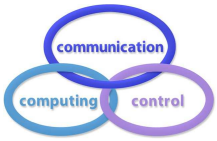
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 20, Issue: 6, Month: December, Year: 2025
Article Number: 6883, https://doi.org/10.15837/ijccc.2025.6.6883



CCC Publications

Cross-Modality Distillation for Multi-View Action Recognition

S. Liu, W. Liu, J. Tian

Siyuan Liu
College of Data Science and Application
Inner Mongolia University of Technology, China
No. 49, Aimin Street, Xincheng District, Huhhot 010080, Inner Mongolia
20221100137@imut.edu.cn

Wenjing Liu*
College of Data Science and Application
Inner Mongolia University of Technology, China
No. 49, Aimin Street, Xincheng District, Huhhot 010080, Inner Mongolia
*Corresponding author: liuwenjing2015@bupt.edu.cn

Jie Tian
Department of Computer Science
New Jersey Institute of Technology, NJ
07102 USA
jt66@njit.edu

Abstract

Behavior recognition provides important help and support in the fields of medical care, security, and intelligent transportation, and thus has received wide attention in the field of practical intelligent applications. However, there remain many challenges in the task of behavior recognition under distributed multi-view video, such as lighting changes under different viewpoints, trunk posture changes, and background noise, which seriously affect the accuracy of behavior recognition. To address these challenges, a multi-view cross-modal distillation behavior recognition method is proposed. Data from two different modalities, skeletal points and RGB, are included to construct teacher and student networks respectively, and KL divergence is used to evaluate cross-modal knowledge transformation to achieve behavior recognition under multiple views. Meanwhile, semi-supervised learning framework is designed to improve the learning performance of the student network through pseudo-labeling. The consistency information of behaviors under different viewpoints is learned among the introduced multiple student networks, which effectively improves the stability and accuracy of multi-view behavior recognition. Experimental results on the behavior recognition datasets NTU RGB+D 60 and NTU RGB+D 120 show that the method outperforms some current popular methods in terms of recognition accuracy. In addition, further experiments conducted in an experimental environment built with real edge devices validate the feasibility of the method for deployment and use in distributed environments.

Keywords: Deep Action recognition, Cross-Model Knowledge Distillation, Contrastive learning, Multi data modality, Edge Intelligence.

1 Introduction

With the continuous proliferation of IoT devices, IoT technology is becoming one of the important development directions in the digital era [**?**]. Alongside the rapid increase in IoT devices, the vast amount of data generated is also gradually increasing. As a result, processing and analyzing this data has become an urgent challenge . Edge computing technology, as an effective solution, enables efficient data processing and analysis on edge devices, significantly reducing latency and bandwidth loss in the data transmission process [18]. Within this paradigm, behavior recognition has gained substantial attention due to its potential to enhance applications such as video surveillance, intelligent human-computer interaction, and healthcare monitoring . In an IoT architecture centered on edge computing, behavior recognition technology—widely regarded as a critical IoT application—has attracted increasing attention from researchers and society alike. This technology can collect users' daily activity data from IoT devices for real-time processing and analysis, enabling the recognition and analysis of user behavior, thus enhancing their quality of life and safety. Therefore, researching and exploring efficient behavior recognition technologies within edge computing holds significant research and application value [24].

The fundamental bottleneck in image-based recognition within edge computing lies in the trade-off between computational efficiency and recognition accuracy, particularly under complex environmental conditions. Factors such as occlusion, viewpoint variability, background noise, and lighting changes significantly affect the accuracy of behavior recognition. To improve recognition accuracy, researchers have explored various methods [22, 31, 32, 33], including the use of skeletal modality data, which enhances recognition precision. However, modality-specific limitations exacerbate the issue: while skeletal data provides robust structural information and scale invariance, it struggles with differentiating actions that share similar motion trajectories or endpoints. On the other hand, RGB data is rich in appearance features but is highly susceptible to environmental variability, making it noisier and less reliable. Effectively utilizing these modalities to achieve robust behavior recognition in distributed edge environments remains an open challenge.

To address these limitations, multimodal data fusion has been proposed as a promising solution , combining the complementary strengths of skeletal and RGB data. However, achieving efficient and scalable fusion suitable for edge environments requires overcoming issues like data inconsistency across modalities, model robustness, and overfitting. To make efficient use of these multimodal data sources, many researchers have proposed the use of Knowledge Distillation (KD). The goal of knowledge distillation is to transfer knowledge from a teacher network to a student network, enabling the fusion of multiple modalities to improve recognition accuracy. It can also help mitigate the limitations edge devices face in accessing certain data modalities due to limited computational resources. Traditional distillation methods minimize output differences between teacher and student networks, transferring useful knowledge to a lightweight student network, which is suitable for RGB-based student networks deployed in edge environments. Various KD approaches have been proposed for behavior recognition, transferring different types of knowledge between models. Examples include transferring knowledge from an optical flow model to an RGB model [9], from an RGB model to a skeletal model [23], and from a skeletal model to an RGB model [4, 12], among others.

However, these methods face the challenge of inconsistency in raw data across different modalities, which can lead to reduced model robustness, data overfitting, decreased learning efficiency, and diminished model generalization. Therefore, a cross-modal contrastive approach is needed to better integrate data from different modalities. This paper explores a novel approach for effectively combining RGB video and skeletal modality, aiming to address the issue of insufficient accuracy in behavior recognition models when distinguishing between actions with similar skeletal structures. While skeletal sequences provide simplified body structure and posture information, are scale-invariant, and are less sensitive to changes in clothing texture and background, they can struggle to differentiate actions with similar motion trajectories and the same start and end points. As shown in Figure 1, actions like drinking, eating, and brushing teeth share similar trajectories, start points, and endpoints from different viewpoints. This indicates that using skeletal sequences alone makes it difficult to distinguish such actions, and thus integrating the rich appearance information from RGB videos offers an effective solution to this problem.

