path configurations [63]. Concerning the parameter settings, τ, α, and β values are set to 8, 8, and 1/8, respectively. The size of the input video is scaled to 88 pixels in width and 176 pixels in height. In addition, the video stream model is also pre-trained on the Kinetics400 dataset, and then a fully connected layer is added to the model generated by pre-training. In this context, the input represents the number of classifications (400) from Kinetics400 and the output indicates the number of classifications (6) from ACSA653. Therefore, the learning rate ranges from 1×10^{-6} to 5×10^{-4}, and the batch size range is between 3 and 6.

Concerning the skeleton stream, joint and bone streams use the same structured network with three AGCN layers stacked in the network, each having 64, 128, and 256 output channels. The sample step length of the input video is set to 3 whereas the topological connections of the human skeleton are initialized according to key points in MSCOCO properties. They will lead to the calculation of vector data and the generation of adjacency matrix. In the training phase, the network is first pre-trained using the Kinetics400 dataset, and the learning rate varies in the range of $(1 \times 10^{-4}\ , 1 \times 10^{-1})$ with a batch size range between 4 and 80.

C. Evaluation of Model Performance

The accuracy is used to measure the trials, responding to the ability of the model to correctly recognize the stereotyped movement in the video. The formula is defined as follows:

$$Accuracy = (TP + TN)\,/ALL$$

where TP is the true positive, TN represents the true negative, and ALL indicates all outcomes.

D. Performance Comparison

We have compared the performance of deep learning algorithms with different input types. The inputs include RGB, Joint, and Multi-Channel. The experimental results are displayed in the Table III, where we have following observations:

- **The dual-stream design of the SlowFast model significantly enhances video analysis capability, particularly in capturing spatiotemporal features.** The SlowFast outperforms ResNet50+LSTM and C3D. This suggests that the dual-stream structure of SlowFast captures dense and sparse temporal information efficiently, thereby capturing key dynamics in videos.
- **The graph-based method can more effectively model human body joints relationship.** When considering the joint inputs, STGCN outperforms BiLSTM and CNN+LSTM methods. This might be attributed to the graph structure created by STGCN that efficiently captures spatial relationships and dynamic changes between joints, helping in recognizing stereotyped movements.
- **APMFNet makes significant improvements on ACSA653, outperforming 2S-AGCN by 9.18% and SlowFast by 3.57%.** These results indicate that APMFNet effectively fuses global prediction information from the vision, joint, and skeleton.
- **Inter-modal fusion generally increases accuracy, however the gains are depending on the specific modalities**

TABLE II
OVERVIEW OF ACSA653.

Dataset	Participants	Action categories	Number of each action videos						#Videos
			Body swing	Hand clap	Head skewed	Over active	Hand dancing	Poke	
ACSA653	6	6	122	45	86	218	98	87	653

TABLE III
PERFORMANCE COMPARISON OF VARIOUS METHODS SELECT TOP-1 ACCURACY AS A MEASURE.

Data	Algorithms	Accuracy	body swing	hands clap	hands dancing	head skewed	over active	poke
RGB	ResNet50+LSTM	0.5255	0.4286	0.2727	0.4286	0.3600	0.6618	0.6552
	C3D	0.6020	0.4286	0.3636	0.4643	0.6000	0.8088	0.5517
	SlowFast	0.7653	0.7143	0.5455	0.7143	0.8400	0.8235	0.7586
Joint	BiLSTM	0.5153	0.5143	0.4545	0.4643	0.4800	0.5735	0.4828
	CNN+LSTM	0.648	0.6000	0.5455	0.6071	0.5600	0.7500	0.6207
	STGCN	0.6327	0.6897	0.6667	0.6364	0.6786	0.5217	0.7576
Multi-Channel	2S-AGCN	0.7551	0.8000	0.6364	0.6786	0.6000	0.8382	0.7586
	SlowFast	0.8112	0.8286	0.7273	0.7500	0.8000	0.8382	0.8276
	APMFNet	**0.8571**	**0.8857**	**0.8182**	**0.7600**	**0.8824**	**0.8966**	**0.8621**

and their complementarity. SlowFast with RGB and flow as inputs performs better than SlowFast with RGB, but performs 4% worse in the third class where the flow modality does not contribute to additional useful information and might even interfere with the learning.

