Joint Learning in the Spatio-Temporal and Frequency Domains for Skeleton-Based Action Recognition

Guyue Hu, *Student Member, IEEE*, Bo Cui, and Shan Yu

***Abstract*—Benefiting from its succinctness and robustness, skeleton-based action recognition has recently attracted much attention. Most existing methods utilize local networks (e.g. recurrent network, convolutional network, and graph convolutional network) to extract spatio-temporal dynamics hierarchically. As a consequence, the local and non-local dependencies, which contain more details and semantics respectively, are asynchronously captured in different level of layers. Moreover, existing methods are limited to the spatio-temporal domain and ignore information in the frequency domain. To better extract synchronous detailed and semantic information from multi-domains, we propose a residual frequency attention (rFA) block to focus on discriminative patterns in the frequency domain, and a synchronous local and non-local (SLnL) block to simultaneously capture the details and semantics in the spatio-temporal domain. In addition, to optimize the whole learning processes of the multi-branch network, we put it under a pseudo multi-task learning paradigm. During training, 1) a soft-margin focal loss (SMFL) is proposed to optimize the intra-branch separated learning process, which can automatically conduct data selection and encourage intrinsic margins in classifiers; 2) A mutual learning policy is also proposed to further facilitate the inter-branch collaborative learning process. Eventually, our approach achieves the state-of-the-art performance on several large-scale datasets for skeleton-based action recognition.**

***Index Terms*—Action recognition, frequency attention, synchronous local and non-local learning, soft-margin focal loss, multi-task learning.**

I. Introduction

HUMAN action recognition is an active topic in the fields of computer vision and multimedia, which is widely applied in video understanding, intelligent surveillance, and human-computer interaction, etc. Human actions can be represented by various of media modalities including RGB video, optical flow, depth, and skeleton [1]–[7]. Due to its succinctness of representation and robustness to variations of viewpoints, appearances and surroundings [8], [9], the skeleton-based human action recognition has recently attracted increasing attention. In this paper, we focus on recognizing human actions from the skeleton-based sequence.

Most of previous works treat skeletal actions as sequences and pseudo-images, then apply Recurrent Neural Networks (RNN) [8], [10], [11] and Convolutional Neural Networks (CNN) [12], [13] to model the temporal evolutions and the spatio-temporal dynamics, respectively. Recently, some works [14]–[16] also feed skeleton graphs into graph convolutional networks (GCN) to exploit the structure information of human body. However, all the aforementioned methods apply stacked local networks to hierarchically extract spatio-temporal features, which lead to two serious problems. 1) The recurrent and convolutional operations are neighborhood-based local operations [17], so the local-range detailed information and non-local semantic information mainly be captured asynchronously in the lower and higher layers respectively, which hinders the fusion of details and semantics in action dynamics. 2) Some human actions have characteristic frequency patterns (See Fig. 1), but previous works are always limited to the spatio-temporal dynamics and ignore the discriminative patterns in the frequency domain.

To move beyond such limitations, we propose a novel model SLnL-rFA to better extract synchronous detailed and semantic information from multi-domains. SLnL-rFA is equipped with synchronous local and non-local (SLnL) blocks for spatio-temporal learning, and a residual frequency attention (rFA) block for frequency-patterns mining. Fig. 2 shows the overall pipeline of our method. Firstly, an adaptive transform network augments and transforms the skeletal actions. Secondly, the residual frequency attention block selects discriminative frequency patterns. Then, M_1 synchronous local and non-local (SLnL) blocks and M_2 local blocks are applied sequentially in the spatio-temporal domain, where SLnL block is designed to simultaneously extract local details and non-local semantics. Thus, we obtained three modalities of feature including branches of position feature, velocity feature and concatenated feature.

Manuscript received July 23, 2019; revised October 16, 2019; accepted November 7, 2019. Date of publication November 13, 2019; date of current version August 21, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0105203, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) under Grant XDB32040200, and in part by the Hundred-Talent Program of CAS (for S.Y.). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wanqing Li. *(Corresponding author: Guyue Hu.)*

G. Hu and B. Cui are with the Brainnetome Center and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: guyue.hu@nlpr.ia.ac.cn; bo.cui@nlpr.ia.ac.cn).

