MMNet: A Model-based Multimodal Network for Human Action Recognition in RGB-D Videos

Bruce X.B. Yu, *Member, IEEE,* Yan Liu, *Member, IEEE,* Xiang Zhang, Sheng-hua Zhong, *Member, IEEE,* and Keith C.C. Chan, *Member, IEEE*

Abstract—Human action recognition (HAR) in RGB-D videos has been widely investigated since the release of affordable depth sensors. Currently, unimodal approaches (e.g., skeleton-based and RGB video-based) have realized substantial improvements with increasingly larger datasets. However, multimodal methods specifically with model-level fusion have seldom been investigated. In this paper, we propose a model-based multimodal network (MMNet) that fuses skeleton and RGB modalities via a model-based approach. The objective of our method is to improve ensemble recognition accuracy by effectively applying mutually complementary information from different data modalities. For the model-based fusion scheme, we use a spatiotemporal graph convolution network for the skeleton modality to learn attention weights that will be transferred to the network of the RGB modality. Extensive experiments are conducted on five benchmark datasets: NTU RGB+D 60, NTU RGB+D 120, PKU-MMD, Northwestern-UCLA Multiview, and Toyota Smarthome. Upon aggregating the results of multiple modalities, our method is found to outperform state-of-the-art approaches on six evaluation protocols of the five datasets; thus, the proposed MMNet can effectively capture mutually complementary features in different RGB-D video modalities and provide more discriminative features for HAR. We also tested our MMNet on an RGB video dataset Kinetics 400 that contains more outdoor actions, which shows consistent results with those of RGB-D video datasets.

Index Terms—Human action recognition, model-based fusion, ensemble learning

1 Introduction

Human action recognition (HAR) is an active research area in computer vision that extends to many practical applications in realms such as healthcare and physical rehabilitation, interactive entertainment, and video understanding. Technological advances in human body skeleton detection have enabled skeleton features to be affordably and easily retrieved, leading to a relatively sparser and more heterogeneous data modality compared with existing RGB or depth modalities. HAR has recently witnessed notable improvements in unimodal methods such as skeleton-based and RGB video–based methods. For instance, skeleton-based methods [1], [2], [3], [4] use graph convolutional models to represent spatiotemporal features of skeleton joints and skeleton bones; these methods lead to performance improvements by aggregating the results of homogeneous input (i.e., skeleton joints and bones). Similarly, approaches using RGB video input [5], [6], [7] are designed to model representations of spatiotemporal features in RGB videos and optical flow streams estimated from such videos.

However, unimodal methods using skeleton or RGB modalities come with obstacles. The major limitation of approaches involving RGB video input is the absence of a 3D structure. Skeleton-based methods are also constrained by the absence of texture and appearance features. Action pairs that have similar skeletal movements in the NTU RGB+D [8] dataset (see Fig. 1), such as "reading" and "writing," "typing" and "writing," or "pointing to something" and "patting other's back," are difficult to distinguish using skeleton-based methods.

Among efforts to better address HAR in RGB-D videos, heterogeneous vision-based multimodal methods that incorporate skeleton and RGB modalities have shown promise in boosting HAR performance [9], [10]. Other attempts have integrated multiple data modalities, such as homogeneous vision-based multimodal (e.g., skeleton joints and bones) HAR methods [2], [4] and even heterogeneous sensor modalities [11], [12], [13]. The core task of multimodal HAR methods is data fusion, which can be further classified as data-level fusion, feature-level fusion, and decision-level fusion [14]. Data-level fusion is rarely adopted when the involved data modalities are intrinsically heterogeneous. Existing data fusion methods usually concatenate feature-level representations at the fully connected layers of modality-specific models or aggregate decision-level results from the final Softmax layers [15], [16], [17]. However, exactly how to effectively fuse data modalities to enhance HAR accuracy in RGB-D videos remains an open question. In addition to data-level, feature-level, and decision-level fusion, [18] summarized another fusion method called co-learning in which knowledge from one data modality facilitates modeling in another data modality; this approach could be applied in multimodal HAR. To advance prior work around co-learning, we propose a novel model-based multimodal network (MMNet) in this paper to model effective knowledge transformation when fusing skeleton and RGB modalities to

- *Bruce Yu, Yan Liu, and Xiang Zhang are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.*
  E-mail: csxbyu, csyliu, csxgzhang@comp.polyu.edu.hk.
- *Sheng-hua Zhong is with the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China.*
  E-mail: csshzhong@szu.edu.cn.
- *Keith Chan was with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong.*
  E-mail: keithccchan@gmail.com.

Manuscript received January 20, 2021; revised June 8, August 3, 2021, and January 10, April 20, 2022; accepted on May 20, 2022
(Corresponding author: Yan Liu)

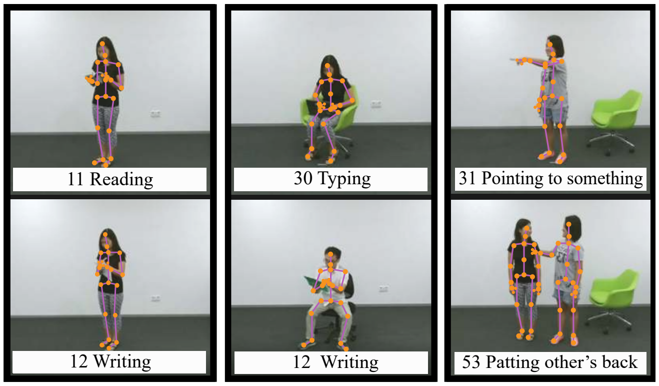


Fig. 1. Difficult action pairs (e.g., Actions 11 and 12, Actions 12 and 30, and Actions 31 and 53) in the NTU RGB+D dataset that confuse skeleton-based models. The goal of this paper is to capture complementary features from the RGB modality to compensate for the limitation of skeleton-based methods.

improve the human action recognition in RGB-D videos.

In our proposed MMNet, we first construct a representation of the RGB modality based on the assumption that an action can be easily recognized by a human when sufficient spatial and appearance features are provided in a temporal manner. To enable machines to simulate human cognition in action recognition, spatial and appearance features should be provided and properly modeled. When human eyes observe an action, the observer develops a general idea of what the subject is doing based on spatial skeleton data. However, human actions typically involve interactions with objects and other human subjects. Narrowing the search space based on objects' appearance features can facilitate machines' action recognition. Object recognition was thus adopted in [15] and [19]. In terms of specifying the objects with which a person is interacting, we focused on areas of the body including the head, hands, and feet, which often convey appearance features of objects and bodily movements. The relationship between an object and a person evolves as an action progresses. As such, we attended to varying appearance features throughout RGB video frames by constructing a spatiotemporal region of interest (ST-ROI) feature map. This strategy alleviates the challenge associated with a vast volume of video data.

When using ST-ROI from the RGB modality, deep learning (DL) models such as VGG nets [20] and ResNet [21] can quickly become overfitted. However, directly applying feeding the ST-ROI to these DL models can not achieve satisfactory single modal and ensemble results. We therefore propose transferring knowledge of the skeleton modality to facilitate action recognition in the RGB modality of our MMNet. In particular, we derived an attention mask from the skeleton joint stream of the proposed MMNet to focus on ST-ROI areas that offer complementary features, which could boost the recognition of human actions in RGB-D videos.

An earlier version of this manuscript appeared in [22]. The present version makes several new contributions. First, we introduce a multimodal DL architecture that fuses different data modalities at the model level with an attention mechanism and uses the skeleton bone stream. Second, our method greatly improves state-of-the-art performance as demonstrated by three benchmarking datasets: NTU RGB+D 120 [10], PKU-MMD [23], and Northwestern-UCLA Multiview [24]. Third, we analyze two key parameters of the proposed MMNet in Sections 4.8 and 4.9 to further validate the method's effectiveness.

The remainder of this paper is organized as follows. Section 2 introduces related work. In Section 3, we detail the proposed MMNet. Section 4 provides ablation results for benchmark datasets and comparisons with state-of-the-art methods. Section 5 concludes the paper.

2 Related Work

HAR has witnessed great progress, from unimodal methods including vision-based [25], ambient sensor–based [26], and wearable sensor–based [27] approaches to the paradigm of multimodal methods. In this section, we discuss research on unimodal HAR and multimodal HAR methods that use data from RGB-D videos.

2.1 Unimodal HAR

2.1.1 Skeleton-based HAR

Skeleton data can be retrieved through vision sensors including depth sensors, stereo cameras, and motion captures [28]. Since the release of RGB-D sensors such as Kinect and RealSense, coupled with advances in human body skeleton detection via RGB cameras, skeleton-based HAR methods have exploded within the computer vision domain. Traditionally, algorithms for skeleton-based HAR focus on modeling geometrical features based on the sequential and spatial characteristics of skeleton sequences. Algorithms including support vector machine, hidden Markov models, and dynamic time warping were common in earlier work [29], which was later dominated by DL algorithms that could automatically learn features from large datasets [30]. Wang et al. [30] reviewed DL models for HAR, such as a deep neural network, convolutional neural network (CNN), and stacked autoencoder.

Available skeleton-based HAR methods appear to emphasize three main directions to improve recognition accuracy. The first direction focuses on data preprocessing and data cleaning. For example, Liu et al. [31] proposed a method that removes skeleton joint noise by learning a model that reconstructs more accurate skeleton data. A similar strategy was proposed by Zhang et al. [32]. The second approach improves HAR benchmarks by proposing novel learning or representation models. For instance, Liu et al. [33] put forth a context-aware LSTM model that could learn which parts of joints contributed to HAR. Since the induction of spatiotemporal graph convolutional networks (ST-GCN) [1], enhanced versions of GCN models have been suggested to improve the results of ST-GCN by considering other physical prior knowledge [4], [34], [35], [36]. The third method involves data augmentation that learns data generation models to produce more training data and provide additional fuel to DL models. Barsoum et al. [37] developed a sequence-to-sequence model for probabilistic human motion prediction, which predicts multiple plausible future human poses from the same input. However, it is not yet clear whether the generated data can be used to enhance HAR models' generalization abilities or accuracy.

2.1.2 RGB Video–Based HAR

Because RGB video data are relatively simple to obtain, large datasets such as UCF-101 [38], HMDB-51 [39], and Kinetics [40] are common benchmarks in video-based HAR. Carreira presented an inflated 3D CNN (I3D) [5] that used a pre-trained Inception-v1 model on ImageNet as its foundation to enhance the performance of UCF-101 and HMDB-51. Two data streams, namely an RGB stream and an optical flow stream (extracted by the TV-L1 algorithm [41]), were applied to vision-based HAR for the two-stream model in [5]. The optical flow stream was found to perform better on UCF-101 and HMDB-51 but was surpassed by the RGB stream on a Kinetics subset. In addition to I3D, Xie et al. [6] considered speed–accuracy trade-offs in video classification and proposed a separable 3D CNN (S3D) model that further improved the performance in [42]. Notably, S3D [6] carried a heavy computational cost, as 3D ConvNets with high training parameters are resource exhaustive.

Intuitively, aggregating the results of S3D with those of skeleton-based methods could boost recognition accuracy. However, for indoor actions with a consistent background as shown in Fig. 1, such video-based methods might not perform well per [43] and [44]. In particular, I3D and S3D are designed for outdoor actions in UCF-101 [38] and Kinetics [40], where the features in background scenes contribute to recognition [45], [46]. The experiments in [47] were performed with a leaning scheme to alleviate the bias in background scenes. Specifically, it penalizes the recognition ability of its model when only background scene information is available. With such a learning scheme, the methods in [47] cannot classify outdoor actions in UCF-101 [38] as effectively as the method in [7]. Comparatively, with proper feature use in background scenes, the methods in [45] and [46] achieve good performance on Kinetics-400. For example, the filters that capture the texture of the water can help the classification of challenging actions such as water skiing and surfing water [45]; while [46] proposed to filter out redundant features in the background scenes to regulate the training process.

The indoor actions in NTU RGB+D [8] may be more challenging to manage using these video-based methods because the actions have relatively less distinguishable information in background scenes. Computational resource limitations pose another barrier, as it needs GPU clusters with 56 GPUs to train.