The outstanding performance of APMFNet can be attributed to the innovative integration of multi-modal data and pattern recognition capabilities. Its core strength lies in the effective fusion of visual and skeleton information. The VML module captures complex spatial and motion features, while the SRM module mines skeleton patterns through a detailed key point relationship graph. This integration is further enhanced by the MF module, ensuring a comprehensive data representation.

E. Ablation Study

To analyze the effect of multi-modality fusion methods on stereotyped movement recognition in children with ASD, recognition alone or in combination accuracy was tested with various modalities either. The input for the skeleton stream is estimated using AlphaPose.

The experimental results in Table IV shows:

- **The fusion of RGB and optical flow in the SlowFast network enhances the recognition of stereotyped movements effectively.** For example, in the categories of "body swing" and "hands dancing", the performance of V(R+O) is significantly better than using V(R) or V(O) individually. This suggests that fusing RGB and the optical flow information can consider both the spatiotemporal features and motional features of actions, thereby complementing each other and improving the accuracy of the stereotypical action recognition.
- **The SRM module, fusing joint and skeleton information, offers two different perspectives for skeleton action analysis.** The fusion of the joint modality (J) and skeleton modality (B) is more efficient than using them individually. For instance, in the "hands dancing" category, the performance of V(R+O) is significantly better compared to using V(R) or V(O) separately. Moreover, knowing the position of the wrist without understanding the structure of the entire arm may not be sufficient to identify a "hands dancing" movement. When being combined, the bone modality provides more context.
- **Vision modality and skeleton modality are fused effectively.** Referring to Table IV, when comparing vision modality (V) or skeleton modality (S) individually, the accuracy rate for every category is further improved.

F. Case Study of Action Pattern Learning

In this section, the ability of the SRM module to extract patterns of ASD children's stereotyped movements is evaluated. All classes from the ACSA653 dataset were selected, specifically "Body Swing", "Hands Clap", "Hands Dancing", "Head Skewed", "Over Active", and "Poke". Moreover, three cases were analyzed for each category.

We effectively visualized the model's attention points on the human skeleton connection diagram by Grad-CAM [24] visualization, as shown in Figure 7. The model's attention to joints is indicated by the size of the red circles on them, while the attention to bones is represented by the thickness of the red lines connecting them. Moreover, the visualization results of the three cases are initialized in each category to validate the capability of the SRM module in extracting stereotyped movement patterns. When evaluating the SRM module across

TABLE IV
EXPERIMENTAL RESULTS OF ABLATION STUDY.

Model	Accuracy	body swing	hands clap	hands dancing	head skewed	over active	poke
V(R)	0.7653	0.7143	0.5455	0.7143	0.8400	0.8235	0.7586
V(O)	0.6173	0.6857	0.8182	0.6071	0.3200	0.5735	0.8276
V(R+O)	0.8112	0.8286	0.7273	0.7500	0.8000	0.8382	0.8276
S(J)	0.6786	0.6571	0.5455	0.6071	0.6800	0.7647	0.6207
S(B)	0.6735	0.6571	0.6364	0.5000	0.6400	0.7941	0.6207
S(J+B)	0.7551	0.8000	0.6364	0.6786	0.6000	0.8382	0.7586
V(R+O)+S(J+B)+F	**0.8571**	**0.8857**	**0.8182**	**0.7600**	**0.8824**	**0.8966**	**0.8621**

the six classes, the findings suggest that the SRM module effectively identifies and classifies the stereotyped movement features of children with ASD.