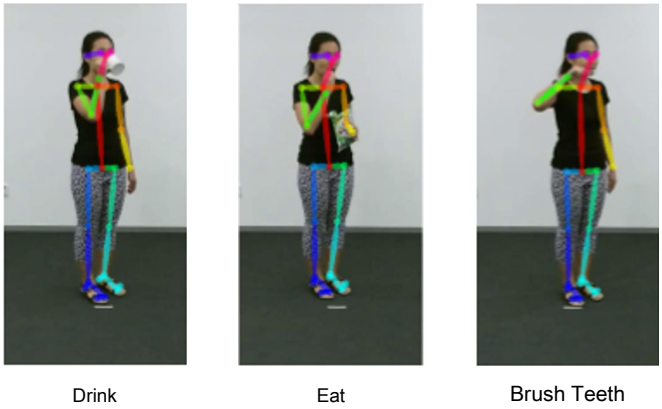


Figure 1: Action with Similar Skeleton.

Based on the above analysis, this paper proposes a Cross-Modal Distillation for Multi-View Action Recognition method (CMMVL) for behavior recognition tasks using multimodal data. The knowledge distillation model in CMMVL effectively combines skeletal and RGB data, two complementary modalities, while comparing multiple student models with multi-view inputs to improve recognition accuracy. The teacher network, with its advantage of extracting most action features and its insensitivity to background, guides the student network to focus on the actions themselves. The extension of the student model with multi-view data allows the learned features to have a more even spatial distribution, which facilitates classification.

This paper's main contributions are threefold:

1. A multi-view human action recognition network model based on cross-modal distillation for edge environments is proposed. This model captures human actions by modeling two types of modality data separately with student and teacher models, making it applicable in real-world scenarios such as video surveillance and intelligent human-computer interaction.

2. The model uses semi-supervised training, where the predictions of the teacher network serve as pseudo-labels for the student network. Under the constraint of KL loss, the student network learns the knowledge provided by the teacher network, enhancing knowledge sharing and transfer between the teacher and student networks.

3. The proposed cross-modal distillation approach expands multiple student networks using multi-view data. The student network learns action consistency information from RGB data, with its rich appearance representation distinguishing different actions with similar skeletons. Through knowledge transfer from the teacher network, it effectively combines the complementary skeletal and RGB modality data.

4. In experimental environments set up on real edge devices, the proposed method outperforms several mainstream methods in recognition accuracy. Experiments also validate the feasibility of deploying and using this method in edge environments.

The remainder of this paper is organized as follows: Chapter 1 discusses related work on multi-view action recognition and knowledge distillation-based action recognition; Chapter 2 presents a multi-view cross-modal action recognition method; Chapter 3 details the experimental implementation, introduces the dataset, and describes and analyzes the experimental results; Chapter 4 provides a summary of this paper.

2 Related Work

2.1 Research on behavior recognition based on multi-view

Nowadays, action recognition methods have achieved considerable success; however, these methods primarily focus on single-view videos. In multi-view environments, the availability of different modalities such as pose and depth has inspired numerous studies aimed at solving the problem of view invariance. In this line of research, most methods leverage depth [1], RGB [19, 25], skeletal data [15],

or multimodal approaches [29] to learn view-independent features.

In multimodal view-independent feature learning, Wang et al. [27] proposed multi-view representation learning network (MRLN) to enhance few-shot action recognition by modeling spatial, temporal, and inter-video relations. Xu et al. [30] presents a two-stage multi-view framework for efficient classification and localization of distracted driving behaviors, addressing high-similarity behaviors and background interference. Dhiman et al. [7] proposed a deep view-invariant human action recognition framework by combining motion and Shape-Time Dynamics (STD) dual streams. The motion stream encapsulates the action's motion content into RGB Dynamic Images (RGB-DI) and processes it using a fine-tuned Inception V3 model. The STD stream employs a series of Long Short-Term Memory (LSTM) and Bi-directional LSTM (BiLSTM) networks to learn long-term view-invariant shape dynamics of actions. The Human Pose Model (HPM) generates view-invariant features based on the Structural Similarity Index Matrix (SSIM) from key depth pose frames. However, the use of multiple LSTM networks results in high computational costs for temporal action features. Liu et al. [16] proposed a new action recognition approach to learn view-independent features, encoding the spatiotemporal information of skeletal joint sequences into a View-Invariant Skeleton Map (VISM) and using a 3D convolutional neural network for 3D action recognition with VISM features. Existing methods learn view-invariant features from skeletal sequences, but these approaches require the availability of 2D/3D pose information. Additionally, the 3D motion modality has also shown effectiveness in action recognition, but obtaining 3D motion is computationally costly, and these methods do not generalize well to the more accessible RGB modality. In summary, while leveraging multimodal data to learn view-invariant features across multiple views can improve recognition performance, further research is needed on multi-view action recognition methods suitable for deployment on edge devices.

This paper focuses on multi-view human action recognition in edge environments, specifically addressing the application of using only RGB data for action recognition. In this study, contrastive learning is employed to efficiently utilize multi-view data to learn view-independent features, enhancing the quality of the learned action features. Through this approach, our method not only improves multi-view action recognition accuracy compared to mainstream methods but is also more suitable for deployment on edge devices.