2.2 Multimodal HAR

2.2.1 Fusion-based Multimodal HAR

Fusion-based multimodal HAR approaches are generally thought to have the potential to boost recognition accuracy and distinguish difficult actions [10], [13]. DL fusion methods can be roughly categorized as joint or coordinated representations [18]. Joint representation is related to model-agnostic approaches that concatenate representations at either the feature or decision level [15], [16], [17]. Yet these fusion methods offer limited improvement and are difficult to enhance further because the relationship between their unimodal networks remains implicit.

Coordinated representation focuses on enforcing either similarity between unimodal representations [43], [48] or more structure on the resulting space, as in correlation-independence analysis [49]. However, in [43] and [48], the authors did not consider whether enforcing similarity between the probability distribution before the Softmax layer could contribute to the ensemble result. Moreover, the methods in [43] and [48] relied on the performance of their cumbersome model, which was aggregated based on the results of submodels, to regularize their modality-specific networks. The correlation analysis in [49] also failed to determine which data modality was best for recognizing which actions. Our multimodal setting is distinct from that in [48], where not all data modalities were available in the testing phase. Instead, we focus on a case where all data modalities were available for the training and testing phases, leading to a multimodal learning setting called multimodal fusion [50].

2.2.2 Model-based Multimodal HAR

Model-based multimodal methods address multimodal HAR at the model level, which is consistent with the concept of co-learning [18] and represents a kind of fusion method. Model-based fusion differs from typical fusion-based methods such as feature- and decision-level fusion, which respectively correspond to the concatenating and adding operations of model-agnostic fusion methods. Our model-based fusion approach is also unique from existing model-level fusion methods that require representation similarity among different modalities. Specifically, our MMNet addresses fusion with co-learning based on a comprehensive understanding of the data structure; specifically, we learned representation from the RGB modality by focusing on body areas that brought mutually complementary features to the skeleton modality.

Similar co-learning attempts were made in [44], [51], and [52], which used skeleton and RGB modalities to capture mutually complementary information. Unlike the approaches proposed in [44], [51], and [52] that focused on appearance features on two hand areas with the help of the skeleton modality, we focused on more body areas including the head, both hands, and both feet in a temporal manner. Another similar method that achieved excellent performance on NTU RGB+D datasets with skeleton and RGB data is VPN [53].

VPN aims to use video-based models (i.e., I3D, S3D, etc.) to improve the recognition of its fused model whereas our method avoids the huge computational cost of video-based models. Meanwhile, our objective is to complement the insufficiency of appearance feature in skeleton data rather than to improve video-based unimodal methods. Besides, our MMNet differs from existing work because it entails a simpler learning structure and fewer loss terms but performs better. We also referred to skeletal features at the decision level because the skeleton bone stream has been shown to be more discriminative than the skeleton joint stream according to [2] and [4].

3 Model-based Multimodal Network

In this section, we introduce the DL architecture of the proposed MMNet from the perspectives of subnetworks used to learn features from the skeleton and RGB modalities.

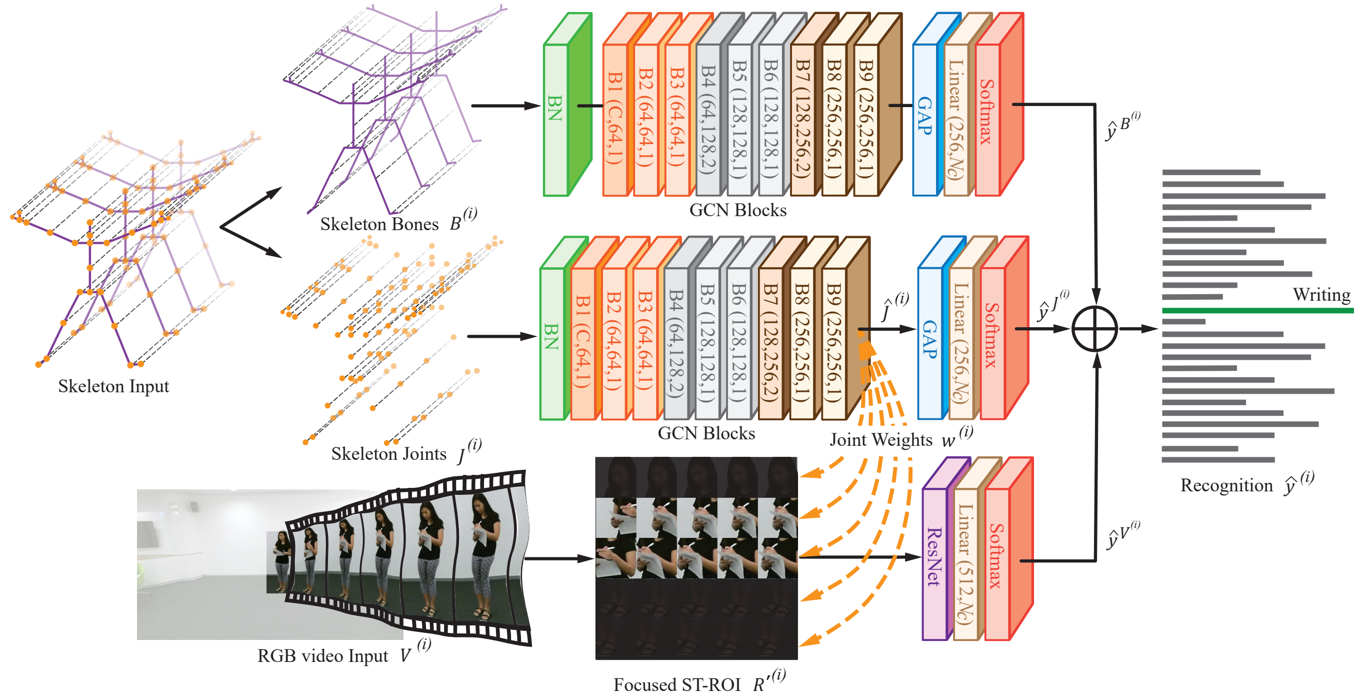


Fig. 2. Architecture of proposed MMNet. $B^{(i)}$, $J^{(i)}$, and $V^{(i)}$ represent the inputs of skeleton bones, skeleton joints, and RGB video, respectively. $w^{(i)}$ are spatial attention weights derived from the graph representation of the skeleton joints $\hat{J}^{(i)}$, which guides the focus of ST-ROI transformed from RGB video input $V^{(i)}$. After this model-based data fusion, the skeleton-focused ST-ROI $R'^{(i)}$ will be fed to the ResNet to generate a modality-specific prediction. $\hat{y}_c^{J^{(i)}}$ and $\hat{y}_c^{B^{(i)}}$ denote respective predictions from skeleton joint and bone streams, which are aggregated through the modality-specific prediction of the RGB modality $\hat{y}_c^{V^{(i)}}$ to deliver the ensemble recognition result.

Then, we elaborate upon the model-based fusion mechanism between the two modalities.

As Fig. 2 indicates, we take skeleton and RGB video data modalities as MMNet input. The proposed MMNet is constructed with three individual networks to learn neural representations from skeleton joints, skeleton bones, and RGB video input. Inspired by [2] and [4], we divide the skeleton input into skeleton joints and skeleton bones. Model-based feature fusion then occurs between the ST-ROI constructed from RGB video input and the joint weights learned from skeleton joints via a GCN model. Given N training samples in a dataset, we denote the features of the ith sample as $\{J^{(i)}, B^{(i)}, V^{(i)}\}$, where $J^{(i)}$ is the skeleton joint input, $B^{(i)}$ is the skeleton bone input, and $V^{(i)}$ is the RGB video input; the corresponding action label is denoted as $y^{(i)}$. The goal is to learn the feature extractors, including for submodels G_J, G_B, and G_V with respective parameters Θ_J, Θ_B, and Θ_V, referring to the action class with an ensemble operation. This operation can be written as

$$\hat{y} = G_J(\Theta_J, J) + G_B(\Theta_B, B) + G_V(\Theta_V, V) \tag{1}$$

3.1 Construct ST-ROI from RGB Modality

Intuitively, video-based models such as I3D [5] and S3D [6] could be top choices to learn discriminative features from the RGB modality. However, these models require vast computational resources in the form of RAM and GPU memory and take a long time to converge. We also observe that early video-based models such as C3D can not perform well on NTU RGB+D 60 due to the limited number of data [43] [44]. Hence, we propose constructing the ST-ROI from the RGB modality and using general CNN models to retrieve effective features from it.

Let us notate $V = \left\{V^{(i)} \mid i = 1, \ldots, N\right\}$ as the RGB modality that has N video samples for training. Then an ordered video sequence of an action in the time interval $[1, T]$ can be represented as $V^{(i)} = (f_1^{(i)}, \ldots, f_t^{(i)}, \ldots, f_T^{(i)})$, where $f_t^{(i)}$ is the frame at time t. To crop the spatial ROI from an action video, we use joints of the skeleton retrieved with the OpenPose tool introduced in [54], which is somewhat more accurate than the skeleton retrieved by the Kinect v2 sensor. Given an RGB frame $f_t^{(i)}$, we define a transformation function g to construct the spatial ROI $R_{tj}^{(i)}$ of a joint as

$$R_{tj}^{(i)} = g\left(f_t^{(i)}, o_{tj}^{(i)}\right), \; j \in (m_1, \ldots, m_{M'_O}), \; M'_O \leq M_O \tag{2}$$

where $o_{tj}^{(i)}$ is the jth joint of the OpenPose skeleton at time t. m_1 to $m_{M'_O}$ are the M'_O indices of the OpenPose skeleton joints we are considering, which are not larger than the total number of OpenPose skeleton joints M_O. Given $V^{(i)} = (f_1^{(i)}, \ldots, f_t^{(i)}, \ldots, f_T^{(i)})$, we perform temporal sampling that selects L representative frames at time $\tau = \{interval \times l \mid l = 1, \ldots, L, \; interval = T/L\}$ and concatenate them into a square ST-ROI as shown in the one-subject case in Fig. 3. For actions that have two subjects, we crop the ST-ROIs of each subject as shown in the two-subject case in Fig. 3. The ST-ROI significantly reduces the data volume of RGB video input while reserving the object's appearance and the movement information of actions. The temporal sub-ROI at time τ will have M' spatial sub-ROIs,

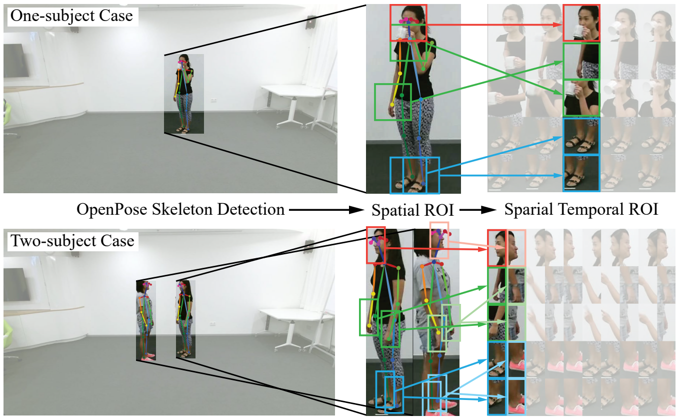


Fig. 3. Construction process of ST-ROI. The top case has one subject, and the bottom case has two subjects. Both cases are based on the OpenPose 2D skeleton.

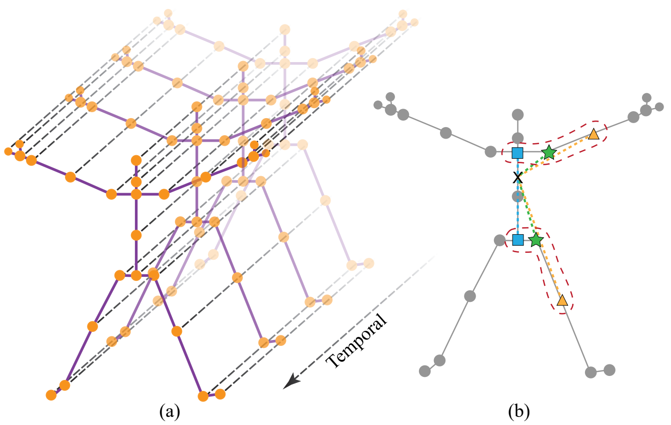


Fig. 4. (a) Structure of spatiotemporal skeleton graph. (b) Spatial sampling strategy of graph convolutional network. Different colors denote different subsets: green stars denote the vertex itself; yellow triangles denote the farther centrifugal subset; blue squares denote the closer centripetal subset.

which can be vertically concatenated and represented as $R_{\tau}^{(i)}$. Conversely, the spatial sub-ROI of the jth joint will have L temporal sub-ROIs and can be horizontally concatenated and represented as $R_{j}^{(i)}$. The ST-ROI of $V^{(i)}$ can then be notated as $R^{(i)}$, which contains $M' \times L$ sub-ST-ROIs denoted as $R_{\tau j}^{(i)}$.