S. Yu is with the Brainnetome Center and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shan.yu@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2953325

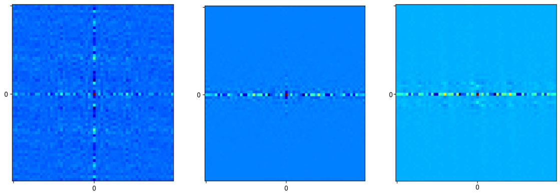


Fig. 1. The mean frequency maps corresponding to three different action classes in the NTU RGB+D dataset. Note that the frequency maps are obtained from the spatio-temporal input features $\boldsymbol{X}'$ of the proposed residual frequency attention (See Fig. 2 for details). To facilitate visualization, the zero-frequency component is shifted to the center of map by function *fftshift*.

In addition, to optimize the whole learning processes of the multi-branch network with multi-modal features, we adopt a pseudo multi-task learning (MTL) manner. During training, a soft-margin focal loss (SMFL) is proposed to optimize the intra-branch learning process, which can automatically conduct data selection and encourage intrinsic margins in classifiers. Besides, a mutual learning policy is also proposed to further facilitate the inter-branch collaboratively learning process, which can encourage the feature branches to mutually aiding each other.

Finally, the main contributions of this paper can be summarized as follows:

1) Moving beyond the spatio-temporal domain, we propose a residual frequency attention block to exploit frequency information for skeleton-based action recognition.
2) We propose a synchronous local and non-local block to simultaneously capture details and semantics in the early-stage layers of the network.
3) We propose a soft-margin focal loss, which can adaptively conduct data selection during training process and encourage intrinsic soft-margins in the classifiers.
4) We put the fusion of multi-modal features under a pseudo multi-task learning paradigm, and further proposed a mutual learning policy to facilitate the collaboration among different feature branches.
5) Our approach consistently outperforms the state-of-the-art methods on several datasets for skeleton-based action recognition, including the NTU RGB+D, Kinetics, N-UCLA, and SYSU datasets.

We note that a preliminary report of this work was published in a conference [18]. As an extension of the preliminary version, we further proposed a mutual learning policy to facilitate the inter-branch aiding during the learning of MTL task, and an adaptive coordinate transform to enrich the skeletal action representation in multiple oblique coordinate systems. We also extensively enrich the experimental datasets, ablation analyses, and visualizations to give more insights on the proposed FA, SLnL and SMFL blocks. Furthermore, the influence of key hyper-parameter choice is also explored. Finally, we present analysis about model complexity and inference efficiency of the proposed framework.

II. Related Works

In this section, we briefly review the previous works which are closely related to the proposed method.

A. Skeleton-Based Action Recognition

Previous skeleton-based action recognition methods can be categorized into two classes: hand-crafted features based methods and deep learning methods. Hand-crafted features based methods include histograms of 3D joint locations [19], actionlet ensemble obtained from data minning [20], covariance matrices of joints trajectories and relative joint positions [20], [21], and joints in a Lie group [22]. Deep learning methods include RNN-based methods [8], [10], [23], CNN-based methods [13], [24], and graph based method [9], [14]–[16], [25], [26]. These approaches progressively use local operations to model spatio-temporal dynamics and have no non-local operation to explore global information in early-stage layers, while our work synchronously fuse local and non-local features in lower layers.

B. Frequency Domain Analysis

Generalized frequency domain analysis contains several large classes of methods such as discret Fourier transform (DFT), short-time Fourier transform (SFT) and wavelet tranform, which are classical and powerful tools in the fields of signal analysis and image processing [27]. Due to the booming of deep learning techniques [28]–[30], methods based on the spatio-temporal domain dominate the field of computer vision, with only a few works paying attention to the frequency domain. For example, frequency domain analysis of critical points trajectories [31] and frequency divergence image [32] are applied for RGB-based action recognition. Scattering convolution network with wavelet filters are used for object classification [33]. Our work will revisit the frequency domain, and exploit discriminative frequency patterns to improve the skeleton-based action recognition.

C. Non-Local Operations

Non-local means is a classical filtering algorithm that allows distant pixels to contribute to the target pixel [34]. Block-matching [35] explores groups of non-local similarity between patches, which is a solid baseline for image denoising. Block-matching is widely used in computer vision tasks like super-resolution [36], image inpainting [37], image denoising [38] etc. The popular self-attention [39] in machine translation can also be viewed as a non-local operation. Recently, different non-local blocks are inserted into CNNs for video classification [17] and RNNs for image restoration [40]. However, their local and non-local operations apply to objects in different level of layers but our SLnL simultaneously operate on the same objects, thus only the proposed SLnL can extract local and non-local information synchronously.