The visualization of joints and bones are shown in the Figure 7, the following observations are derived:

- Specifically, in the "Body Swing" class, the SRM module shows accurate capture of the unique motion patterns of ASD children. For example, in case 1, despite the smaller extent of "body swing", the SRM module accurately focuses on the movements of the head, shoulders, and arms. In other cases, with a larger extent of "body swing", it equally focuses on the movements of the head, shoulders, arms, and torso.
- In cases of "Hands Clap", whether in joint or bone modality, the model focuses its attention on the movements of hands and arms, accurately capturing the key patterns of the "hands clap" movement.
- Similarly, in "Hands Dancing", the attention of model focuses on the hands and arms, demonstrating the effectiveness of mining the patterns of hand-dancing movements.
- For the "Head Skewed" class, the model focuses on the head and torso movements, central to understanding this motion pattern.
- In the "Over Active" class, model focuses on parts where the movement extent is larger, including the legs and arms, aligning with the features of "Over Active".
- In the "Poke" class, even with indirect hand movements, the model consistently concentrates on the hands and arms, effectively mining the "poke" motion patterns.

These findings collectively prove the SRM module's efficiency on pattern mining and accuracy in recognizing the stereotypical movements of ASD children.

G. Error Analysis

Based on the above analysis, we find that each module of the APMFNet plays a pivot role in alleviating the issue of the inadequacy of video and skeleton fusion. In this section, we will delve into the complementary approaches to multimodal fusion through error analysis. We randomly selected stereotyped movements from four ASD children found in the ACSA653 dataset and observed the attention areas of the model on these samples using the Grad-CAM visualization method. Therefore, Figure 8 illustrates the model's attention area variations on random samples from the ASCA653 dataset.

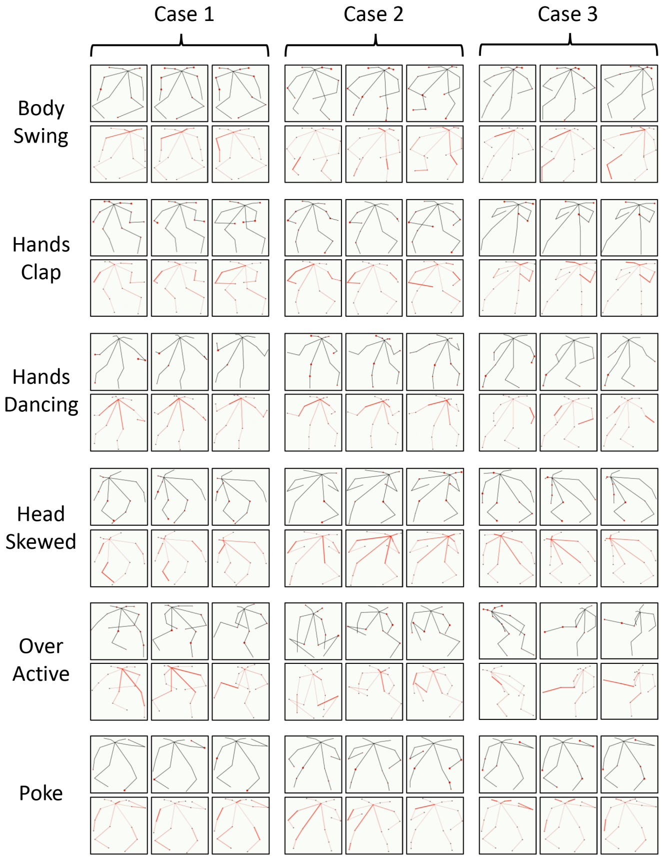


Fig. 7. Visualization of model attention on skeleton modality.

In case 1, both the vision and skeleton modalities were capable of capturing the movement features of ASD children in the "hands dancing" class. When both modalities accurately captured the motion features, it is recognized to the visual data being clear without significant obstructions or noise, and the skeleton data accurately reflecting the dynamic movement of key points. Moreover, modal fusion enhanced the model's robustness in fine-grained reasoning.