3.2 Learn Joint Weights from Skeleton Modality

For the skeleton modality, the ith training sample that starts at time $t = 1$ and ends at time T with skeleton frames collected at regular intervals can be represented as a sequence of T skeleton frames $J^{(i)} = (J_1^{(i)}, \ldots, J_t^{(i)}, \ldots, J_T^{(i)})$. We denote the corresponding sequence of skeleton bones transformed from the skeleton joints as $B^{(i)} = (B_1^{(i)}, \ldots, B_t^{(i)}, \ldots, B_T^{(i)})$. Given a set of M joints in a skeleton frame observed at time t, we represent it as $J_t^{(i)} = (J_{t1}^{(i)}, \ldots, J_{tj}^{(i)}, \ldots, J_{tM}^{(i)})$ with $J_{tj}^{(i)} \in \mathbb{R}^C$ that has C attributes. We then construct a spatiotemporal graph to represent the spatial and temporal structure of $J^{(i)}$. The structure of the GCN follows [1] and [2]. Fig. 4(a) illustrates the structure of the spatiotemporal skeleton graph, where the joints and bones of a single skeleton frame are depicted by graph vertices (orange circles in Fig. 4[a]) and their natural connections (purple lines in Fig. 4[a]), respectively. Sequentially, two adjacent skeletons are connected by edges between the joints (dashed black lines in Fig. 4[a]). The attribute of a graph vertex can be the corresponding 3D coordinates of each joint. The skeleton graph of a skeleton input $J^{(i)}$ can thus be symbolized as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the joints and bones, respectively. In this graph, the node set $\mathcal{V} = \{v_{tj} \mid v_{tj} = J_{tj}^{(i)}, t = 1, \ldots, T, j = 1, \ldots, M\}$ contains all joints of the skeleton input. Meanwhile, the edge set $\mathcal{E} = \{\varepsilon_t \mid \varepsilon_t = B_t^{(i)} = (v_{tj} - v_{tk}), t = 1, \ldots, T, j, k = 1, \ldots, M\}$ represents all bones of the skeleton input.

3.2.1 Graph Convolutional Operation

To represent the sampling area of convolutional operations, a neighbor set of a node v_{ti} is defined as $N(v_{ti}) = \{v_{tj} \mid d(v_{ti}, v_{tj}) \le D\}$, where D is the maximum path length of $d(v_{ti}, v_{tj})$. Fig. 4(b) displays this strategy, where × represents the skeleton's center of gravity. The sampling area $N(v_{ti})$ is enclosed by the curve. In detail, this strategy empirically uses 3 spatial subsets: the vertex itself (the green star in Fig. 4[b]); the centripetal subset, which contains neighboring vertices closer to the center of gravity (the blue square in Fig. 4[b]); and the centrifugal subset, which contains neighboring vertices farther from the gravity center (the yellow triangle in Fig. 4[b]). Suppose there is a fixed number of K subsets in the neighbor set; they will be labeled numerically with a mapping $l_{ti} : N(v_{ti}) \to \{0, \ldots, K-1\}$. Temporally, the neighborhood concept is extended to temporally connected joints as $N(v_{ti}) = \{v_{qj} \mid d(v_{tj}, v_{ti}) \le K, |q - t| \le \Gamma/2\}$, where Γ is the temporal kernel size that controls the temporal range of the neighbor set. Then the graph convolution can be computed as

$$\hat{v}_{ti} = \sum\nolimits_{v_{tj} \in N(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \, \mathbf{w}(l(v_{tj})) \qquad (3)$$

where $f_{in}(v_{tj})$ is the feature map to acquire the attribute vector of v_{tj}, and $\mathbf{w}(l(v_{tj}))$ is a weight function $\mathbf{w}(v_{ti}, v_{tj}) : N(v_{ti}) \to \mathbb{R}^C$ that can be implemented with a tensor of (C, K) dimension. $Z_{ti}(v_{tj}) = |v_{tk}| l_{ti}(v_{tk}) = l_{ti}(v_{tj})|$ is equal to the cardinality of the corresponding subset, which serves as a normalization term.

3.2.2 Joint Weights

Upon applying graph convolution to the skeleton modality, the output of each vertex on the graph can be used to infer the importance of the corresponding skeleton joint. The feature map of the skeleton sequence can be represented by a tensor of (C, T, M) dimensions, where C denotes the number of attributes of the joint vertex, T denotes the temporal length, and M denotes the number of vertices. This partitioning strategy can be represented by an adjacent matrix $\mathbf{A}$ with its elements indicating whether a vertex v_{tj} belongs to a subset of $N(v_{ti})$. The graph convolution can then be implemented using a $1 \times \Gamma$ classical 2D convolution and by multiplying the resulting tensor by the normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{A} \mathbf{\Lambda}^{-\frac{1}{2}}$ on the second dimension.

With K partitioning strategies $\sum_{k=1}^{K}\mathbf{A}_k$, Equation 3 can be transformed into

$$\hat{J}^{(i)} = \sum_{k=1}^{K} \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{A}\mathbf{\Lambda}^{-\frac{1}{2}} f_{in}\left(J^{(i)}\right)\mathbf{W}_k \odot \mathbf{M}_k \tag{4}$$

where $\mathbf{\Lambda}_k^{ii} = \sum_j(\mathbf{A}_k^{ij}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid empty rows. $\mathbf{W}_k$ is a weight tensor of the 1×1 convolutional operation with $(C_{in},\ C_{out},\ 1,\ 1)$ dimensions, which represents the weight function of Equation 3. $\mathbf{M}_k$ is an attention map with the same size as A_k, demonstrating the importance of each vertex. $\odot$ denotes the element-wise product between two matrices. $\hat{J}^{(i)}$ is a tensor of size $(c,\ t,\ M)$ with c as the number of output channels, t as the output temporal length, and M as the number of vertices. This tensor can be used to infer the action class and can be transformed into joint weights to provide attention knowledge for the RGB modality. The joint weights that represent their corresponding body area importance can be calculated as

$$w^{(i)} = \frac{1}{ct}\sum_{1}^{c}\sum_{1}^{t}\sqrt{\left(\hat{J}_{ct}^{(i)}\right)^2} \tag{5}$$

where t and c are output dimensions of the convolutional graph denoting the temporal length and output channels, respectively. $w^{(i)}$ is a vector that contains the weights of M different skeleton joints.

3.3 Model-based Fusion

We propose a spatial weight mechanism for RGB frames to enable the machine to focus on RGB features that will provide discriminative information. More explicitly, the machine will be more capable as it intuitively mimics action recognition of the human eye. Researchers have aimed to derive an attention weight from the RGB modality itself. For instance, [55] tested four variants of attention mechanisms based on convolutional LSTM [56], but results showed few to no performance improvements. Hence, we have not continued to explore the contributions of self-attention mechanisms in this work; instead, we chose to use joint weights from the skeleton modality and multiply them by the ST-ROI to regularize the RGB modality. The skeleton-focused ST-ROI (denoted as $R'^{(i)}$) of the ith training sample can be mapped from $R^{(i)}$ with a function h defined as

$$R'^{(i)} = h\left(R_j^{(i)}, w_j^{(i)}\right),\ j = m'_1,\ \ldots,\ m'_{M'}, M' < M \tag{6}$$

where w_j is the weight of the jth joint, and $R_j^{(i)}$ is the sub-spatial ROI of the corresponding body area. While $m'_1,\ \ldots,\ m'_{M'}$ are the indices of M' different skeleton joints corresponding to body areas that we propose to focus on. The value of M' equals to that of M'_O in Equation 2. Fig. 5 shows the data fusion process of Equation 6.

3.4 Objective Function

We build the end-to-end format of our MMNet with the sum of a collection of loss terms supervised by the action label, which is represented as

$$\mathcal{L} = \mathcal{L}_J\left(\hat{y}^J, y\right) + \mathcal{L}_B\left(\hat{y}^B, y\right) + \mathcal{L}_V\left(\hat{y}^V, y\right) \tag{7}$$

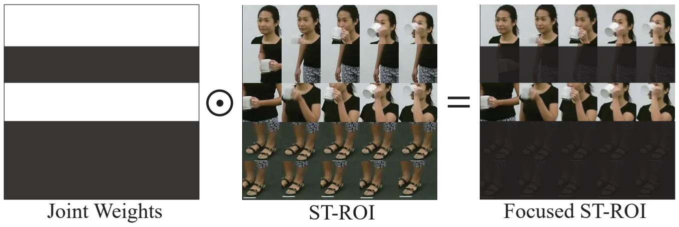


Fig. 5. Model-based fusion scheme of our MMNet. It constructs a skeleton-focused representation of the RGB modality by multiplying joint weights by the ST-ROI.

where $\mathcal{L}_J$, $\mathcal{L}_B$, and $\mathcal{L}_V$ are the loss terms of skeleton joints, skeleton bones, and RGB video input, respectively. We further explain how to obtain modality-specific predictions below.

The skeleton joint input is fed into the graph convolution model introduced in Equation 4. Thus, the prediction of skeleton joints can be defined as

$$\hat{y}^{J^{(i)}} = \sigma\left(G_J\left(\Theta_J, J^{(i)}\right)\right) \tag{8}$$

where G_J represents the graph convolutional operation defined in Equation 4. Θ_J denotes the learnable parameters of the GCN submodel. $J^{(i)}$ is a data sample of skeleton joint input. While σ denotes a linear layer that transforms the shape of the submodel output to a one hot representation, which is also used in Equations 8 and 9.

The skeleton bone input is essentially a transformation of skeleton joint input. Recall that in the graph, the edge set is defined as $\mathcal{E} = \{(\upsilon_{ti}-\upsilon_{tj})|\upsilon_{ti}, \upsilon_{tj} = J_{tj}, t = 1, \ldots, T, i,\ j = 1, \ldots, M\}$, which includes all combinations of joint pairs represented in the adjacency matrix $\mathbf{A}$. Based on the actual structure of skeleton bones in the specific dataset, we follow the transformation method in [2], [4] to build skeleton bones. For example, given two joint vectors $\upsilon_{t1} = (x_1, y_1, z_1)$ and $\upsilon_{t2} = (x_2, y_2, z_2)$, the bone vector can be calculated as $B_{t1} = \varepsilon_{t1} = \upsilon_{t1} - \upsilon_{t2} = (x_1 - x_2, y_1 - y_2, z_1 - z_2)$. We apply the same graph convolutional operation method to skeleton bone input, which can be represented as

$$\hat{y}^{B^{(i)}} = \sigma\left(G_B\left(\Theta_B, B^{(i)}\right)\right) \tag{9}$$

where G_B represents the graph convolutional operation defined in Equation 4. Θ_B denotes learnable parameters of the GCN submodel. $B^{(i)}$ is a data sample of skeleton bone input.

Recall that we have proposed the ST-ROI as the transformed form of RGB video input, which can substantially reduce the data volume and maintain core discriminative information for HAR. As the ST-ROI is intrinsically a 2D feature map, we adopt the ResNet proposed by He et al. [21] to learn features from it. The one hot representation of ResNet can be formulated as

$$\hat{y}^{V^{(i)}} = \sigma\left(G_V\left(R'^{(i)}, \Theta_V\right) + R'^{(i)}\right) \tag{10}$$

where $G_V\left(R'^{(i)}, \Theta_V\right)$ represents the residual mapping to be learned, and Θ_V denotes learnable parameters based on the number of ResNet layers [21].