2.2 Research on behavior recognition based on knowledge distillation

Knowledge distillation [11] was initially proposed to extract knowledge from a large model to a smaller model, enhancing the performance of the smaller model during testing. In the field of action recognition, distillation learning has recently gained significant attention. Knowledge distillation is used to transfer knowledge from a teacher network to a lightweight student network. Currently, knowledge distillation is applied to action recognition learning, with frameworks increasingly suited to cross-modal knowledge distillation. Wang et al. [26] proposes a knowledge distillation framework using a generative model to enhance spatial-temporal feature semantics for video tasks. Garcia et al. [9] proposed a distillation framework comprising teacher and student networks, which can hallucinate depth features from RGB features. Crasto et al. [3] introduced MARS, which trains an RGB stream with standard cross-entropy loss while mimicking the feature learning of the optical flow stream. This mimicry is achieved through a distillation loss that minimizes the Euclidean distance between the features learned by the two streams. Xiao et al. [28] extracted fine-grained motion representations from Temporal Gradients (TG) and enforced consistency between different modalities (i.e., RGB and TG). Many distillation methods in action recognition research not only focus on optical flow and RGB but also explore RGB and skeletal information. The method in [12] is specifically designed to combine cross-modal information from RGB and skeletons. By injecting skeletal information into the RGB stream through feature-level and attention-level distillation mechanisms, this approach provides a practical model for combining RGB and 3D poses. However, methods that combine RGB and skeletons do not explicitly consider similar skeletal sequences that represent human actions. Cross-modal data inconsistency between RGB and skeletal features remains a major hurdle, as these modalities often capture different aspects of motion, leading to misalignments during distillation. While multi-stream models (e.g., optical flow and RGB, or skeleton and RGB) have been explored, few methods have leveraged multi-view consistency in learning from these modalities. This paper shows that this issue

can be addressed by expanding multiple student models.

This paper investigates and designs a cross-modal distillation network model for knowledge transfer from skeletal sequences to RGB videos. The goal is to leverage the advantages of the 3D skeletal teacher model, which can extract most action features and is less sensitive to background, to guide the student model in focusing on the actions themselves. Expanding actions in the student model across multiple views results in more spatially uniform feature distribution, enhancing action recognition accuracy.

3 Multi-perspective cross-modal behavior recognition method

3.1 Cross-modal distillation data input

To efficiently utilize RGB video data, the data preprocessing method follows Duan et al. [8], which uses a 2D image-based human pose feature learning approach to process videos. The preprocessed data is then fed into the teacher network for learning. The video preprocessing process is as follows: first, video frames are extracted from the RGB video, and a top-down pose estimator is used to generate 2D poses; second, the 2D pose heatmaps are obtained using a human pose feature learning approach based on 2D images; finally, the 2D poses are stacked along the temporal dimension to create 3D heatmaps, which are then input into the student network for learning. The principles of data processing are explained as follows:

The dimensions of the 2D pose representation are $K \times H \times W$, where K is the number of joints, and H and W are the height and width of the frame, respectively. When using heatmaps generated by the top-down pose estimator directly as target heatmaps, zero padding is applied to the target heatmap to match the original frame, given the corresponding bounding box. For existing joint coordinate triplets (x_k, y_k, c_k), joint heatmaps J are obtained by synthesizing K Gaussian maps centered on each joint:

$$J_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2\times\sigma^2}} \times \mathrm{c}_k, \tag{1}$$

whrer θ controls the spread of the variance of the Gaussian map, and (x_k, y_k) and c_k represent the position and confidence score of the k-th joint, respectively. The limb heatmap L is as follows:

$$L_{kij} = e^{-\frac{D((i,j),\mathrm{seg}[a_k,\ b_k])^2}{2\times\sigma^2}} \times \min\left(c_{a_k}, c_{b_k}\right). \tag{2}$$

The k-th limb is located between two joints, a_k and b_k. The function D calculates the distance from point i, j to the segment $[(x_{a_k}, y_{a_k}), (x_{b_k}, y_{b_k})]$. Finally, the 3D heatmap is obtained by stacking all heatmaps (J or L) along the temporal dimension, resulting in dimensions of $K \times T \times H \times W$. Two techniques—subject-centered cropping and uniform sampling—are applied, as in reference [19], to further reduce redundancy in the 3D heatmap. The processed 3D heatmap is then fed into the model for training.

3.2 Cross-modal action recognition framework based on knowledge distillation

CMMVL consists of a teacher network and multiple student networks. Figure 2 illustrates the architecture of the CMMVL framework, which consists of a teacher network and multiple student networks. In this example, three student networks—Student Model 1, 2, and 3—process RGB action videos from three different viewpoints. The teacher model, which is based on skeleton sequences corresponding to the same RGB action videos, guides the student networks in learning action representations. The CMMVL framework is organized into three key components. In the first stage, knowledge distillation occurs between the teacher model and Student Model 3, where the teacher's predictions are transferred to the student network. In the second stage, pseudo-labels are generated for the student networks based on the action categories predicted by the teacher model, facilitating semi-supervised learning. The final stage applies contrastive learning to ensure consistency across multiple views by aligning action representations between the student models, promoting robust multi-view feature learning.