In case 2, the vision modality failed to fully focus on the movement features of ASD children, and the visual noise led to a dispersion of attention. In this case, the model's attention was attracted to other children in the background, with the vision modality only partially focusing on the movement features. However, due to the skeleton modality's

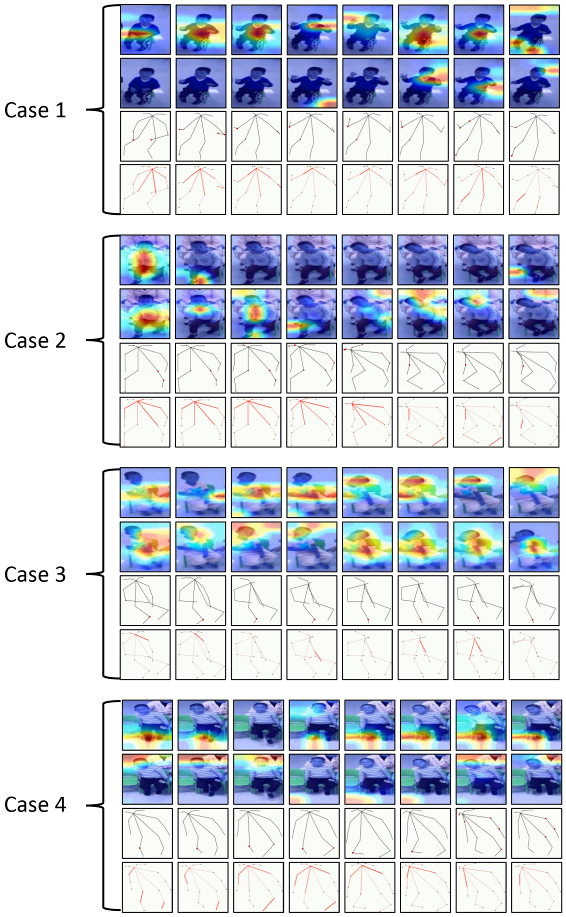


Fig. 8. Visualization of model attention, RGB Features, Optical Flow Features, Joint Features and Bone Features represent the the model attention in different modalities.

insensitivity to background changes, it was still able to provide clear information regarding the main body movements and key point locations, which helps reduce the model's susceptibility to visual noise.

In case 3, the skeleton modality exhibits limited efficacy in modeling the children's motion features, especially when the arms and legs of children with ASD overlapped, the model encountering obstacles in learning the associations between different nodes. Meanwhile, the vision modality successfully captured the children's motion features. In such situations, the fusion of vision and skeleton modalities provided crucial information to resolve the issue of overlapping body parts.

In case 4, neither the vision nor skeleton modalities captured the movement features of ASD children effectively. The vision modality was disrupted by background noise when extracting the child's motion features due to concurrent movement of the ASD child and a doctor in the background, and the skeleton modality is hard to construct a key point relationship graph accurately due to some skeleton nodes being obscured. However, the optical flow and bone modalities captured some movement features, with optical flow capturing the motion features of the child's head, and bone capturing the movement features of the shoulders and arms, representing key patterns of the "Body Swing" movement. Moreover, in complex situations with motion interference, although some modalities were unable to provide effective information, multimodal fusion was still effective in capturing some of the child's movement features, yielding in valuable information for stereotyped movements recognition.

VII. Conclusion

In this study, we constructed an ASD stereotyped movements dataset, called ACSA653. It includes 653 videos across six different classes. ACSA653 provides more samples, classes and modalities than other stereotyped movements datasets.

We designed APMFNet, a model comprising the VML, SRM, and MF modules. APMFNet processes video-based RGB streams, optical flow streams, as well as joint and bone streams derived from skeleton data. The model effectively captures patterns of stereotyped movements, thereby improving the recognition performance of ASD-related movements. By efficiently integrating scene details and contextual information with the human body's structural and motion patterns, the model enhances both predictive accuracy and generalization capabilities. This fusion is particularly important for addressing the challenges posed by complex scenes in diagnosing stereotyped movements in autism spectrum disorder and resolving issues arising from similarities between different classes of movements.

Concerning the future work, we will further attempt to employ more advanced network designs to fuse the modalities at different time stages. We also plan to explore the way to align features effectively through network architectures or training strategies, aiming to optimize the flow of information between different modalities. This will enhance the efficiency and accuracy of stereotyped movement recognition in ASD.