Given the definitions of the above submodel predictions, we formulate the optimization problem per the following objectives:

$$\underset{\Theta_B}{\arg\min} - \sum\nolimits_{i=1}^{N} \sum\nolimits_{c=1}^{N_c} \underbrace{y_c log\left(\hat{y}_c^{B^{(i)}}\right)}_{\mathcal{L}_B} \tag{11}$$

$$\underset{\Theta_J}{\arg\min} - \sum\nolimits_{i=1}^{N} \sum\nolimits_{c=1}^{N_c} \underbrace{y_c log\left(\hat{y}_c^{J^{(i)}}\right)}_{\mathcal{L}_J} \tag{12}$$

$$\underset{\Theta_V}{\arg\min} - \sum\nolimits_{i=1}^{N} \sum\nolimits_{c=1}^{N_c} \underbrace{y_c log\left(\hat{y}_c^{V^{(i)}}\right)}_{\mathcal{L}_V} \tag{13}$$

where $\mathcal{L}_J$, $\mathcal{L}_B$ and $\mathcal{L}_V$ are cross-entropy losses that enforce the prediction ability from the skeleton joints, skeleton bones, and RGB video, respectively. N_c is the number of action classes in a specific dataset. N denotes the number of samples in the training set.

3.5 Training and Optimization

Several other loss terms could be adopted for joint weights to pursue high recognition accuracy. For instance, according to the findings in [44], both the loss that encourages joint weights to maintain diversity and the loss that leads to joint weights with temporal variance can elicit slight recognition improvements. To ease the process of validating the effectiveness of our MMNet, we avoided using such fine-tuning and hyperparameter-tuning skills. Rather, we adopted a vanilla implementation of joint weights that acts as spatial attention on the RGB modality to verify the effectiveness of our novel model-based data fusion mechanism. Given the objective function, we solved Equations 11, 12, and 13 using stochastic gradient descent (SGD). Note that the network G_J can be either pretrained or simultaneously trained with G_V to derive spatial attention weights for feature fusion. Therefore, the submodels G_J and G_V can be trained end-to-end by tuning the Θ_J together with Θ_V or simply by updating the Θ_V with the Θ_J being fixed. Meanwhile, the network G_B for skeleton bones is trained separately and aggregated to the results of G_J and G_V to deliver the ensemble prediction. Specific training steps are illustrated in Algorithm 1.

Algorithm 1 MMNet Optimization

```
Input:
    $J = \left\{J^{(i)} \mid i = 1, \dots, N\right\}$: skeleton joints
    $B = \left\{B^{(i)} \mid i = 1, \dots, N\right\}$: skeleton bones
    $V = \left\{V^{(i)} \mid i = 1, \dots, N\right\}$: RGB videos
    $M'$: the number of spatial ROIs
    $L$: the number of temporal ROIs
Procedure:
 1: Train $G_J$ with skeleton joints $J$.
 2: Construct a $M' \times L$ ST-ROI $R$ from the RGB video $V$.
 3: Extract joint weights $w$ by feeding $J$ to the trained $G_J$.
 4: Construct skeleton-focused ST-ROI $R'$ from the results
    of Steps 2 and 3.
 5: Train $G_V$ with $R'$.
 6: Train $G_B$ with skeleton bones $B$.
 7: Aggregate prediction results in Steps 1, 5, and 6.
Output:
    Trained MMNet including submodels: $G_J$, $G_B$, and $G_V$
```

4 Experiments

We evaluated the proposed method on five public indoor HAR datasets: NTU RGB+D 60 [8], NTU RGB+D 120 [10], PKU-MMD [23], Northwestern-UCLA Multiview [24], and Toyota Smarthome [57]. To the best of our knowledge, the first three datasets constitute the top 3 largest datasets collected with Microsoft Kinect v2 [58]. Microsoft Kinect v2 is capable of tracking up to six human body skeletons, each of which has 25 skeleton joints. Northwestern-UCLA Multiview was gathered via Microsoft Kinect v1 and contains 20 skeleton joints. We did not perform experiments on other, relatively smaller datasets such as RGBD-HuDaAct [59], MSR Daily Activity 3D [60], or 3D Action Pairs [61] because they were nearly 100% recognized by [49]. We conducted extensive ablation experiments on each dataset to verify the contribution of the proposed fusion scheme and to identify the best empirical practice for training our MMNet as follows:

1) "Skeleton Joint": This model is the GCN submodel implemented with ST-GCN [1] for the skeleton joint stream, by which joint weights for the RGB modality can be learned.
2) "Skeleton Bone": This model is another GCN submodel implemented using ST-GCN [1] for the skeleton bone stream transformed from the skeleton joint modality.
3) "RGB Video (No Joint Weights)": This submodel is implemented with ResNet18 [21] for the RGB modality of the proposed MMNet without knowledge of the skeleton modality.
4) "RGB Video (Dynamic Weights)": This submodel is implemented with ResNet18 [21] for the RGB modality with joint weights of the skeleton modality. Here, "Dynamic" refers to tuning the GCN model together with "ResNet18."
5) "RGB Video (Fixed Weights)": This model is implemented with ResNet18 [21] for the RGB modality with joint weights from the pretrained "GCN-Joints." Here, "Fixed" means training "ResNet18" without updating the parameters of "GCN-Joints."
6) Further improvements to the skeleton modality: This implementation is intended to further verify the contribution of the proposed MMNet in enhancing the performance of more advanced skeleton-based methods by aggregating the results of the RGB modality with those of skeleton-based models (i.e., 2s-AGCN [2] and MS-G3D [4]) that use both skeleton joint and bone streams.

4.1 Evaluation Datasets

NTU RGB+D 60 dataset [8] contains $56,880$ samples of 60 different actions including individual activities, interactions between multiple people, and health-related events. The actions were performed by 40 subjects and recorded from 80 viewpoints.

NTU RGB+D 120 dataset [10] extends NTU RGB+D 60 to 120 action classes with an additional 57,600 samples of 60 extra action classes, which makes it relatively more difficult. The dataset has $114,480$ samples captured from 106 different subjects with 155 distinct viewpoints.

PKU-MMD dataset [23] contains 1076 long, untrimmed video and skeleton sequences. The dataset was performed by 66 subjects from three camera views. With 51 annotated action categories, we retrieved $21,545$ valid action sequences and 6 invalid samples that had no skeleton frames.

Northwestern-UCLA Multiview dataset [24] has 12 action categories, each performed by 10 actors. It contains $1,494$ total samples: 518 from View 1, 509 from View 2, and 467 from View 3.

Toyota Smarthome dataset [57] is a real-world dataset that has 31 actions performed by 18 subjects. It contains $16,115$ samples collected from 7 viewpoints.

4.2 Implementation Details

In our experiments, we performed sampling using the RGB video sequence to build the ST-ROI. This sampling strategy not only reduced the large data volume of the RGB modality but also made the modality suitable for feature extraction with ordinary CNN models. Moreover, this approach enabled us to vary the input data via random selection. Given that we wished to obtain object and movement information from body areas including the hands, feet, and head, we set M' to 5 to construct the spatial sub-ST-ROI. For the temporal dimension, we empirically set the L to 5 to effectively cover variations in temporal appearance, as larger values of L will lead to redundant appearance information.

For the RGB modality, the height and width of sub-ST-ROIs of action sequences in NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD were each 96 pixels. For Northwestern-UCLA and Toyota Smarthome, the height and width of sub-ST-ROIs were 48 pixels, as data were collected with Kinect v1. Therefore, the input size for the NTU RGB+D 60, NTU RGB+D 120, PKU-MMD, Northwestern-UCLA, and Toyota Smarthome datasets were 480×480, 480×480, 480×480, 240×240, and 240×240, respectively. The ST-ROIs of the four datasets were resized to 225×225 and normalized before being fed into ResNet. As the data volume of the Northwestern-UCLA and Toyota Smarthome datasets were relatively small, we performed random selection on the RGB video frames and randomly flipped them. We adopted ResNet18, which has 18 layers, for all datasets. For NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD, we evenly selected frames based on the video length for training and testing.

For the submodel of the skeleton modality, we used the GCN implementation in [1], [2], and [4] for all datasets. The preprocessing method in [4] was also used for all datasets. To calculate spatial weights, we adopted the GCN model in [1]. Then, to alleviate the effect of smoothing out joint weights by the mean values of temporal positions, we empirically selected the top 15 valued temporal positions to calculate the joint weight for Equation 5.

The SGD optimizer was used for all implementations with the initial learning rate set to 0.1, which was divided by 10 at the 10th and 50th epochs. The training process was terminated at the 80th epoch. The minibatch size was set to 64. All experiments were conducted on a workstation with 4 GTX 1080 Ti GPUs.

TABLE 1
Ablation study for NTU RGB+D with X-Sub and X-View protocols.
* denotes our implementation. † uses the Kinect v2 2D skeleton.

#	Methods	X-Sub	X-View
1	Skeleton Joint [1]	80.4%	90.1%
2	Skeleton Bone [1]	84.4%	93.1%
3	Ensemble (#1+#2)	85.8%	93.3%
4	ST-ROI (No Joint Weights)	72.7%	81.3%
5	ST-ROI (Dynamic Weights)	73.8%	85.2%
6	ST-ROI (Fixed Weights)	76.8%	86.2%
7	ST-ROI (Fixed Weights)†	72.0%	75.4%
8	Ensemble (#3+#4)	90.7%	96.5%
9	Ensemble (#3+#5)	90.8%	96.6%
10	Ensemble (#3+#6)	91.2%	97.0%
11	2s-AGCN (Joint+Bone) [2]	88.5%	95.1%
12	MS-G3D (Joint+Bone) [4]	91.5%	96.2%
13	CTR-GCN* (Joint+Bone) [62]	92.2%	96.1%
14	Ensemble (#11+#6)	92.4%	97.3%
15	Ensemble (#12+#7)	92.7%	97.0%
16	Ensemble (#12+#6)	93.9%	**98.0%**
17	Ensemble (#13+#6)	**94.2%**	97.8%

4.3 Experiments on NTU RGB+D 60

The NTU RGB+D dataset provides two evaluation protocols, namely cross-subject (X-Sub) and cross-view (X-View) [8]. For the X-Sub protocol, half of the subjects were used for training and the other half were used for testing. For the X-View evaluation protocol, samples of $2/3$ viewpoints were used for training, and those of the remaining $1/3$ unseen viewpoints were used for testing. Table 1 shows the evaluation results of our ablation study based on the X-Sub and X-View evaluation protocols.

In Table 1, findings on different training strategies for the submodel of the RGB modality appear in rows #4, #5, and #6. The ensemble results of these training strategies are illustrated in rows #8, #9, and #10. We observed that the joint weights could improve submodel performance for the RGB modality, such that training with fixed joint weights outperformed training with dynamic weights. Moreover, the ensemble results of MMNet demonstrated consistent findings in rows #8, #9, and #10 compared with those in rows #4, #5, and #6. Furthermore, the ensemble results from rows #13 and #15 indicate that our MMNet could significantly improve the existing representative skeleton-based methods 2s-AGCN [2] and MS-G3D [4]. Precisely, our approach enhanced the results of 2s-AGCN [2] by 3.9% and 2.2% for the X-Sub and X-View evaluation protocols, respectively. It improved the results of MS-G3D [4] by 2.4% and 1.8% for the X-Sub and X-View evaluation protocols, respectively. It is also worth noting that, upon comparing rows #14 and #15 of Table 1, the implementation of ST-ROI with 2D skeleton data retrieved from OpenPose can alleviate noise in the 2D skeleton from the Kinect v2 sensor as this implementation achieves better performance.

To further verify the performance boost using the proposed MMNet, we calculated performance improvements

TABLE 2
Improvements in actions of NTU RGB+D that are difficult to address using skeleton-based methods.