D. Reformed Softmax Loss

The softmax loss [41], consisted of the last fully connected layer, the softmax function, and the cross-entropy loss, is widely applied in supervised learning due to its simplicity and clear probabilistic interpretation. However, recent works [41]–[43] have exposed its limitations on feature discriminability and have stimulated two types of improvements. One type directly refines or combines the cross-entropy loss with other losses like

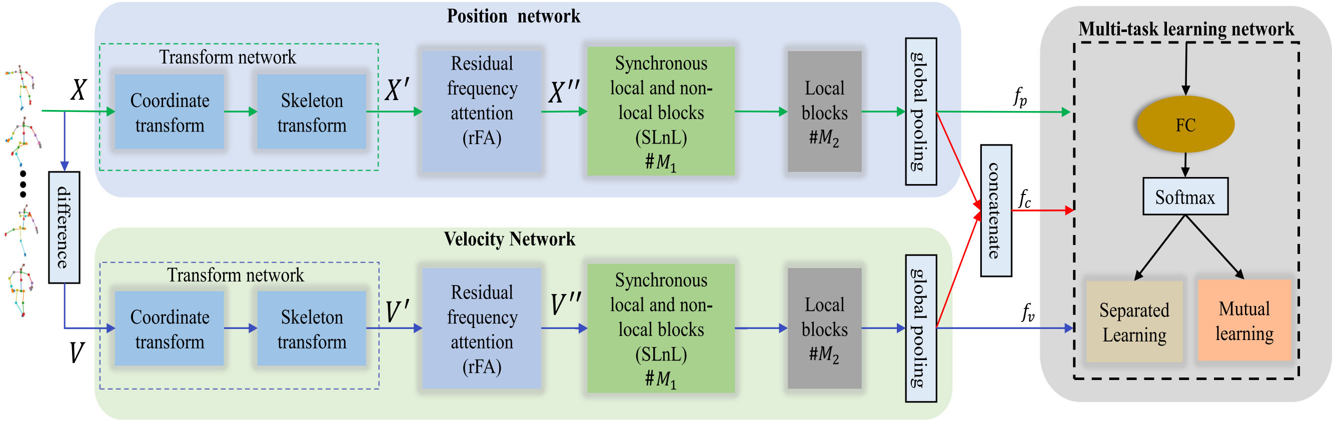


Fig. 2. The overall pipeline of the proposed method. The position and velocity information of human joints are fed into a tranform network, a residual frequency attention network, M_1 synchronous local and non-local blocks, and M_2 local blocks sequentially. Treated as a pseudo multi-task learning task, the model is optimized by conventional separated learning and the proposed mutual learning policy, see Fig. 6(d) for details.

contrastive loss, triplet loss, etc [42], [44]. The other type reformulates the softmax function with geometrical or algebraic margin [41], [42] to encourage intra-class compactness and inter-class separability of feature learning, which completely destroys the probabilistic meaning of the original softmax function. Contrastively, our SMFL not only conducts data selection but also encourages intrinsic soft-margins in classifiers with a clear probabilistic interpretation, which will be proved in the *Methods* section.

E. *Adaptive Data Selection*

The contributions of easy data and hard data are different among the training processes of neural networks, thus adaptive data selection strategy significantly impact the model performance and training efficiency [45]. Some previous studies adopt heuristic rules to adjust the sampling probabilities of the train data, such as curriculum learning [46], self-paced learning [47], online batch selection [48], etc. Fan *et al.* [45] also uses deep reinforcement learning framework to automatically learn what data to learn. However, the aforementioned methods require extra data selection networks or complex modifications to the mainstream shuffle-based training pipeline, while the focal loss [49] introduces only simple modification to the loss function that can encourage effective data selection. Thus our soft-margin focal loss also falls into this paradigm.

III. Methods

The overall pipeline have been introduced in the Introduction section. In this section, we will dig into the details of each component separately.

A. *Preliminary*

A skeletal action $\boldsymbol{X} \in \mathbb{R}^{d\times T\times N}$ is represented by d dimensional locations of N body joints in a T frames video sequence. Directly taking action $\boldsymbol{X}$ as a d-channels spatio-temporal image will lose structural information among skeletons. Following Li *et al.* [24], each skeleton $\boldsymbol{S} \in \mathbb{R}^{N\times d}$ with the structureless permutation is adaptively augmented and rearranged as an optimal permutation $\boldsymbol{S}' \in \mathbb{R}^{N'\times d}$ through a transform function $\boldsymbol{S}' = \boldsymbol{W}_s\boldsymbol{S}$, where $\boldsymbol{W}_S \in \mathbb{R}^{N'\times N}$ is the transform matrix and N' is the number of new joints. As a result, the transform can adaptively learn an optimal permutation of joints, and augment the joint number from N to N' where each new joint is a linear combination of original N joints.