Ethics Statement

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Weihai Maternal and Child Health Care Hospital (No.WFEY-QR-XZ-230). When this study involved human participants, informed consent was received from each individual.

Acknowledgment

This work was partially supported by the Ministry of Education of Humanities and Social Science project (22C10422009), the National Natural Science Foundation of China (6227071327; 62007021), and the Young Scholars Program of Shandong University, Weihai (20820211005).

References

[1] J. Barbaro and C. Dissanayake, "Autism spectrum disorders in infancy and toddlerhood: a review of the evidence on early signs, early identification tools, and early diagnosis," *Journal of Developmental & Behavioral Pediatrics*, vol. 30, no. 5, pp. 447–459, 2009.

[2] R. Cooper, *Diagnosing the diagnostic and statistical manual of mental disorders*. Routledge, 2018.

[3] F. Negin, B. Ozyer, S. Agahian, S. Kacdioglu, and G. T. Ozyer, "Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders," *Neurocomputing*, vol. 446, pp. 145–155, 2021.

[4] C.-H. Yoo, J.-H. Yoo, M.-K. Back, W.-J. Wang, and Y.-G. Shin, "A unified framework to stereotyped behavior detection for screening autism spectrum disorder," *Pattern Recognition Letters*, 2024.

[5] A. Coronato, G. De Pietro, and G. Paragliola, "A situation-aware system for the detection of motion disorders of patients with autism spectrum disorders," *Expert Systems with Applications*, vol. 41, no. 17, pp. 7868–7877, 2014.

[6] Y. Zhang, Y. Tian, P. Wu, and D. Chen, "Application of skeleton data and long short-term memory in action recognition of children with autism spectrum disorder," *Sensors*, vol. 21, no. 2, p. 411, 2021.

[7] Y. Tian, X. Min, G. Zhai, and Z. Gao, "Video-based early asd detection via temporal pyramid networks," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 272–277.

[8] N. M. Rad, S. M. Kia, C. Zarbo, T. van Laarhoven, G. Jurman, P. Venuti, E. Marchiori, and C. Furlanello, "Deep learning for automatic stereotypical motor movement detection using wearable sensors in autism spectrum disorders," *Signal Processing*, vol. 144, pp. 180–191, 2018.

[9] M. S. Goodwin, M. Haghighi, Q. Tang, M. Akcakaya, D. Erdogmus, and S. Intille, "Moving towards a real-time system for automatically recognizing stereotypical motor movements in individuals on the autism spectrum using wireless accelerometry," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 2014, pp. 861–872.

[10] F. Albinali, M. S. Goodwin, and S. Intille, "Detecting stereotypical motor movements in the classroom using accelerometry and pattern recognition algorithms," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 103–114, 2012.

[11] F. Albinali, M. S. Goodwin, and S. S. Intille, "Recognizing stereotypical motor movements in the laboratory and classroom: a case study with children on the autism spectrum," in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 71–80.

[12] M. S. Goodwin, S. S. Intille, F. Albinali, and W. F. Velicer, "Automated detection of stereotypical motor movements," *Journal of autism and developmental disorders*, vol. 41, no. 6, pp. 770–782, 2011.

[13] M. B. Shaikh and D. Chai, "Rgb-d data-based action recognition: A review," *Sensors*, vol. 21, no. 12, p. 4246, 2021.

[14] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: research and evaluation challenges," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 650–663, 2014.

[15] R. Yue, Z. Tian, and S. Du, "Action recognition based on rgb and skeleton data sets: A survey," *Neurocomputing*, 2022.

[16] A. Ali, F. F. Negin, F. F. Bremond, and S. Thümmler, "Video-based behavior understanding of children for objective diagnosis of autism," in *VISAPP 2022-17th International Conference on Computer Vision Theory and Applications*, 2022.

[17] N. Muty and Z. Azizul, "Detecting arm flapping in children with autism spectrum disorder using human pose estimation and skeletal representation algorithms," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. IEEE, 2016, pp. 1–6.