	Action	#3	#3+#6	#11	#11+#6	#12	#12+#6
X-Sub (%)	11	43.6	61.5 (+17.9)	55.7	67.0 (+11.4)	71.1	77.7 (+6.6)
	12	46.0	65.8 (+19.9)	53.7	69.5 (+15.8)	57.4	68.4 (+11.0)
	30	59.3	78.5 (+19.3)	70.2	85.8 (+15.6)	76.4	88.7 (+12.4)
	31	73.6	84.1 (+10.5)	76.8	85.1 (+8.3)	85.9	88.4 (+2.5)
	53	91.7	96.4 (+4.7)	89.5	96.4 (+6.9)	92.8	97.1 (+4.3)
X-View (%)	11	67.0	75.9 (+8.9)	74.3	79.7 (+5.4)	83.8	89.2 (+5.4)
	12	65.1	85.4 (+20.3)	64.8	84.4 (+19.7)	72.1	85.4 (+13.3)
	30	69.9	93.4 (+23.4)	79.1	94.9 (+15.8)	82.9	97.5 (+14.6)
	31	95.2	97.5 (+2.2)	94.6	96.8 (+2.2)	95.6	97.5 (+1.9)
	53	92.4	97.5 (+5.1)	92.1	98.1 (+6.0)	96.5	98.7 (+2.2)

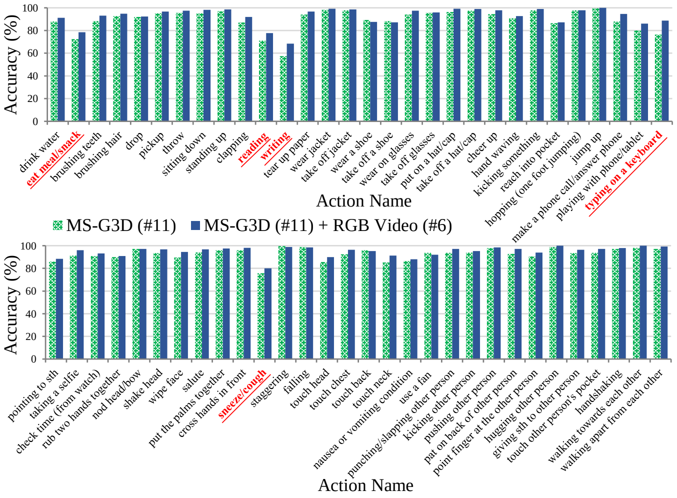


Fig. 6. Recognition accuracy per action on NTU RGB+D 60 (X-Sub). Action names in red and underlined are relatively more difficult actions.

for the challenging actions indicated in Fig. 1. Table 2 illustrates performance improvements for these actions compared with skeleton-based models #3, #11, and #12, corresponding to those in Table 1. Of note, the recognition accuracy for difficult actions on NTU RGB+D was substantially improved under the two evaluation protocols. Moreover, our MMNet enhanced the recognition accuracy of these actions and of other actions more generally (see Fig. 6).

Table 3 presents a comparison of MMNet with state-of-the-art methods on the NTU RGB+D dataset. Results show that our method greatly outperformed existing unimodal methods and performed competitive with existing multimodal methods. Regarding skeleton-based approaches, our method exceeded the performance of MS-G3D [4] by 2.4% and 1.8% for the X-Sub and X-View evaluation protocols, respectively. Regarding RGB video–based methods, our method outperformed Glimpse Clouds [44] by 7.3% and 4.8% for the X-Sub and X-View evaluation protocols, respectively. Among existing multimodal methods, our findings also surpassed the TSMF [22] by 1.4% and 0.6% for the X-Sub and X-View evaluation protocols, respectively. In addition to the vanilla implemention using ResNet18 in [22], we observed that the Inception-v3 [72] that factorized $n \times n$ convolution to $1 \times n$ and $n \times 1$ asymmetric convolutions can better represent the discontinuous data form of ST-ROI. Other more advanced backbones such as EfficientNet [73] and Swin-Transformer [74] can further improve the performance via our ST-ROI, which is ranked the second best among state-of-the-art methods. Compared with VPN++, our method achieves better performance on the larger version of NTU RGB+D 60 (i.e., NTU RGB+D 120) and Northwestern-UCLA Multiview (see Tables 7 and 11).

TABLE 3
Comparison of NTU RGB+D with X-Sub and X-View protocols. S and R denote skeleton and RGB modalities, respectively. Bold accuracy indicates the best. The second best is underlined.

Methods	S	R	X-Sub	X-View
Lie Group [63]	√	-	50.1%	52.8%
Dynamic Skeletons [64]	√	-	60.2%	65.2%
Part-aware LSTM [8]	√	-	62.9%	70.3%
ST-LSTM [9]	√	-	69.2%	77.7%
STA-LSTM [65]	√	-	73.4%	81.2%
GCA-LSTM [33]	√	-	74.4%	82.8%
View-invariant [31]	√	-	80.0%	87.2%
ST-GCN [1]	√	-	81.5%	88.3%
CNN-based [66]	√	-	83.2%	89.3%
DPRL+GCNN [67]	√	-	83.5%	89.8%
HCN [35]	√	-	86.5%	91.1%
2s-AGCN [2]	√	-	88.5%	95.1%
AGC-LSTM [3]	√	-	89.2%	95.0%
SRNet [68]	√	-	87.3%	91.3%
DGNN [36]	√	-	89.9%	96.1%
MS-G3D Net [4]	√	-	91.5%	96.2%
CTR-GCN [62]	√	-	92.4%	96.8%
C3D [7]	-	√	63.5%	70.3%
Glimpse Clouds [44]	-	√	86.6%	93.2%
ST-LSTM [9]	√	√	73.2%	80.6%
DSSCA - SSLM [49]	√	√	74.9%	-
STA-Hands [51]	√	√	82.5%	88.6%
Hands Attention [52]	√	√	84.8%	90.6%
S-Res-LSTM [69]	√	√	90.0%	96.3%
Body Pose Evolution Map [70]	√	√	91.7%	95.3%
TSMF [22]	√	√	92.5%	97.4%
VPN [53] (I3D)	√	√	93.5%	96.2%
VPN [53] (RNX3D101)	√	√	95.5%	98.0%
VPN++ + 3D Poses [71] (RNX3D101)	√	√	**96.6%**	**99.1%**
Our MMNet (ResNet18, $L = 5$)	√	√	94.2%	97.8%
Our MMNet (Inception-v3, $L = 5$)	√	√	95.3%	98.4%
Our MMNet (Swin-Transformer-B, $L = 5$)	√	√	95.6%	98.7%
Our MMNet (EfficientNet-B7, $L = 5$)	√	√	<u>96.0%</u>	<u>98.8%</u>

4.4 Experiments on NTU RGB+D 120

The NTU RGB+D 120 dataset provides two evaluation protocols: cross-subject (X-Sub) and cross-setup (X-Set) [10]. For the X-Sub protocol, $63,026$ samples collected from 53 subjects were used for training while the remaining $50,919$ samples were used for testing. For the X-Set protocol, $54,468$ samples from the first half of camera setups were used for training, and $59,477$ samples from the second half of camera setups were applied in testing. Table 5 lists the evaluation results of our ablation study with the X-Sub and X-Set evaluation protocols.

In Table 5, findings based on different training strategies for the submodel of the RGB modality are displayed in rows #4, #5, and #6. The ensemble results for these training strategies appear in rows #8, #9, and #10. The joint weights were found to improve submodel performance for the RGB modality, such that training with fixed joint weights outperformed training with dynamic weights. Moreover, the ensemble results for MMNet were consistent in rows #8, #9, and #10 compared with findings in rows #4, #5, and #6. Furthermore, the ensemble results in rows #13 and #15 indicate that our MMNet could significantly enhance the

TABLE 4
Action recognition improvements on NTU RGB+D 120 dataset by aggregating the results of ensemble (#12+#6) compared with the top 10 accurate and confused actions of skeleton-based method MS-G3D.

Protocol	Top 10 accurate actions (ID)	MS-G3D	#12+#6	Top 10 confused actions (ID)	MS-G3D	#12+#6
X-Sub	1. walking towards each other (59)	100.0%	100.0% (+0.0%)	1. staple book (73)	34.9%	35.6% **(+0.7%)**
	2. jump up (27)	99.6%	100.0% **(+0.4%)**	2. counting money (74)	57.0%	58.1% **(+1.1%)**
	3. staggering (42)	99.6%	99.6% (+0.0%)	3. make victory sign (72)	59.7%	59.7% (+0.0%)
	4. arm swings (98)	99.5%	99.5% (+0.0%)	4. make OK sign (71)	60.9%	63.1% **(+2.3%)**
	5. hugging other person (55)	99.3%	100.0% **(+0.7%)**	5. writing (12)	62.5%	78.7% **(+16.2%)**
	6. cheers and drink (113)	99.1%	99.7% **(+0.5%)**	6. playing with phone/tablet (29)	67.3%	87.3% **(+20.0%)**
	7. wear jacket (14)	98.9%	100.0% **(+1.1%)**	7. hit with object (106)	67.7%	74.1% **(+6.4%)**
	8. high-five (112)	98.8%	99.3% **(+0.5%)**	8. cutting nails (75)	68.9%	85.1% **(+16.2%)**
	9. falling (43)	98.5%	99.3% **(+0.7%)**	9. cutting paper (76)	70.5%	79.2% **(+8.7%)**
	10. arm circles (97)	98.4%	99.0% **(+0.5%)**	10. blow nose (105)	71.1%	80.5% **(+9.4%)**
X-Set	1. walking towards each other (59)	98.8%	99.8% **(+1.0%)**	1. staple book (73)	57.0%	56.4% (-0.6%)
	2. wear jacket (14)	98.6%	99.2% **(+0.6%)**	2. writing (12)	60.0%	78.1% **(+18.1%)**
	3. standing up (9)	98.6%	99.6% **(+1.0%)**	3. cutting paper (76)	62.3%	71.5% **(+9.2%)**
	4. nod head/bow (35)	98.4%	99.2% **(+0.8%)**	4. make victory sign (72)	66.7%	67.8% **(+1.0%)**
	5. hopping (one foot jumping) (26)	98.4%	99.6% **(+1.2%)**	5. counting money (74)	67.3%	66.3% (-1.0%)
	6. arm circles (97)	98.2%	99.4% **(+1.2%)**	6. reading (11)	68.4%	74.6% **(+6.2%)**
	7. staggering (42)	98.0%	99.2% **(+1.2%)**	7. yawn (103)	69.3%	82.1% **(+12.8%)**
	8. cheers and drink (113)	98.0%	98.0% (+0.0%)	8. cutting nails (75)	71.5%	83.4% **(+11.9%)**
	9. cross toe touch (101)	97.8%	99.8% **(+2.0%)**	9. make OK sign (71)	72.0%	71.6% (-0.4%)
	10. arm swings (98)	97.8%	99.2% **(+1.4%)**	10. blow nose (105)	72.1%	86.4% **(+14.3%)**

TABLE 5
Ablation study for NTU RGB+D 120 with X-Sub and X-Set protocols. * denotes our implementation. † uses the Kinect v2 2D skeleton.

#	**Methods**	**X-Sub**	**X-Set**
1	Skeleton Joint [1]	79.0%	81.3%
2	Skeleton Bone [1]	81.0%	82.4%
3	Ensemble (#1+#2)	83.5%	85.2%
4	ST-ROI (No Joint Weights)	67.2%	71.7%
5	ST-ROI (Dynamic Weights)	69.7%	74.2%
6	ST-ROI (Fixed Weights)	71.7%	74.4%
7	ST-ROI (Fixed Weights)†	60.3%	58.8%
8	Ensemble (#3+#4)	88.2%	90.5%
9	Ensemble (#3+#5)	88.3%	90.5%
10	Ensemble (#3+#6)	88.6%	90.7%
11	2s-AGCN* (Joint+Bone) [2]	84.2%	86.0%
12	MS-G3D* (Joint+Bone) [4]	87.2%	88.4%
13	Ensemble (#11+#6)	88.9%	91.0%
14	Ensemble (#12+#7)	88.9%	89.7%
15	Ensemble (#12+#6)	**90.3%**	**92.1%**

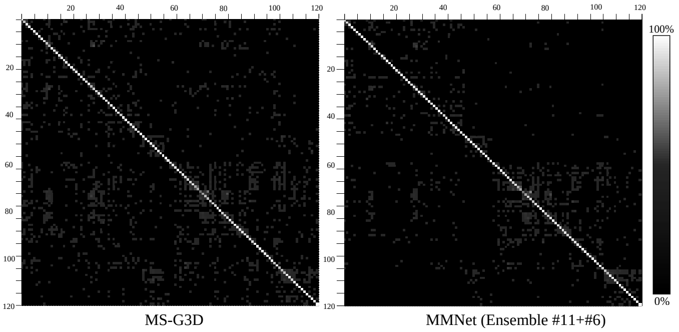


Fig. 7. Confusion matrices of MS-G3D and ensemble (#12+#6) on NTU RGB+D 120 with X-Sub protocol. Darker color in off-diagonal areas on the right side confusion matrix comparing with the left one indicates the improvements.

representative skeleton-based methods 2s-AGCN [2] and MS-G3D [4]. More precisely, our method improved the results of 2s-AGCN [2] by 3.9% and 2.2% for the X-Sub and X-View evaluation protocols, respectively. It improved the results of MS-G3D [4] by 2.4% and 1.8% for the X-Sub and X-View evaluation protocols, respectively. The results of implementing ST-ROI with 2D skeleton data from Kinect v2 (see rows #7 and #14 of Table 5) is consistent with those for NTU RGB+D 60.