Similarly, we propose a coordinate transform to transfer the original joint representation $\boldsymbol{J}_0 \in \mathbb{R}^d$ in the single rectangular coordinate system to rich representations $\boldsymbol{J}_1, \boldsymbol{J}_2, \ldots, \boldsymbol{J}_K$ in K oblique coordinate systems. $\boldsymbol{J}_i = \boldsymbol{C}_i^T\boldsymbol{J}_0$, where $\boldsymbol{C}_i$ is the transition matrix from the original coordinate system to a new coordinate system i. For convenience, the K coordinates are concatenated as $\boldsymbol{J} = [\boldsymbol{J}_1, \boldsymbol{J}_2, \ldots, \boldsymbol{J}_K]^T$, and similar for the transition matrices $\boldsymbol{C} = [\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_K]^T$. Therefore, the expressions of human actions are enriched to K oblique coordinate systems by the concatenated transform matrix $\boldsymbol{C}$.

The whole transform network in Fig. 2 is implemented with fully connected layers and corresponding transpose, flatten, and concatenate operations. As a result, a new adaptive expression $\boldsymbol{X'} \in \mathbb{R}^{Kd\times T'\times N'}$ is formed for each action.

B. *Residual Frequency Attention*

Previous works always concentrate on the spatio-temporal domain, but many actions contain inherent frequency-sensitive patterns, such as *shaking hands*, and *brushing teeth*. The gap motivates us to revisit the frequency domain. The classical operations in the frequency domain, such as high-pass, low-pass, and band-pass filters, only have a few parameters that are far from enough, thus we propose a more general frequency attention block (Fig. 3) equipped with abundant learnable parameters to adaptively select frequency components.

Given a transformed action after the transform network $\boldsymbol{X'} \in \mathbb{R}^{C'\times T'\times N'}$ ($C' = Kd$, $T' = T$), the 2D discret Fourier

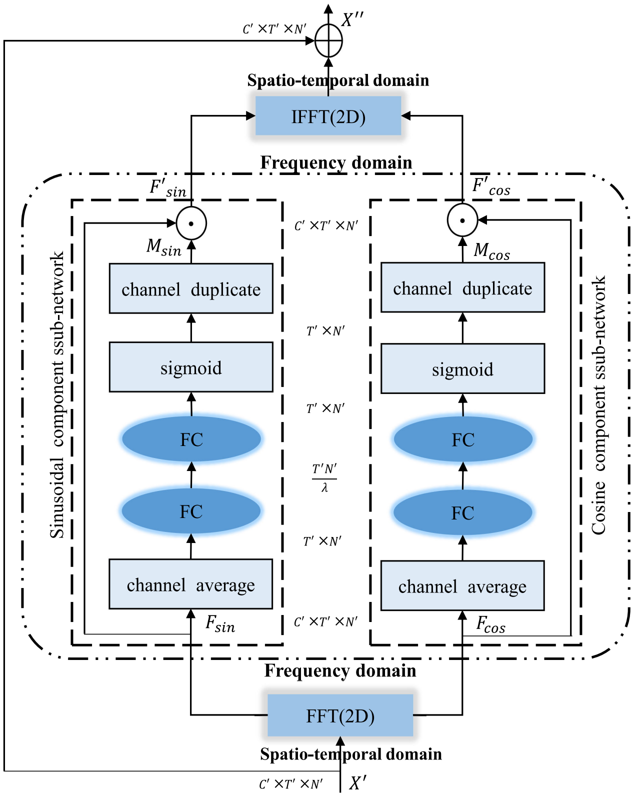


Fig. 3. The detailed structure of residual frequency attention. The spatio-temparal domain and frequency domain are switched conveniently through 2D-FFT and 2D-IFFT. The attention for the sinusoidal and cosine components ($\boldsymbol{F}_{\sin}, \boldsymbol{F}_{\cos}$) are conducted in the frequency domain, and the residual component is applied in the spatio-temporal domain.