[18] W. Liu, Y. Zhang, B. Zhang, Q. Xiong, H. Zhao, S. Li, J. Liu, and Y. Bian, "Self-guided dmt: Exploring a novel paradigm of dance movement therapy in mixed reality for children with asd," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–10, 2024.

[19] J. Liu, Y. Bian, Y. Yuan, Y. Xi, W. Geng, X. Jin, W. Gai, X. Fan, F. Tian, X. Meng *et al.*, "Designing and deploying a mixed-reality aquarium for cognitive training of young children with autism spectrum disorder," *Science China Information Sciences*, vol. 64, pp. 1–3, 2021.

[20] J. Liu, Y. Bian, Y. Xi, Y. Zheng, J. Huang, W. Gai, C. Yang, and X. Meng, "Evaluating the role of mixed reality in cognitive training of children with asd: Evidence from a mixed reality aquarium," *International journal of human-computer studies*, vol. 162, p. 102815, 2022.

[21] S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 755–761.

[22] M. Jazouli, S. Elhoufi, A. Majda, A. Zarghili, and R. Aalouane, "Stereotypical motor movement recognition using microsoft kinect with artificial neural network," *International Journal of Computer and Information Engineering*, vol. 10, no. 7, pp. 1270–1274, 2016.

[23] P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, and M. A. Armin, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, 2023.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[25] X. Xiao, D. Xu, and W. Wan, "Overview: Video recognition from handcrafted method to deep learning method," in *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, 2016, pp. 646–651.

[26] C. Zhang, Y. Xu, Z. Xu, J. Huang, and J. Lu, "Hybrid handcrafted and learned feature framework for human action recognition," *Applied Intelligence*, vol. 52, no. 11, pp. 12 771–12 787, 2022.

[27] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[28] C. I. Orozco, E. Xamena, M. E. Buemi, and J. J. Berlles, "Human action recognition in videos using a robust cnn lstm approach," *Ciencia y Tecnología*, pp. 23–36, 2020.

[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[31] L. Xia and W. Ma, "Human action recognition using high-order feature of optical flows," *The Journal of Supercomputing*, vol. 77, no. 12, pp. 14 230–14 251, 2021.

[32] I. Dave, R. Gupta, M. N. Rizve, and M. Shah, "Tclr: Temporal contrastive learning for video representation," *Computer Vision and Image Understanding*, vol. 219, p. 103406, 2022.

[33] Y. Tian, Y. Yan, G. Zhai, L. Chen, and Z. Gao, "Clsa: a contrastive learning framework with selective aggregation for video rescaling," *IEEE Transactions on Image Processing*, vol. 32, pp. 1300–1314, 2023.

[34] Y. Tian, G. Lu, Y. Yan, G. Zhai, L. Chen, and Z. Gao, "A coding framework and benchmark towards low-bitrate video understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[35] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[36] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[37] Y. Tian, G. Lu, G. Zhai, and Z. Gao, "Non-semantics suppressed mask learning for unsupervised video semantic compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 610–13 622.

[38] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[39] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[40] T. Soo Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20–28.

[41] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *arXiv preprint arXiv:1705.08106*, 2017.

[42] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Body joint guided 3-d deep convolutional descriptors for action recognition," *IEEE transactions on cybernetics*, vol. 48, no. 3, pp. 1095–1108, 2017.

[43] M. Majd and R. Safabakhsh, "Correlational convolutional lstm for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, 2020.

[44] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based lstm networks," *Applied soft computing*, vol. 86, p. 105820, 2020.

[45] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[46] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[47] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[48] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[49] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[50] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[51] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[52] P. Khaire, P. Kumar, and J. Imran, "Combining cnn streams of rgb-d and skeletal data for human activity recognition," *Pattern Recognition Letters*, vol. 115, pp. 107–116, 2018.

[53] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Skeleton-indexed deep multi-modal feature learning for high performance human action recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[54] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[56] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018.

[57] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[59] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[62] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.