As Table 6 shows, we conducted further ablation on NTU RGB+D 120 to compare our method with VPN [53] regarding inference time, the numbers of model parameters, and floating point operations (FLOPs). We tested $1,000$ on a single GTX 1080 Ti with the batch size of 1, and reported the average inference time, which also includes the processing time of OpenPose [54] tool (36ms per RGB frame). We used fvcore[1] to calculate the FLOPs. Although our submodel for the RGB modality implemented with ResNet18 does not perform as well as I3D, it can effectively contribute to the ensemble results with smaller number of model parameters. More precisely, the simplest version of our method (i.e., the method in row #8 of Table 6, which only uses the skeleton joint stream and the basic GCN model) can perform better than the state-of-the-art multimodal method VPN. It is also worth noting that VPN relies on the video-based model I3D which requires 64 RGB video frames; our method uses the relatively smaller model ResNet18 and requires only 5 RGB video frames. Compared with VPN, our MMNet is relatively more lightweight and achieves better performance with shorter inference time (see Table 6). Without considering the models size, our MMNet can be further improved by implementing the RGB modality with Inception-v3. Regarding the weakness of our approach, our MMNet does not rely on video-based methods that incorporate features into the

[1] fvcore: https://github.com/facebookresearch/fvcore.git

TABLE 6
Ablation study for comparison with VPN on NTU RGB+D 120 under X-Sub and X-Set protocols. L and M' indicate the numbers of RGB frames and body parts, respectively. Better performances are in bold based on using skeleton joint only (i.e., no skeleton bone).

#	Methods	Backbone	L	M'	Inference Time	Parameters	FLOPs	X-Sub	X-Set
1	Skeleton Joint	GCN [1]	-	-	0.013s	3.1M	17.2G	79.0%	81.3%
2	Skeleton Bone	GCN [1]	-	-	0.013s	3.1M	17.2G	81.0%	82.4%
3	Ensemble (#1+#2)	GCN [1]	-	-	0.026s	6.2M	34.4G	83.5%	85.2%
4	ST-ROI (Fixed Weights)	ResNet18 [21]	5	5	0.270s	14.4M	19.2G	71.7%	74.4%
5	ST-ROI (Fixed Weights)	Inception-v3 [72]	5	5	0.329s	27.8M	23.0G	79.9%	82.0%
6	RGB Video [53]	I3D [5]	64	25	0.3s	12.1M	107.9G	77.0%	80.1%
7	VPN [53]	GCNs+I3D	64	25	65s	24.0M	-	86.3%	87.8%
8	Ensemble (#1+#4)	GCN+ResNet18	5	5	0.283s	14.4M	19.2G	**86.6%**	**88.7%**
9	Ensemble (#2+#4)	GCN+ResNet18	5	5	0.283s	17.5M	36.4G	87.1%	89.4%
10	Ensemble (#3+#4)	GCN+ResNet18	5	5	0.296s	17.5M	36.4G	88.6%	90.7%
11	Ensemble (#3+#5)	GCN+Inception-v3	5	5	0.355s	30.9M	40.2G	91.5%	93.2%
12	TSMF [22]	MS-G3D+ResNet18	5	5	0.359s	20.8M	85.4G	87.0%	89.1%
13	MMNet (Inception-v3)	MS-G3D+Inception-v3	5	5	0.418s	34.2M	89.2G	92.9%	94.4%

TABLE 7
Comparison of NTU RGB+D 120 with X-Sub and X-Set protocols. S and R denote skeleton and RGB modalities, respectively. * denotes our implementation.

Methods	S	R	X-Sub	X-Set
Spatiotemporal LSTM [75]	√	-	55.7%	57.9%
Internal Feature Fusion [9]	√	-	58.2%	60.9%
GCA-LSTM [9]	√	-	58.3%	59.2%
Multi-Task Learning Network [76]	√	-	58.4%	57.9%
FSNet [77]	√	-	59.9%	62.4%
ST-GCN* (Joint+Bone) [1]	√	-	83.5%	85.2%
2s-AGCN* (Joint+Bone) [2]	√	-	84.2%	86.0%
MS-G3D [4]	√	-	86.9%	88.4%
CTR-GCN [62]	√	-	88.9%	90.6%
Baseline [10]	√	√	61.2%	63.1%
Two-Stream Attention LSTM [78]	√	√	61.2%	63.3%
Multi-Task CNN with RotClips [79]	√	√	62.2%	61.8%
VPN [53]	√	√	86.3%	87.8%
TSMF [22]	√	√	87.0%	89.1%
VPN++ + 3D Poses [71]	√	√	90.7%	92.5%
Our MMNet (ResNet18, $L = 5$)	√	√	90.3%	92.1%
Our MMNet (Inception-v3, $L = 5$)	√	√	**92.9%**	**94.4%**

TABLE 8
Ablation study for PKU-MMD with X-Sub and X-View protocols. * denotes our implementation.

#	Methods	X-Sub	X-View
1	Skeleton Joint [1]	91.5%	92.4%
2	Skeleton Bone [1]	93.4%	95.1%
3	Ensemble (#1+#2)	94.6%	96.3%
4	ST-ROI (No Joint Weights)	81.3%	77.4%
5	ST-ROI (Dynamic Weights)	81.6%	76.2%
6	ST-ROI (Fixed Weights)	83.0%	82.2%
7	Ensemble (#3+#4)	95.8%	97.1%
8	Ensemble (#3+#5)	95.9%	97.2%
9	Ensemble (#3+#6)	96.0%	97.5%
10	2s-AGCN* (Joint+Bone) [2]	94.7%	96.8%
11	MS-G3D* (Joint+Bone) [4]	95.5%	97.1%
12	Ensemble (#10+#6)	96.1%	97.8%
13	Ensemble (#11+#6)	**96.3%**	**98.0%**

background scenes. Our method thus has limitations similar to skeleton-based methods for outdoor actions according to our discussion in Section 5. VPN relies on the I3D backbone, which makes it possible to improve performance on the Kinetics dataset.

Following the analysis in [10], we plotted a confusion matrix to analyze the effectiveness of our method. Fig. 7 depicts the confusion matrices corresponding to results for the MS-G3D and ensemble (#12+#6) methods on NTU RGB+D 120 with the X-Sub evaluation protocol. Improvements were apparent across several areas of the two confusion matrices. To extend our analysis based on the confusion matrices in Fig. 7, we conducted action-wise analysis for the proposed MMNet. In particular, we analyzed the top 10 actions that were accurately recognized and the top 10 actions that confused the state-of-the-art skeleton-based model (i.e., MS-G3D); results are listed in Table 4. Although the top 10 recognized actions exhibited high recognition accuracy, our MMNet could further improve highly recognized actions in the X-Sub and X-Set evaluation protocols. Based on the top 10 confused actions in Table 4, we found that the recognition accuracy for Actions 11 and 12 (i.e., "reading" and "writing") improved substantially when using our proposed MMNet. Most other confused actions, such as "playing with phone/tablet," "cutting nails," "yawn," and "blow nose," were also recognized with significant improvements. The relatively lower (or nonexistent) improvement associated with actions such as "staple book" and "counting money" could be due to a lack of discriminative features in the skeleton and RGB video modalities. For other challenging actions such as "make victory sign" and "make OK sign," limited or nonexistent improvement may have occurred because these actions are more fine-grained and require higher resolution in the RGB video modality for recognition.

Table 7 shows a comparison of our method with state-of-the-art approaches on NTU RGB+D 120. Our approach greatly outperformed existing unimodal and multimodal methods. Regarding skeleton-based methods, our method exceeded the state-of-the-art performance of CTR-GCN [62] by 4.0% and 3.8% for the X-Sub and X-Set evaluation protocols, respectively. Regarding multimodal methods, our method outperformed VPN++ [71] by 2.2% and 1.9% for the X-Sub and X-Set evaluation protocols, respectively.

4.5 Experiments on PKU-MMD

The PKU-MMD dataset [23] provides evaluation protocols similar to NTU RGB+D 60, specifically cross-subject (X-Sub)

TABLE 9
Comparison of PKU-MMD with X-Sub and X-View protocols. S and R denote skeleton and RGB modalities, respectively.

Methods	**S**	**R**	**X-Sub**	**X-View**
JCRRNN [80]	√	-	32.5%	53.3%
Skeleton boxes [81]	√	-	54.8%	94.2%
STA-LSTM [65]	√	-	86.9%	92.6%
CNN-based [66]	√	-	90.4%	93.7%
HCN [35]	√	-	92.6%	94.2%
SRNet [68]	√	-	93.1%	97.0%
TSMF [22]	√	√	95.8%	97.8%
Our MMNet (ResNet18, $L = 5$)	√	√	96.3%	98.0%
Our MMNet (Inception-v3, $L = 5$)	√	√	97.2%	98.1%
Our MMNet (Swin-Transformer-B, $L = 5$)	√	√	97.3%	98.1%
Our MMNet (EfficientNet-B7, $L = 5$)	√	√	**97.4%**	**98.6%**

and cross-view (X-View). As Table 8 shows, the ablation study for PKU-MMD revealed consistent results compared with NTU RGB+D and NTU RGB+D 120 as illustrated in Table 1 and Table 5, respectively. The PKU-MMD, NTU RGB+D 60, and NTU RGB+D 120 datasets were similar, as they were collected using the same sensor and share the same data characteristics.

Compared with existing methods, our MMNet appeared to achieve the best performance on PKU-MMD under the X-Sub and X-View evaluation protocols. Table 9 presents a comparison based on PKU-MMD with state-of-the-art methods. Given that our previous version in [22] already achieved high recognition accuracy under the X-Sub and X-View protocols, our method continued to boost the accuracy in the current study to 97.4% and 98.6% for the X-Sub and X-View protocols, respectively.

4.6 Experiments on Northwestern-UCLA Multiview

Northwestern-UCLA Multiview were gathered via Kinect v1 from three views. We followed the cross-view evaluation protocols defined by [24]. Table 10 shows an ablation study using the Northwestern-UCLA Multiview dataset with three cross-view evaluation protocols: $V^3_{1,2}$, $V^2_{1,3}$, and $V^1_{2,3}$. Here, $V^3_{1,2}$ indicates the use of samples from the first two views for training, whereas samples in the third view were used for testing. The results in Table 10 are consistent with those of the other three datasets examined in this paper. Notably, the recognition accuracy for Northwestern-UCLA Multiview data was not as strong overall as for the other three larger datasets. Essentially, our method is data-driven and relies on a large amount of training data. Another interesting finding is that the performance on $V^1_{2,3}$ was not as good as for the other two evaluation protocols. We found that View 1 was in the middle of the data collection environment, while Views 2 and 3 were at either side of the data collection environment. This difference may explain the dataset shift between the training and test cases for $V^1_{2,3}$ and thus why recognition was more challenging than for the other two evaluation protocols.

TABLE 10
Ablation study for Northwestern-UCLA Multiview with three cross-view settings. * denotes our implementation.

#	**Methods**	$\boldsymbol{V^3_{1,2}}$	$\boldsymbol{V^2_{1,3}}$	$\boldsymbol{V^1_{2,3}}$
1	Skeleton Joint [1]	84.2%	82.1%	69.2%
2	Skeleton Bone [1]	80.8%	83.5%	70.9%
3	Ensemble (#1+#2)	86.2%	83.9%	74.0%
4	ST-ROI (No Joint Weights)	39.7%	50.1%	35.5%
5	ST-ROI (Dynamic Weights)	53.1%	20.9%	15.9%
6	ST-ROI (Fixed Weights)	61.8%	70.2%	49.4%
7	Ensemble (#3+#4)	87.3%	85.1%	73.5%
8	Ensemble (#3+#5)	86.8%	84.7%	74.4%
9	Ensemble (#3+#6)	87.7%	85.1%	76.6%
10	2s-AGCN* (Joint+Bone) [2]	87.3%	81.3%	69.2%
11	MS-G3D* (Joint+Bone) [4]	92.7%	89.7%	82.4%
12	Ensemble (#10+#6)	88.8%	82.9%	73.3%
13	Ensemble (#11+#6)	**93.3%**	**91.1%**	**83.7%**

TABLE 11
Comparison of Northwestern-UCLA Multiview with three cross-view settings. S and R denote skeleton and RGB modalities, respectively.