transform (DFT) transforms the pseudo spatio-temporal image $\boldsymbol{X'}$ in each channel to $\boldsymbol{Y'} \in \mathbb{R}^{C'\times T'\times N'}$ in the frequency domain via

$$\begin{aligned}\boldsymbol{Y'}[c,u,v] &= \sum_{t=0}^{T'-1}\sum_{n=0}^{N'-1}\boldsymbol{X'}[c,t,n]\cos\left(-2\pi\left(\frac{ut}{T'}+\frac{vn}{N'}\right)\right)\\ &+ j\sum_{t=0}^{T'-1}\sum_{n=0}^{N'-1}\boldsymbol{X'}[c,t,n]\sin\left(-2\pi\left(\frac{ut}{T'}+\frac{vn}{N'}\right)\right)\\ &= \boldsymbol{F}_{\cos}[c,t,n] + j\boldsymbol{F}_{\sin}[c,t,n], \qquad (1)\end{aligned}$$

where u, v and c are frequencies and channel of spatio-temporal image respectively. $\boldsymbol{F}_{\cos}$ and $\boldsymbol{F}_{\sin}$ denote the cosine and sinusoidal component, respectively. The frequency spectrum $\boldsymbol{F}_A = (\boldsymbol{F}_{\cos}^2 + \boldsymbol{F}_{\sin}^2)^{1/2}$ and the phase spectrum $\boldsymbol{F}_\phi = arctan(-\frac{\boldsymbol{F}_{\sin}}{\boldsymbol{F}_{\cos}})$. In practice, the DFT and its inverse (IDFT) are computed through the fast Fourier transform (FFT) algorithm and its inverse (IFFT).

For each action, the attention weights $\boldsymbol{M}_{\cos}$ and $\boldsymbol{M}_{\sin}$ are complex functions of its cosine and sinusoidal components in the frequency domain, i.e.

$$\boldsymbol{M}_i = dup(\sigma(\boldsymbol{W}_{i1}(\boldsymbol{W}_{i2}(Avg(\boldsymbol{F}_i)) + b_{i1}) + b_{i2})), \qquad (2)$$

where $i \in \{cos, sin\}$. Specifically, after a channel averaging operation, each component is fed into two fully connected layers (FC) to learn adaptive weights for each frequency, followed by a sigmoid transfom function. The first FC layers serve as a bottleneck layer [29] for dimensionality reduction with a ratio factor λ. Then, the learned attention weights are duplicated to every channel to pay attention to the input frequency image via

$$\boldsymbol{F}'_{\sin} = \boldsymbol{F}_{\sin} \odot \boldsymbol{M}_{\sin}, \qquad (3)$$

$$\boldsymbol{F}'_{\cos} = \boldsymbol{F}_{\cos} \odot \boldsymbol{M}_{\cos}, \qquad (4)$$

where $\odot$ denotes the element-wise multiplication. Eventually, to avoid severely destroying information in the spatio-temporal domain when strengthening the key frequent patterns, a spatio-temporal residual trick is applied to obtain the final output $\boldsymbol{X''} \in \mathbb{R}^{C'\times T'\times N'}$ after attention, i.e.

$$\boldsymbol{X''} = \boldsymbol{X'} + ifft2(\boldsymbol{F}'_{\sin},' \boldsymbol{F}'_{\cos}), \qquad (5)$$

where $ifft2$ denotes the efficient 2-dimensional IFFT.

C. Synchronous Local and Non-Local Learning in the Spatio-Temporal Domain

1) Non-Local Module: A general non-local operation takes a multi-channel signal $\boldsymbol{X} \in \mathbb{R}^{M\times P}$ as its input and generates a multi-channel output $\boldsymbol{Y} \in \mathbb{R}^{M\times Q}$. Here P and Q are channels, and M is the number of Ω, where Ω is the set that enumerates all positions of the signal (image, video, feature map, etc.). Let $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ denote the i-th row vector of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, the non-local operation is formulated as follows:

$$\boldsymbol{y}_i = \frac{1}{\mathcal{Z}_i(\boldsymbol{X})}\sum_{j\in\Omega}\phi(\boldsymbol{x}_i, \boldsymbol{x}_j)g(\boldsymbol{x}_j), \qquad \forall i \in \Omega \qquad (6)$$

where the multi-channel unary transform $g(\boldsymbol{x}_j)$ computes the embedding of x_j, the multi-channel binary transform $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j)$ computes the affinity between the positions i and j, and $\mathcal{Z}(\boldsymbol{X})$ is a normalization factor. With different choices of ϕ and g, such as Gaussian, embedded Gaussian and dot product, various of non-local operations could be constructed. For simplicity, we only consider ϕ and g in the form of linear embedding and embedded Gaussian respectively, and set $\mathcal{Z}_i(\boldsymbol{X}) = \sum_{j\in\Omega}\phi(\boldsymbol{x}_i, \boldsymbol{x}_j)$, i.e.

$$g(\boldsymbol{x}_j) = (\boldsymbol{W}_g\boldsymbol{x}_j^T)^T, \qquad \forall j \qquad (7)$$

where $\boldsymbol{W}_g \in \mathbb{R}^{Q\times P}$ are learnable transform parameters.