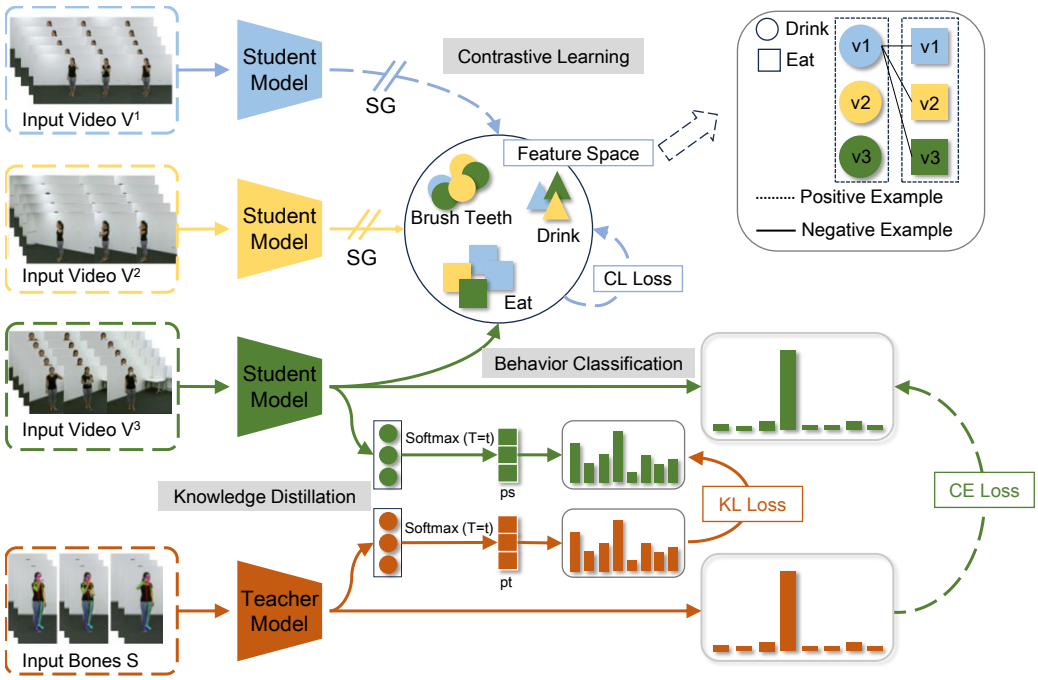


Figure 2: Framework of CMMVL.

The teacher network is a network model that uses the skeleton modality. After pre-training, it predicts the probabilities of the input data, and the class probability distribution for all data in the teacher network is denoted as P_S. The parameters of Student Network 3 are then optimized to make the estimated class probability distribution P_{V^3} as consistent as possible with P_S. In reference [11], KL divergence is proposed as the loss for knowledge transfer between two networks of the same modality. Although the teacher network and the student network use two different modalities of data, the probability distributions of the modality data are consistent, allowing KL divergence to be used as the calculation for the knowledge transfer loss between the teacher network and the student network:

$$\mathrm{KL}\left(P_S^\tau, P_{V^3}^\tau\right) = \sum_i P_S^\tau \left(\log \frac{P_S^\tau(i)}{P_{V^3}^\tau(i)}\right), \tag{3}$$

where P_S^τ and $P_{V^3}^\tau$ represent the target probability distributions of the teacher and student networks, respectively, after being softened by the addition of a temperature parameter. As the temperature parameter changes, the target probability distributions returned by the teacher and student networks become smoother, providing additional information on other classes:

$$P^\tau(n) = \frac{e^{z_n/\tau}}{\sum_i^M e^{z_i/\tau}}. \tag{4}$$

A temperature value $\tau > 1$ generates a smoother probability distribution, with specific recommended values suggested in the literature [9] to avoid overfitting. However, using the loss function (3) alone is not optimal for cross-modal knowledge transfer, as finding an optimal τ is challenging and largely depends on the student network. To address this uncertainty, the teacher network's predictions are used as pseudo-labels, and cross-entropy (CE) loss minimizes the discrepancy between the student network's predictions and the pseudo-labels generated by the teacher network.

$$\mathrm{CE}\left(\hat{P}_V, P_V\right) = -\log\left(P_S\left(\hat{P}_V\right)\right), \tag{5}$$

where $\hat{P}_V$ is the predicted value of the student network and P_V is the predicted value of the teacher network.

3.3 View-independent feature learning with multiple student models

In the context of cross-modal action recognition, Zhang et al. [34] proposed a deep mutual learning strategy. This strategy collaboratively trains a group of student networks, enabling each network to learn by mimicking the teacher network's probability distribution and matching the probability estimates of its peer networks. Experimental validation shows that using the mutual learning strategy can improve the performance of cross-modal knowledge distillation models. However, to reduce the computational overhead of training the student models, the deep mutual learning strategy has been refined.

In the multi-student network learning of cross-modal knowledge distillation, not every student network learns to mimic the teacher network's probability distribution. The training strategy for the student networks is as follows: train a group of N student networks, designating student network N as the primary student network to learn by imitating the teacher network's probability distribution. The N student networks share parameters and optimize the class probability distributions obtained through learning. For all student networks except network N, a gradient stopping operation is performed before generating the class probability distribution.