Methods	**S**	**R**	$\boldsymbol{V^3_{1,2}}$	$\boldsymbol{V^2_{1,3}}$	$\boldsymbol{V^1_{2,3}}$
Lie Group [63]	√	-	74.2%	-	-
HBRNN-L [82]	√	-	78.5%	-	-
View-invariant [31]	√	-	86.1%	-	-
Ensemble TS-LSTM [83]	√	-	89.2%	-	-
AGC-LSTM [3]	√	-	93.3%	-	-
CTR-GCN [62]	√	-	**96.5%**	-	-
Hankelets [84]	-	√	45.2%	-	-
nCTE [85]	-	√	68.6%	68.3%	52.1%
NKTM [86]	-	√	75.8%	73.3%	59.1%
Glimpse Clouds [44]	-	√	90.1%	89.5%	83.4%
VPN [53]	√	√	93.5%	-	-
VPN++ + 3D Poses [71]	√	√	93.5%	-	-
Our MMNet (ResNet18, $L = 5$)	√	√	93.3%	**91.1%**	**83.7%**
Our MMNet (ResNet18, $L = 7$)	√	√	$\underline{93.7\%}$	89.9%	82.6%

Table 11 presents a comparison of the Northwestern-UCLA Multiview dataset with state-of-the-art methods. These results appear to verify the effectiveness of MMNet, which achieved state-of-the-art recognition accuracy for the last two cross-view protocols and the second best for the first protocol. The improvement over existing methods was not as significant as for the other three datasets because Northwestern-UCLA Multiview had an insufficient size for our data-driven method. Additionally, the dataset was gathered with the Kinect v1 sensor, which could provide neither skeleton data as accurately as Kinect v2 nor RGB video data at a resolution as high as Kinect v2. Based on this cross-dataset comparison, using Kinect v2 for data collection together with a larger dataset could influence action recognition accuracy in RGB-D videos.

4.7 Experiments on Toyota Smarthome

The Toyota Smarthome dataset provides three evaluation protocols: cross-subject (CS), cross-view 1 (CV_1) and cross-view 2 (CV_2) [57]. Following the three evaluation protocols in [57], we conducted ablations as shown in Table 12. The ablation study of evaluation protocol CS shows consistent results with those of other datasets. For evaluation protocols CV_1 and CV_2, advanced GCN models such as 2s-AGCN and CTR-GCN cannot gain single modal performance improvements due to the small amount of training data. Note that, excluding empty skeleton samples, there are $1,877$ and $7,735$ training samples in evaluation protocols CV_1 and CV_2, respectively. While evaluation protocol CS has $10,614$ training samples, which is relatively more on par with those of other datasets.

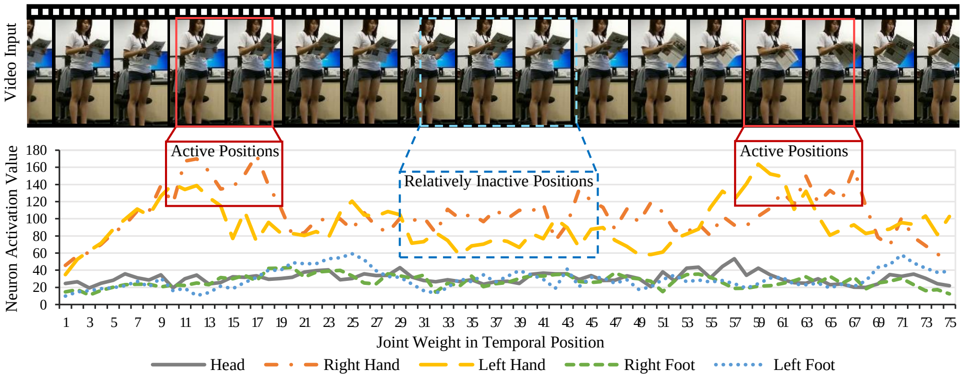


Fig. 8. Visualization of neuron activation values for different skeleton joints along their temporal positions and cropped sample frames of video input. This visualization shows the idea of selecting top-t positions to calculate joint weights for their corresponding body areas.

TABLE 12
Ablation study for Toyota Smarthome with three evaluation protocols. * denotes our implementation. † uses the original 2D skeleton.

#	Methods	CS	CV_1	CV_2
1	Skeleton Joint [1]	66.6%	44.1%	53.8%
2	Skeleton Bone [1]	66.3%	36.5%	53.1%
3	Ensemble (#1+#2)	70.8%	44.4%	59.1%
4	ST-ROI (No Joint Weights)	54.6%	47.8%	46.0%
5	ST-ROI (Dynamic Weights)	56.7%	35.3%	26.2%
6	ST-ROI (Fixed Weights)	60.2%	39.1%	28.1%
7	ST-ROI (Fixed Weights)†	57.0%	37.0%	32.3%
8	Ensemble (#3+#4)	75.3%	52.2%	62.4%
9	Ensemble (#3+#5)	75.3%	44.5%	60.2%
10	Ensemble (#3+#6)	76.7%	48.8%	60.7%
11	2s-AGCN* (Joint+Bone) [2]	71.3%	42.4%	53.2%
12	MS-G3D* (Joint+Bone) [4]	71.1%	37.0%	54.6%
13	CTR-GCN* (Joint+Bone) [62]	79.9%	**64.7%**	62.1%
14	Ensemble (#11+#6)	76.5%	47.7%	57.1%
15	Ensemble (#12+#7)	74.7%	41.7%	57.7%
16	Ensemble (#12+#6)	77.5%	47.8%	57.4%
17	Ensemble (#13+#6)	**82.1%**	58.5%	**62.9%**

Compared with existing methods, as shown in Table 13, our MMNet achieved the second best performance under the CS evaluation protocol by using the vision backbone EfficientNet-V2-L [87]. Improvements of per-class accuracy are illustrated in Figure 9. For the other two evaluation protocols, our method also achieved competitive performance.

4.8 Analysis of Joint Weights

The GCN submodel can learn the importance of skeleton joints at a specific time, meaning that the time-specific effect on a skeleton joint fluctuates during the progression of an action. Fig. 8 shows changes in this time-specific effect of different skeleton joints, reflecting the importance of a node on the graph. As depicted, a specific action is associated with active and inactive positions on the temporal dimension. Inactive positions indicate that the subject has performed the action and is in an almost static state. For the action in Fig. 8, some skeleton joints (e.g., both hands) are more active (i.e., have higher neuron activation values) than the foot areas. Directly taking the mean value of all temporal positions to compute Equation 5 will smooth out the time-specific effect, as inactive positions will affect the active positions used to calculate the joint-specific weight.

TABLE 13
Comparison of Toyota Smarthome with three evaluation protocols. S and R denote skeleton and RGB modalities, respectively. Results are mean per-class accuracy. The second best is underlined.

Methods	S	R	CS	CV_1	CV_2
5C-AGCN+SSTA-PRS [88]	√	-	62.1%	22.8%	54.0%
I3D [84]	-	√	53.4%	34.9%	45.1%
AssembleNet++ [89]	-	√	63.6%	-	-
TSMF (Pose_V1.2) [22]	√	√	53.8%	16.9%	28.9%
VPN (Pose_V1.1) [53]	√	√	60.8%	**43.8%**	53.1%
VPN (Pose_V1.2) [53]	√	√	65.2%	-	54.1%
VPN++ + Poses (Pose_V1.2) [53]	√	√	**71.0%**	-	**58.1%**
Our MMNet (Pose_V1.2, ResNet18, $L = 5$)	√	√	65.1%	27.4%	33.4%
Our MMNet (Pose_V1.2, EfficientNet-V2-L, $L = 5$)	√	√	<u>70.1%</u>	<u>37.4%</u>	46.6%

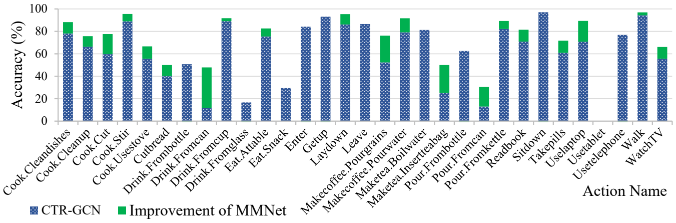


Fig. 9. Improvement of our MMNet over baseline model CTR-GCN on the Toyota Smarthome dataset (21 of 31 actions are improved).

The implementation of the joint weight in Equation 5 can be empirically determined based on the top 15 valued positions. Intuitively, computing a fused representation at a fine-grained temporal level (i.e., a time-specific structural-appearance connection) could be an effective strategy; however, we found that that the performance was poor because temporal positions for the RGB and skeleton modalities were not ideally associated. Thus, we implemented the top 15 valued positions of a skeleton joint for a specific body part. Table 14 shows a comparison of different numbers of top valued temporal positions, where we tested selections from the top 5 to top 25 skeleton joint positions with an interval of 5. Results for the top 15 positions tended to be the most empirically compelling.

4.9 Analysis of Skeleton-Focused Representation

For the RGB video modality, our MMNet could focus on varied lengths of RGB video frames, which could influence performance. On one hand, taking a larger number of

TABLE 14
Comparison of results when selecting different top-t valued temporal positions in GCN features to retrieve joint weights.

Top-t	NTU 60		NTU 120		PKU-MMD		N-UCLA Multiview			Toyota Smarthome		
	X-Sub	X-View	X-Sub	X-Set	X-Sub	X-View	$V_{1,2}^3$	$V_{1,3}^2$	$V_{2,3}^1$	CS	CV_1	CV_2
$t = 5$	93.5%	97.2%	89.6%	91.8%	95.4%	97.6%	93.1%	90.0%	82.1%	76.2%	37.5%	55.2%
$t = 10$	93.6%	97.9%	90.0%	**92.2%**	95.8%	**98.0%**	**93.5%**	90.3%	82.4%	76.2%	38.4%	57.0%
$t = 15$	**93.9%**	**98.0%**	90.3%	92.1%	**96.3%**	**98.0%**	93.3%	**91.1%**	**83.7%**	**77.5%**	**47.8%**	**57.4%**
$t = 20$	93.6%	97.7%	**90.5%**	91.8%	**96.3%**	97.8%	93.3%	90.3%	83.0%	76.3%	45.4%	55.5%
$t = 25$	93.8%	97.7%	90.3%	91.8%	95.9%	97.1%	93.2%	90.1%	83.2%	76.2%	45.6%	55.3%

TABLE 15
Comparison of results when selecting different numbers of RGB frames to construct the ST-ROI.

L	NTU 60		NTU 120		PKU-MMD		N-UCLA Multiview			Toyota Smarthome		
	X-Sub	X-View	X-Sub	X-Set	X-Sub	X-View	$V_{1,2}^3$	$V_{1,3}^2$	$V_{2,3}^1$	CS	CV_1	CV_2
$L = 1$	92.6%	97.3%	89.2%	90.7%	95.8%	97.2%	92.1%	89.4%	82.1%	75.4%	37.7%	56.3%
$L = 3$	92.9%	97.6%	89.7%	91.5%	96.1%	97.9%	93.1%	90.1%	82.6%	76.1%	38.4%	56.9%
$L = 5$	93.9%	**98.0%**	90.3%	**92.1%**	**96.3%**	98.0%	93.3%	**91.1%**	**83.7%**	**77.5%**	**47.8%**	57.4%
$L = 7$	93.5%	97.7%	**90.4%**	92.0%	96.1%	**98.2%**	**93.7%**	89.9%	82.6%	76.8%	40.0%	54.0%
$L = 9$	**94.1%**	97.5%	**90.4%**	91.8%	96.2%	98.0%	92.9%	90.1%	82.6%	76.1%	39.5%	**59.8%**

frames could lead to redundant features for the RGB modality and cause the model to struggle to focus on important frames and variance in this modality. On the other hand, using fewer frames could prevent the model from capturing useful features. Fig. 10 displays different potential choices of RGB frames.

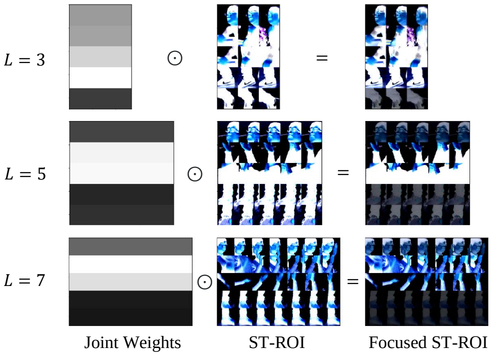


Fig. 10. Visualization of skeleton-focused representations with different sampling lengths of the RGB modality (ST-ROI is normalized).