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\varphi(\boldsymbol{x}_i)^T\psi(\boldsymbol{x}_j)}, \quad \forall i, j \qquad (8)$$

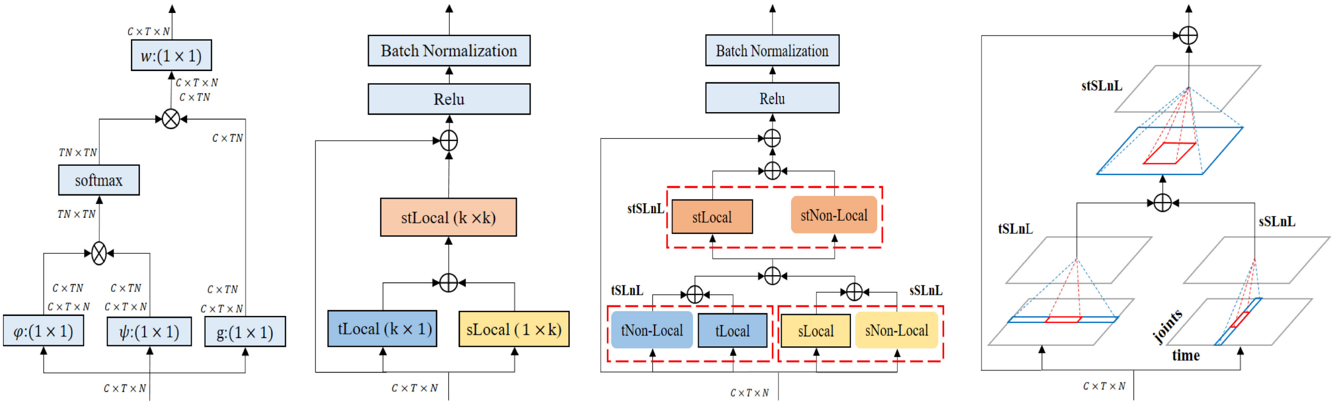
$$\varphi(\boldsymbol{x}_i) = (\boldsymbol{W}_\varphi\boldsymbol{x}_i^T)^T, \quad \forall i \qquad (9)$$

$$\psi(\boldsymbol{x}_j) = (\boldsymbol{W}_\psi\boldsymbol{x}_j^T)^T, \quad \forall j \qquad (10)$$

where $\boldsymbol{W}_\varphi, \boldsymbol{W}_\psi \in \mathbb{R}^{L\times P}$, and L denotes the embedding channel. To weigh how important the non-local information is when compared to local information, a weighting function is appended, i.e.

$$w(\boldsymbol{y}_i) = (\boldsymbol{W}_w(\boldsymbol{y_i})^T)^T, \qquad (11)$$

where $\boldsymbol{W}_w \in \mathbb{R}^{Q\times Q}$. Note that the non-local modules can be drop-in pretrained model without breaking its initial behavior by initializing $\boldsymbol{W}_w$ as 0. A non-local module with a d-dimensional input can be completed with some transpose operations, some



(a) 2D Non-local module (b) Baseline local block (c) SLnL block (d) The affinity field of SLnL

Fig. 4. (a) A 2D example of non-local module. (b) The structure of the baseline local block. (c) The structure of the proposed synchronous local and non-local (SLnL) block. (d) The affinity field of synchronous local and non-local block. Note that the *affinity field* is a more general concept than the receptive field of CNNs. The red and blue in (d) represent local and non-local modules, repectively.

convolutional layers with the kernels of 1, and a softmax layer, Fig. 4(a) shows a 2D example.