Multiple student networks are applied to the same modality, using KL loss with a temperature value τ. However, to compensate for the lack of rich appearance information in skeletal data and address the challenge of classifying similar skeletal patterns, CL loss is used to fuse multi-view RGB data, optimizing the class probability distribution. When $N = 3$, the cross-modal knowledge distillation network framework for multi-view fusion through mutual learning is shown in Figure 2. The CL loss functions between student networks 1, 2, and 3 -denoted as $CL^{(1,2)}$, $CL^{(2,3)}$, and $CL^{(1,3)}$ - are given by the following equations:

$$\begin{aligned} CL^{(1,2)} &= \frac{1}{2}D\left(\mathrm{z}^2, stopgrad\left(h^1\right)\right) + \frac{1}{2}D\left(z^1, stopgrad\left(h^2\right)\right), \\ CL^{(2,3)} &= \frac{1}{2}D\left(z^3, stopgrad\left(h^2\right)\right) + \frac{1}{2}D\left(z^2, stopgrad\left(h^3\right)\right), \\ CL^{(1,3)} &= \frac{1}{2}D\left(z^3, stopgrad\left(h^1\right)\right) + \frac{1}{2}D\left(z^1, stopgrad\left(h^3\right)\right). \end{aligned} \tag{6}$$

Here, h and z represent the feature vectors learned by the backbone network and multilayer perceptron in the student network, respectively. $D(z, h)$ denotes the negative cosine similarity between the feature vectors z and h, and $stopgrad()$ represents the gradient stopping operation.

The proposed method can be extended to include more student networks. For N student networks, the optimized CL loss function for the network is given by the following equation:

$$CL^N = CL^{(1,2)} + CL^{(1,3)} + CL^{(2,3)} + \cdots + CL^{(N-1,N)}. \tag{7}$$

Combining the KL loss and CE loss described in Section 3.2, the loss function for optimizing cross-modal knowledge distillation across N student networks is given by the following equation:

$$L^N = KL(P_S^\tau, P_{V^3}^\tau) + CE(\hat{P}_V, P_V) + \lambda CL^N, \tag{8}$$

where is the weighting factor of CL loss. The multi-view distillation model optimization algorithm is shown in Algorithm 1.

In Algorithm 1, during the training process of the distillation model, the teacher and student networks iteratively perform cross-modal knowledge transfer and pseudo-label learning using KL loss and CE loss, while CL loss is used between student networks to fuse multi-view data, compensating for the limitations of the skeletal modality. Thus, CMMVL not only learns to transfer pose knowledge to RGB but also learns discriminative representations through multi-view pose fusion. During testing, we use only the student network, taking RGB video as input to compute action class scores, thereby avoiding issues with missing skeletal modality data.

Algorithm 1 : Multi-view distillation model optimization

Require: $S = \{S^{(i)} \mid i = 1, 2, \ldots, M\}$: Skeletal sequences
1: $V = \{V^{(i)} \mid i = 1, 2, \ldots, M\}$: RGB video streams
2: M: Total number of samples
3: N: Number of action viewpoints
4: P: Probability distribution of actions

Ensure: Trained multi-view distillation model, including the teacher and student networks.
5: Pre-train the teacher model using the skeletal sequences S.
6: Freeze the teacher model parameters and use the skeletal sequences to predict class probabilities, obtaining the target class label $\hat{c}_S$ and soft target class probability P_s.
7: Train N student models using N viewpoints of RGB video streams $V_1, V_2, \ldots, V_N$. The N student models share parameters and each produces predictions $\hat{P}_{V_N}$ and soft target class probabilities $P_{V_1}, P_{V_2}, \ldots, P_{V_N}$.
8: Calculate the KL loss between P_s and P_{V_N} using the values of P_s and P_{V_N} from Steps 2 and 3.
9: Calculate the CE loss between $\hat{P}_{V_N}$ and the true values P_{V_N} using the predictions from Step 3.
10: Calculate the CL loss between each of the N student models as obtained in Step 3.
11: Sum the losses calculated in Steps 4, 5, and 6 and iteratively update the model until convergence.

4 Experiment

4.1 Experimental Setup

4.1.1 Models and Dataset

The CMMVL teacher network employs the PoseC3D network model, while the student network uses I3D as its backbone network. The proposed CMMVL and related detailed ablation studies were evaluated on commonly used datasets (specifically NTU-RGB+D). All experiments were conducted on skeletal modality and corresponding RGB videos, with evaluation performed on the respective RGB validation sets.

(1) The PoseC3D network model [8] is an open-source 3D-CNN-based skeletal action recognition framework provided by the Chinese University of Hong Kong. PoseC3D achieves both excellent recognition accuracy and efficiency, reaching state-of-the-art (SOTA) performance on multiple skeletal action datasets including FineGYM, NTU RGB+D, and Kinetics-skeleton. Unlike traditional GCN methods based on 3D human skeletons, PoseC3D achieves superior recognition performance using only stacked 2D human skeleton heatmaps as input.

(2) The I3D network model [2] is a video action recognition model proposed by DeepMind in 2017. The I3D network model serves as a feature extraction mechanism, where different tasks essentially correspond to different feature space mappings - different tasks can be performed by simply changing the labels, such as video emotion recognition and micro-expression recognition. In RGB video-based action recognition, the I3D network demonstrates stable performance and good effectiveness, and continues to be widely used as a baseline network for video understanding tasks.

(3) The NTU-RGB+D dataset [20], provided by Nanyang Technological University, contains 60 action classes with approximately 56,000 video clips. These are divided into three major categories: 40 daily actions (drinking, eating, reading, etc.), 9 health-related actions (sneezing, staggering, falling down, etc.), and 11 interactive actions (punching/kicking, hugging, etc.). The RGB videos have a resolution of 1920x1080, while both depth maps and infrared videos are 512x424. The 3D skeletal data includes three-dimensional coordinates of 25 body joints per frame. The dataset employs two different evaluation protocols: Cross-subject (CSub) and Cross-view (CView). The NTU-RGB+D 120 dataset [17] expands upon the NTU-RGB+D dataset by adding another 60 classes with approximately 57,600 video samples. This brings the total to 120 classes and 114,480 samples in the NTU-RGB+D 120 dataset. This expanded dataset employs two different evaluation protocols: Cross-subject (CSub) and Cross-setup (CSet).