We conducted experiments to investigate options for the number of RGB video frames used to construct the ST-ROI. Table 15 shows experimental results with different values of L. Findings reveal that, when using 5 RGB video frames, the proposed MMNet generally achieved competitive recognition accuracy. Occasionally, the model performed better with different options for L (e.g., 98.2% for PKU X-View and 93.7% for N-UCLA $V_{1,2}^3$), but the enhancement was not as noteworthy compared with when $L = 5$.

5 Discussion

As discussed in Section 2.1.2, indoor actions vary considerably from outdoor actions regarding feature differences in their background scenes. To further explore whether our model design can be validated on outdoor actions, we conduct experiments on the Kinetics 400 dataset [40], which is based on OpenPose 2D skeleton samples available in [1] and downloadable videos of Kinetics 400 [40] at the time experiments were performed. Table 16 shows that the results on Kinetics 400 coincide with those of the other four datasets considered in this research, further confirming that the design of our MMNet is reasonable: this approach can effectively alleviate the lack of appearance features in the skeleton modality.

The state-of-the-art accuracy of the skeleton-based method on Kinetics 400 lags far behind that of RGB video-based methods, such that the former is achieved by MS-G3D [4] at an accuracy of 38.5% and the latter is achieved by a series of SlowFast models [46] (the highest accuracy is 79.8%). Video-based methods also cannot perform as competitively as skeleton-based methods on indoor actions in NTU RGB+D. It is worth noting that background scenes contribute to the recognition of outdoor actions in Kinetics 400. For example, [45] and [46] each indicated that the background scenes of some actions (e.g., “playing tennis”, “playing badminton”, “playing cricket”) in Kinetics 400 play an important role in recognition. Our MMNet is designed for indoor actions in RGB-D videos, where background scene information is not used. This circumstance leads to a limitation similar to skeleton-based methods when recognizing outdoor actions.

6 Conclusion

We have proposed a multimodal DL architecture called MMNet for HAR in RGB-D videos using a model-based multimodal data fusion mechanism. This method borrows the attention feature from the skeleton modality and contributes to the RGB modality’s performance, thus enhancing ultimate ensemble performance. The proposed MMNet achieved very competitive performance on five representative large datasets (NTU RGB+D 60/120, PKU-MMD, Northwestern-UCLA Multiview, and Toyota Smarthome) compared with skeleton-based, RGB video–based, and multimodal methods. The results of the RGB modality when

TABLE 16
Ablation study for the Kinetics 400 dataset.

#	Methods	Top-1
1	MS-G3D (Skeleton Joint) [4]	36.4%
2	MS-G3D (Skeleton Bone) [4]	36.0%
3	Ensemble (#1+#2)	38.5%
4	ST-ROI (No Joint Weights)	21.7%
5	ST-ROI (Dynamic Weights)	22.8%
6	ST-ROI (Fixed Weights)	23.3%
7	Ensemble (#1+#6)	40.7%
8	Ensemble (#3+#4)	42.7%
9	Ensemble (#3+#5)	43.0%
10	Ensemble (#3+#6)	43.5%

using a fixed attention mechanism were better than that using dynamic weights and performed better in terms of ensemble results when aggregated with findings from the skeleton modality.

In the future, we intend to further investigate other aspects (e.g., depth and optical flow streams) that can affect the performance of multimodal HAR by designing architectures with more prior knowledge and by making our models more explainable and improvable. Additionally, for outdoor actions, we will work on incorporating background scene information to expand our method to more challenging real-world datasets such as Kinetics [40].

Acknowledgments

The work is supported by The Hong Kong Polytechnic University for the Project: Music Therapy for Dementia Prevention Using Algorithmic Composition under Grant No.: P0008740. The author also would like to thank anonymous reviewers who help improve this manuscript.

References

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *32nd AAAI conference on artificial intelligence*, 2018, Conference Proceedings.

[2] 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 Conference on Computer Vision and Pattern Recognition*, 2019, Conference Proceedings, pp. 12 026–12 035.

[3] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, Conference Proceedings, pp. 1227–1236.

[4] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, Conference Proceedings, pp. 143–152.

[5] 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, Conference Proceedings, pp. 6299–6308.

[6] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, Conference Proceedings, pp. 305–321.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, Conference Proceedings, pp. 4489–4497.

[8] 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, Conference Proceedings, pp. 1010–1019.

[9] J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3007–3021, 2017.

[10] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, pp. 2684–2701, 2019.

[11] H. Sagha, S. T. Digumarti, J. d. R. Millán, R. Chavarriaga, A. Calatroni, D. Roggen, and G. Tröster, "Benchmarking classification techniques using the opportunity human activity dataset," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*. IEEE, 2011, Conference Proceedings, pp. 36–40.

[12] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, Conference Proceedings, pp. 168–172.

[13] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[14] W. Elmenreich, "An introduction to sensor fusion," *Vienna University of Technology, Austria*, vol. 502, pp. 1–28, 2002.

[15] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1165–1179, 2017.

[16] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[17] B. Pan, J. Sun, W. Lin, L. Wang, and W. Lin, "Cross-stream selective networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, Conference Proceedings, pp. 0–0.

[18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[19] Y. Zhang, P. Tokmakov, M. Hebert, and C. Schmid, "A structured model for action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, Conference Proceedings, pp. 9975–9984.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] 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, Conference Proceedings, pp. 770–778.

[22] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher-student network for indoor action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3199–3207.

[23] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for skeleton-based human action understanding," in *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*. ACM, 2017, Conference Proceedings, pp. 1–8.

[24] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, Conference Proceedings, pp. 2649–2656.

[25] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[26] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2013.

[27] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[28] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[29] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[30] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[31] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[32] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, Conference Proceedings, pp. 2117–2126.

[33] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, Conference Proceedings, pp. 1647–1656.

[34] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, Conference Proceedings, pp. 3595–3603.

[35] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, Conference Proceedings, pp. 786–792.

[36] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, Conference Proceedings, pp. 7912–7921.

[37] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," *arXiv preprint arXiv:1711.09561*, 2017.

[38] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, Conference Proceedings, pp. 2556–2563.

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

[41] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, Conference Proceedings, pp. 214–223.

[42] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, Conference Proceedings, pp. 4768–4777.

[43] Z. Luo, J.-T. Hsieh, L. Jiang, J. Carlos Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, Conference Proceedings, pp. 166–183.

[44] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, Conference Proceedings, pp. 469–478.

[45] B. Martinez, D. Modolo, Y. Xiong, and J. Tighe, "Action recognition with spatial-temporal discriminative filter banks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5482–5491.

[46] J. Weng, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, X. Jiang, and J. Yuan, "Temporal distinct representation learning for action recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 363–378.

[47] J. Choi, C. Gao, J. C. Messou, and J.-B. Huang, "Why can't i dance in the mall? learning to mitigate scene bias in action recognition," in *Advances in Neural Information Processing Systems*, 2019, Conference Proceedings, pp. 851–863.

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

[49] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+ d videos," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[50] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[51] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, Conference Proceedings, pp. 604–613.

[52] F. Baradel, C. Wolf, and J. Mille, "Human activity recognition with pose-driven attention to rgb," in *BMVC 2018 - 29th British Machine Vision Conference*, 2018, Conference Proceedings, pp. pp.1–14.

[53] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in *European Conference on Computer Vision*. Springer, 2020, pp. 72–90.

[54] 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, Conference Proceedings, pp. 7291–7299.

[55] L. Zhang, G. Zhu, L. Mei, P. Shen, S. A. A. Shah, and M. Bennamoun, "Attention in convolutional lstm for gesture recognition," in *Advances in Neural Information Processing Systems*, 2018, Conference Proceedings, pp. 1953–1962.

[56] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, Conference Proceedings, pp. 802–810.

[57] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 833–842.

[58] A. Corti, S. Giancola, G. Mainetti, and R. Sala, "A metrological characterization of the kinect v2 time-of-flight camera," *Robotics and Autonomous Systems*, vol. 75, pp. 584–594, 2016.

[59] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, Conference Proceedings, pp. 1147–1153.

[60] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, Conference Proceedings, pp. 1290–1297.

[61] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, Conference Proceedings, pp. 716–723.

[62] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 359–13 368.

[63] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, Conference Proceedings, pp. 588–595.

[64] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, Conference Proceedings, pp. 5344–5352.

[65] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based lstm networks for 3d action recognition and detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3459–3471, 2018.

[66] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, Conference Proceedings, pp. 597–600.

[67] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, Conference Proceedings, pp. 5323–5332.

[68] W. Nie, W. Wang, and X. Huang, "Srnet: Structured relevance feature learning network from skeleton data for human action recognition," *IEEE Access*, vol. 7, pp. 132 161–132 172, 2019.

[69] S. Song, J. Liu, Y. Li, and Z. Guo, "Modality compensation network: Cross-modal adaptation for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3957–3969, 2020.

[70] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, Conference Proceedings, pp. 1159–1168.

[71] S. Das, R. Dai, D. Yang, and F. Bremond, "Vpn++: Rethinking video-pose embeddings for understanding activities of daily living," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[72] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[73] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[74] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[75] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[76] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[77] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 6, pp. 1453–1467, 2019.

[78] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[79] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2842–2855, 2018.

[80] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *European Conference on Computer Vision*. Springer, 2016, Conference Proceedings, pp. 203–220.

[81] B. Li, H. Chen, Y. Chen, Y. Dai, and M. He, "Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, Conference Proceedings, pp. 613–616.

[82] 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, Conference Proceedings, pp. 1110–1118.

[83] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, Conference Proceedings, pp. 1012–1020.

[84] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, Conference Proceedings, pp. 1362–1369.

[85] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, "3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, Conference Proceedings, pp. 2601–2608.

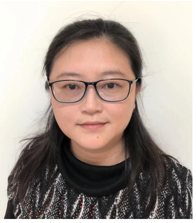
[86] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, Conference Proceedings, pp. 2458–2466.

[87] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[88] D. Yang, R. Dai, Y. Wang, R. Mallick, L. Minciullo, G. Francesca, and F. Bremond, "Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2363–2372.

[89] M. S. Ryoo, A. Piergiovanni, J. Kangaspunta, and A. Angelova, "Assemblenet++: Assembling modality representations via attention connections," in *European Conference on Computer Vision*. Springer, 2020, pp. 654–671.

Bruce X.B. Yu obtained Ph.D. in the Department of Computing from The Hong Kong Polytechnic University, Hong Kong. He is now with the Hong Kong Polytechnic University as a Research Associate. His research interests include big data analytics, machine learning, deep learning, human motion analysis and healthcare.



Yan Liu obtained Ph.D. degree in Computer Science from Columbia University in the US. In 2005, she joined The Hong Kong Polytechnic University, Hong Kong, where she is currently an Associate Professor with the Department of Computing and the director of Cognitive Computing Lab. Her research interests span a wide range of topics, ranging from brain modeling and cognitive computing, image/video retrieval, computer music to machine learning and pattern recognition.

Xiang Zhang is currently a PhD candidate at the department of computing in The Hong Kong Polytechnic University, Hong Kong. She graduated as Bachler of Engineering in 2016 at the College of Computer Science and Technology in Zhejiang University, Zhejiang, China. Her major research interests include chatbot, natural language processing, developmental robotics, algorithmic composition, and affective computing.

Sheng-hua Zhong received her Ph.D. from Department of Computing, The Hong Kong Polytechnic University in 2013. She worked as a Postdoctoral Research Associate in Department of Psychological & Brain Sciences at The Johns Hopkins University from 2013 to 2014. Currently, she is an associate professor in College of Computer Science & Software Engineering at Shenzhen University in Shenzhen. Her research interests include multimedia content analysis, brain science, and machine learning.

Keith C.C. Chan obtained Ph.D. degrees in systems design engineering from the University of Waterloo, Waterloo, ON, Canada. He then worked as a Software Analyst for the development of multimedia and software engineering tools with the IBM Canada Laboratory, Toronto, ON, Canada. In 1994, he joined The Hong Kong Polytechnic University, Hong Kong, where he was a Professor with the Department of Computing. He has authored or coauthored more than 250 publications.