2) Baseline Local Block: The local operation is defined as

$$\boldsymbol{y}_i = \frac{1}{\mathcal{Z}_i(\boldsymbol{X})} \sum_{j \in \delta_i} \phi(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_j), \qquad \forall i \in \Omega \tag{12}$$

where δ_i is the local neighbor set of target position i, $\delta_i \ll \Omega$. And $\mathcal{Z}_i(\boldsymbol{x})$ is the local normalization factor within δ_i. The convolution is a typical local operation with identity affinity $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1$, liner transform $g(\boldsymbol{x}_j) = \boldsymbol{w}_j \boldsymbol{x}_j$, identity normalization factor $\mathcal{Z}_i(\boldsymbol{X}) = 1$, and δ_i is the neighbors around target center i with a same shape of kernel. Our baseline local block is constructed from convolution operation. As shown in Fig. 4(b), two convolutional layers with kernel $k \times 1$ and $1 \times k$ are applied to learn temporal local (tLocal) features and spatial local (sLocal) features respectively, and a $k \times k$ convolutional layer for spatial-temporal local (stLocal) features. The block also contains a residual path, a rectified linear unit (ReLU) and a batch normalization (BN) layer.

3) Synchronous Local and Non-Local Block (SLnL): In order to synchronously exploit local details and non-local semantics in human actions, three non-local modules are parallel merged into the above baseline local block. As shown in Fig. 4(c), two 1D non-local modules to explore temporal non-local (tNon-Local) and spatial non-local (sNon-Local) information respectively, followed by a 2D non-local module for spatio-temporal non-local (stNon-Local) patterns. We define the *affinity field* as the representation of the range of pixel indices that could contribute to the target position in the next layer of the local or non-local modules, which is a more general concept than the *receptive field* of CNNs. The affinity field in Fig. 4(d) clearly shows our SLnL can mine local details and non-local semantics synchronously in every layer. Note that our SLnL is significantly different from the methods [17], [40] which only inserted a few non-local modules after stacked local networks, thus the local and non-local operations are still separately conducted in different layers having different resolutions. Contrastively, our SLnL simultaneously captures local and non-local patterns in every layer (Fig. 4(d)).

D. Soft-Margin Focal Loss

A common challenge for classification tasks is that the discrimination difficulties are different across samples and classes, but most previous works for skeleton-based action recognition use the *softmax loss* that haven't taken it into consideration. There are two possible measures to alleviate it, i.e. data selection and margin encouraging.

Intuitively, the larger predicted probability a sample has, the farther away from the decision boundary it might be, and vice versa. Motivated by this intuition, we construct a soft-margin (SM) loss term as follows:

$$\mathcal{L}_{SM}(p_t) = log\left(e^m + (1 - e^m)p_t\right), \tag{13}$$

where p_t is the estimated posterior probability corresponding to ground truth class, and m is a margin parameter. $\mathcal{L}_{SM} \in [0, m]$ for the fact that $p_t \in [0, 1]$. As Fig. 5 shows when the posterior probability p_t is small, the sample is more likely close to the boundary, thus we penalize it with a relative large margin loss. Otherwise, a small margin loss is imposed. To further illustrate the idea, we introduce the $\mathcal{L}_{SM}$ term into the cross entropy loss leading to a soft-margin cross entropy (SMCE) loss,

$$\begin{aligned} \mathcal{L}_{SMCE}(p_t) &= \mathcal{L}_{SM} + \mathcal{L}_{CE} \\ &= log\left(e^m + (1 - e^m)p_t\right) - log(p_t). \end{aligned} \tag{14}$$

Assuming that $\boldsymbol{x} \in \mathbb{R}^d$ is the features before the last FC layer, the FC layer transforms it into score $\boldsymbol{z} = [z_1, z_2, \ldots, z_C]^T \in \mathbb{R}^C$ of C classes by multiplying $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_C] \in \mathbb{R}^{d \times C}$, where $\boldsymbol{w}_c$ is the parameter of the linear classifier corresponding to the class c, i.e. $z_c = \boldsymbol{w}_c^T \boldsymbol{x}$. Followed with a softmax layer, $p_t = \frac{e^{\boldsymbol{w}_t \boldsymbol{x}}}{\sum_{c=1}^{C} e^{\boldsymbol{w}_c \boldsymbol{x}}}$ and $(1 - p_t) = \frac{\sum_{c \neq t}^{C} e^{\boldsymbol{w}_c \boldsymbol{x}}}{\sum_{c=1}^{C} e^{\boldsymbol{w}_c \boldsymbol{x}}}$, then the SMCE can