4.1.2 Environment Setup

All experiments in this section are conducted under the Python deep learning framework on Ubuntu 16.04, utilizing two NVIDIA Tesla P100 GPUs. The training input used both 3D skeletal data and corresponding RGB videos provided by the NTU-RGB+D dataset. The backbone model for the teacher network is the PoseC3D backbone proposed in reference [6]. The student network employs the I3D RGB backbone network proposed in reference [2], which was pre-trained on ImageNet [5] and Kinetics-400 [14].

A test bed for efficiency evaluation of our training strategy is constructed. Without loss of generality, there are two types of subordinates with 100M links connect their edge server, laptop and Raspberry Pi. Two edge nodes collect videos from different viewpoints and transmit them to the edge computing server. The two edge nodes are Raspberry Pi E1 and E2, with the edge computing server designated as S1. The configurations of edge computing server S1 and Raspberry Pi nodes E1 and E2 are shown in Tables 1 and 2. The edge computing server S1 is equipped with an i7-9700 (3.00GHz, 3.00GHz) CPU, NVIDIA GeForce RTX 2060 GPU, and 16GB of memory, while the Raspberry Pi nodes E1 and E2 use a 1.5GHz quad-core Broadcom BCM2711 B0 (Cortex-A72) processor.

Table 1: Detail Information of Central Server

hardware	Central Server
CPU	i7-9700(3.00GHz 3.00 GHz)
Memory	16GB DDR4
GPU	NVIDIA GeForce RTX 2060
Wireless network	802.11ac(2.4/5GHz) Bluetooth5.0
Wired network	Gigabit Ethernet
Idle power	120W
CPU full load power	95W
GPU full load power	160W

Table 2: Detail Information of Raspberry Pi 4B

hardware	A type of edge device
CPU	1.5GHz quad-core Broadcom BCM2711BO (Cortex A-72)
Memory	2GB DDR4
GPU	500MHz VideoCore VI
Wireless network	802.11ac(2.4/5GHz) Bluetooth5.0
Idle power	10W
full load power	15W
Average Power	12W

Raspberry Pi nodes E1 and E2 collect videos from viewpoints that differ by 45 degrees and transmit them to edge computing server S1, where the server S1 performs model training and action recognition.

4.1.3 Implementation details

The hyperparameters are carefully selected as follows. A value of $\tau = 2$ is chosen to smooth class probabilities effectively, aiding in knowledge transfer while avoiding overfitting. We use 3 student networks balances computational complexity with accuracy gains.

4.2 Result and Analysis

In this section, experimental evaluations are conducted to verify the accuracy and effectiveness of CMMVL. CMMVL is compared with current state-of-the-art single skeletal modality and multi-modal action recognition methods. The impact of the multi-student network distillation method on model performance is validated.

To evaluate CMMVL's model performance, this paper compares it with current mainstream deep learning algorithms on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets, as shown in Tables 3 and 4. CMMVL is compared with advanced methods that use only skeletal modality. Previous approaches [21, 22, 31] primarily used spatiotemporal graph convolutional networks to learn action features. As shown in the tables, CMMVL achieves superior action recognition performance under different partition criteria on the NTU-RGB+D dataset compared to previous methods, indicating that RGB and skeletal modalities complement each other to some extent in action recognition tasks, and better exploitation of the relationship between these two modalities can achieve improved recognition results.

Table 3: Comparison with state-of-the-art methods on NTU-RGB+D 60

Method	Pose	RGB	NTU-60(CView)	NTU-60(CSub)
ST-GCN [31]	✓	✗	88.3	81.5
RA-GCN [22]	✓	✗	93.6	87.3
2s-AGCN [21]	✓	✗	95.1	88.5
SGN [33]	✓	✗	94.5	89.0
MMFF [35]	✓	✓	91.6	85.4
3s-AimCLR [10]	✓	✓	92.8	86.9
CMMVL	✓	✓	95.6	89.1

Furthermore, compared to advanced methods [10, 35] that use both RGB and skeletal modalities, CMMVL achieves better results through its cross-modal multi-student network knowledge distillation model. As shown in Table 3, CMMVL's two experimental results on the NTU-RGB+D 60 dataset surpass the state-of-the-art methods by 0.5% and 0.1% respectively. In Table 4, CMMVL's two experimental results on the NTU-RGB+D 120 dataset exceed previous methods by 2.1% and 1.3% respectively. This demonstrates that the cross-modal approach using multi-student network knowledge distillation better integrates the two data modalities and can learn superior action representation information.

Table 4: Comparison with state-of-the-art methods on NTU-RGB+D 120

Method	Pose	RGB	NTU-120(CSet)	NTU-120(CSub)
ST-GCN [31]	✓	✗	73.2	70.7
RA-GCN [22]	✓	✗	82.7	81.1
SGN [33]	✓	✗	81.5	79.2
3s-AimCLR [10]	✓	✓	80.9	80.1
CMMVL	✓	✓	84.8	82.4

The NTU-RGB+D 120 dataset has greater complexity due to its larger number of classes and increased inter-class similarity. The CMMVL framework's ability to leverage complementary modalities (skeletal and RGB) and its multi-student structure make it particularly effective for handling such complexity. This is less critical for NTU-RGB+D 60, where the skeletal modality alone achieves near-optimal results for many classes.

Table 5: Comparison of the Backbone Network only used the main student model

Knowledge Distillation	NTU-60(CView)	NTU-60(CSub)	NTU-120(CSet)	NTU-120(CSub)
Not used	87.3	85.5	80.1	77.0
Used	95.6	89.1	84.8	82.4

Table 6 presents a comparison of inference time across different methods, including the proposed CMMVL framework, ST-GCN, RA-GCN, SGN, and 3s-AimCLR. The x-axis represents the methods, and the y-axis denotes the inference time in seconds. The proposed CMMVL framework exhibits a

Table 6: Inference time of various methods.

Method	CMMVL	ST-GCN	RA-GCN	SGN	3s-AimCLR
Time(s)	0.09	0.13	0.11	0.16	0.15

competitive inference time, achieving a value below 0.1 seconds per sample. This demonstrates its computational efficiency, especially in real-time applications.

4.3 Ablation Experiment

To verify the impact on model performance of knowledge distillation, mutual learning between multiple student models, multi-student model learning strategies, multi-view fusion learning, and hyperparameter selection, the following ablation experiments were conducted on the NTU-RGB+D dataset:

The results of the ablation study on the impact of knowledge distillation algorithm on model performance are shown in Table 5, comparing the effectiveness with and without knowledge distillation. The model without knowledge distillation, which uses only the backbone network of the main student model on the RGB modality, represents the performance of the I3D network on the NTU-RGB+D dataset. Comparison of the data in the table shows that the network model incorporating knowledge distillation outperforms the I3D network alone in recognition effectiveness across all metrics. This demonstrates that the introduction of the knowledge distillation algorithm effectively combines skeletal data with RGB data, and the complementary learning of these two modalities in action representation improves action recognition performance.

Table 7: Comparison of the Backbone Network only used the main student model

Knowledge Distillation	NTU-60(CView)	NTU-60(CSub)	NTU-120(CSet)	NTU-120(CSub)
Not used	87.3	85.5	80.1	77.0
Used	95.6	89.1	84.8	82.4

To investigate the effectiveness of mutual learning between multiple student models, a comparison was made between models with and without multiple student models, as shown in Table 6. The model in Table 6 uses only the main student model, with all viewpoint data input into the main student model, without inter-viewpoint feature fusion computation. The comparison of data in the table shows that the multi-student model achieves better recognition performance than the single student model. This indicates that the multi-student model learned viewpoint consistency information when processing RGB data from multiple viewpoints, effectively improving model recognition performance.

Table 8: Comparison of the Backbone Network only used the main student model

Multi-student model	NTU-60(CView)	NTU-60(CSub)	NTU-120(CSet)	NTU-120(CSub)
✗	94.8	87.7	83.3	81.1
✓	95.6	89.1	84.8	82.4

To study how multi-view RGB modality compensates for the limitations in distinguishing similar actions due to similar skeletal and motion patterns, classification results for similar actions such as eating and drinking are shown in Table 7. The table compares classification performance between using only skeletal modality input with the teacher model backbone network and the CMMVL approach. The data comparison shows that CMMVL's recognition performance surpasses that of the teacher model backbone network on most datasets, with only one metric showing similar performance. This demonstrates that multi-view RGB modality can, to some extent, compensate for the limitations of skeletal modality data. Training the model using knowledge distillation provides a feasible premise for actual model deployment and expands the use of multiple data modalities in edge environments.

To investigate the impact of hyperparameter selection on model performance, a quantitative analysis was conducted on the weighting factor of CL loss, with λ ranging from 0 to 1. The impact of

Table 9: Comparison of PoseC3D and CMMVL

Model	NTU-60(CView)	NTU-60(CSub)	NTU-120(CSet)	NTU-120(CSub)
Teacher Network	95.1	88.5	84.8	82.1
CMMVL	95.6	89.1	84.8	82.4

Table 10: The Influence of Different λ on the Accuracy.

λ	0.0005	0.005	0.01	0.05	0.1	0.2	0.5	0.9
Accuracy	58	**87**	59	57	50	52	55	53

different λ values on accuracy under the CSub evaluation protocol of the NTU-RGB+D 60 dataset is shown in Table 10. The graph indicates that optimal performance is achieved when $\lambda = 0.005$.

5 Conclusion

To address issues such as data heterogeneity in multi-view action recognition in practical edge environments, this paper proposes a multi-view human action recognition model based on cross-modal distillation. The model primarily expands the use of multiple data modalities in edge environments and resolves the challenge of distinguishing between actions with similar skeletal features. The proposed model employs knowledge distillation algorithms, using existing high-quality skeletal models to guide student models in learning human action features. Specifically, the multi-student network collectively learns potential spatial relationships between multiple viewpoints of different actions, with knowledge transfer occurring only between the main student network and the teacher model, thereby improving distillation learning efficiency. This model not only learns viewpoint-invariant features across multiple viewpoints but also constructs student networks suitable for deployment in practical edge scenarios. Experiments based on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets, compared with current mainstream action recognition methods, showed improvements of 0.5%, 0.1%, 2.1%, and 1.3% respectively, demonstrating the superiority of the proposed method. Additionally, significant improvements were achieved through optimizations based on a self-supervised end-to-end multi-view human action recognition model using contrastive learning. Multiple ablation studies validate the effectiveness of the multi-student network distillation model for multi-view action recognition. The proposed model holds strong potential for real-world applications. In smart transportation, where accurate and real-time action recognition is crucial for monitoring and controlling traffic behavior, the model can be deployed on edge devices for efficient analysis of driver behavior, pedestrian actions, or even detecting distracted driving. In medical monitoring, the ability to recognize subtle movements or actions—such as patient rehabilitation exercises or identifying fall events—could improve patient care and safety. The lightweight nature of the model, combined with its high accuracy, makes it suitable for these applications, where both computational efficiency and real-time processing are essential.

Funding

This work was supported by the National Science Foundation of China (61962045, 61502255, 61650205), the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT23104), the Natural Science Foundation of Inner Mongolia Autonomous Region (2023LHMS06023) and the Basic Scientific Research Program of Universities of Inner Mongolia Autonomous Region (JY20220273, JY20240002, JY20240061)

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Ashraf, N., Sun, C., & Foroosh, H. (2014). View invariant action recognition using projective depth. *Computer Vision and Image Understanding*, 123, 41–52.

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

[3] Crasto, N., Weinzaepfel, P., Alahari, K., & Schmid, C. (2019). Mars: Motion-augmented rgb stream for action recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7882–7891.

[4] Das, S., Dai, R., Yang, D., & Bremond, F. (2021). Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (12), 9703–9717.

[5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

[6] Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2969-2978).

[7] Dhiman, C., & Vishwakarma, D. K. (2020). View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on Image Processing*, 29, 3835–3844.

[8] Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2969–2978.

[9] Garcia, N. C., Morerio, P., & Murino, V. (2018). Modality distillation with multiple stream networks for action recognition. *Proceedings of the European Conference on Computer Vision (ECCV)*, 103–118.

[10] Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., & Ding, R. (2022). Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (1), 762–770.

[11] Hinton, G. (2015). Distilling the knowledge in a neural network. *arXiv preprint* arXiv:1503.02531.

[12] Hong, J., Fisher, M., Gharbi, M., & Fatahalian, K. (2021). Video pose distillation for few-shot, finegrained sports action recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9254–9263.

[13] Hornos, M. J., & Quinde, M. (2024). Development methodologies for iot-based systems: Challenges and research directions. *Journal of Reliable Intelligent Environments*, 1–30.

[14] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al (2017). The kinetics human action video dataset. *arXiv preprint* arXiv:1705.06950.

[15] Li, Y., Xia, R., & Liu, X. (2020). Learning shape and motion representations for view invariant skeleton-based action recognition. *Pattern Recognition*, 103, 107293.

[16] Liu, J., & Xu, D. (2021). Geometrymotion-net: A strong two-stream baseline for 3d action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31 (12), 4711–4721.

[17] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10), 2684-2701.

[18] Modupe, O. T., Otitoola, A. A., Oladapo, O. J., Abiona, O. O., Oyeniran, O. C., Adewusi, A. O., Komolafe, A. M., & Obijuru, A. (2024). Reviewing the transformational impact of edge computing on real-time data processing and analytics. *Computer Science & IT Research Journal*, 5 (3), 693–702.

[19] Rani, S. S., et al (2020). Self-similarity matrix and view invariant features assisted multi-view human action recognition. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 1–6.

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

[21] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12026–12035.

[22] Song, Y.-F., Zhang, Z., Shan, C., & Wang, L. (2020). Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31 (5), 1915–1925.

[23] Thoker, F. M., & Gall, J. (2019). Cross-modal knowledge distillation for action recognition. *2019 IEEE International Conference on Image Processing (ICIP)*, 6–10.

[24] Tsai, Y.-H., & Hsu, T.-C. (2024). An effective deep neural network in edge computing enabled internet of things for plant diseases monitoring. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 695–699.

[25] Ullah, A., Muhammad, K., Hussain, T., & Baik, S. W. (2021). Conflux lstms network: A novel approach for multi-view action recognition. *Neurocomputing*, 435, 321–329.

[26] Wang, G., Zhao, P., Shi, Y., Zhao, C., & Yang, S. (2024, March). Generative Model-Based Feature Knowledge Distillation for Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 14, pp. 15474-15482).

[27] Wang, X., Lu, Y., Yu, W., Pang, Y., & Wang, H. (2024). Few-shot Action Recognition via Multi-view Representation Learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

[28] Xiao, J., Jing, L., Zhang, L., He, J., She, Q., Zhou, Z., Yuille, A., & Li, Y. (2022). Learning from temporal gradient for semi-supervised action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3252–3262.

[29] Xiao, Y., Chen, J., Wang, Y., Cao, Z., Zhou, J. T., & Bai, X. (2019). Action recognition for depth video using multi-view dynamic images. *Information Sciences*, 480, 287–304.

[30] Xu, Y., Jiang, S., Cui, Z., & Su, F. (2024). Multi-View Action Recognition for Distracted Driver Behavior Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7172-7179).

[31] Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI conference on artificial intelligence*, 32 (1).

[32] Zhang, D., Du, C., Peng, Y., Liu, J., Mohammed, S., & Calvi, A. (2024). A multi-source dynamic temporal point process model for train delay prediction. *IEEE Transactions on Intelligent Transportation Systems*.

[33] Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., & Zheng, N. (2020). Semantics-guided neural networks for efficient skeleton-based human action recognition. *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1112–1121.

[34] Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.

[35] Zhu, X., Zhu, Y., Wang, H., Wen, H., Yan, Y., & Liu, P. (2022). Skeleton sequence and rgb frame based multi-modality feature fusion network for action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18 (3), 1–24.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Liu, S.; Liu, W.; Tian, J. (2025). Cross-Modality Distillation for Multi-View Action Recognition, *International Journal of Computers Communications & Control*, 20(6), 6883, 2025.
https://doi.org/10.15837/ijccc.2025.6.6883