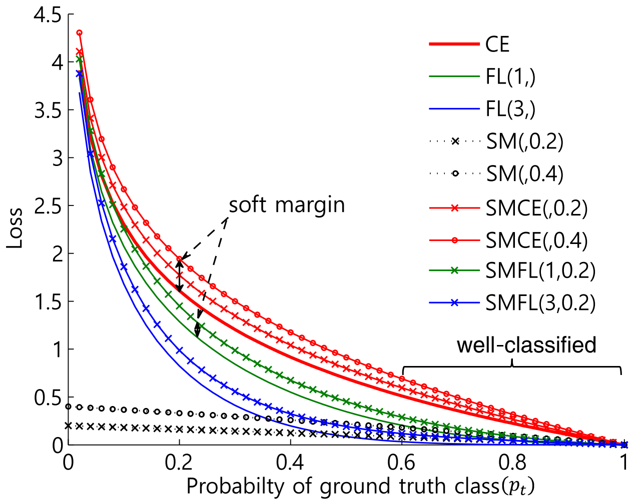


Fig. 5. Comparisons among our soft-margin focal loss (SMFL), the soft-margin cross entropy (SMCE) loss, the cross-entropy (CE) loss, the focal loss (FL), and the soft-margin loss (SM). The focusing parameter γ and the margin parameter m of losses are expressed as (γ, m).

be rewritten as

$$\begin{aligned}\mathcal{L}_{SMCE} &= log\left(p_t + e^m \cdot (1-p_t)\right) - log(p_t)\\ &= log\left(\frac{e^{\boldsymbol{w}_t\boldsymbol{x}} + e^m \cdot \sum_{c\neq t}^{C} e^{\boldsymbol{w}_c\boldsymbol{x}}}{\sum_{c=1}^{C} e^{\boldsymbol{w}_c\boldsymbol{x}}}\right)\\ &\quad - log\left(\frac{e^{\boldsymbol{w}_t\boldsymbol{x}}}{\sum_{c=1}^{C} e^{\boldsymbol{w}_c\boldsymbol{x}}}\right)\\ &= -log\left(\frac{e^{\boldsymbol{w}_t\boldsymbol{x}}}{e^{\boldsymbol{w}_t\boldsymbol{x}} + e^m \cdot \sum_{c\neq t}^{C} e^{\boldsymbol{w}_c\boldsymbol{x}}}\right)\\ &= -log\left(\frac{e^{\boldsymbol{w}_t\boldsymbol{x}-m}}{e^{\boldsymbol{w}_t\boldsymbol{x}-m} + \sum_{c\neq t}^{C} e^{\boldsymbol{w}_c\boldsymbol{x}}}\right).\end{aligned} \tag{15}$$

Comparing the standard *softmax loss* with Eq.15, only the score of the ground truth class $\boldsymbol{w}_t\boldsymbol{x}$ is replaced by $\boldsymbol{w}_t\boldsymbol{x} - m$. Optimizing model with SMCE, we will obtain classifiers that meet the constraint $\boldsymbol{w}_t\boldsymbol{x} - m \geq \boldsymbol{w}_{c\neq t}\boldsymbol{x}$. As a result, an intrinsic margin m between the positive (belonging to a specific class) samples and the negative (not belonging to the specific class) samples of each class will be formed in classifiers by adding the SM loss term into the loss function.

In addition, the focal loss [49] defined as

$$\mathcal{L}_{FL}(p_t) = -(1-p_t)^{\gamma} log(p_t), \tag{16}$$

where γ is a focusing parameter, can encourage adaptive data selection without any damage to the original model structure and training processes. As Fig. 5 shows the relative loss for well-classified easy samples is reduced by FL when compared to CE. Although FL pays more attention to hard samples, it has no margin around the decision boundary. Similar to SMCE, we introduce the $\mathcal{L}_{SM}$ term into FL to obtain the soft-margin focal loss (SMFL) as follows:

$$\begin{aligned}\mathcal{L}_{SMFL}(p_t) &= \mathcal{L}_{SM} + \mathcal{L}_{FL}\\ &= log\left(e^m + (1-e^m)p_t\right) - (1-p_t)^{\gamma} log(p_t).\end{aligned} \tag{17}$$

Finally, the proposed SMFL can encourage intrinsic margins in classifiers and maintain FL's advantage of adaptive data selection as well.

E. Pseudo Multi-Task Learning

The two-stream network produces features from both position and velocity information (Fig. 2), thus it is vital to explore an effective multi-modal feature fusion policy. Most of the existing works directly sum or concatenate the position feature $\boldsymbol{f}_p$ and the velocity feature $\boldsymbol{f}_v$, as shown in Fig. 6(a) and Fig. 6(b). In these policies, the two-modal information are completely entangled with each other that is hard to optimize. Contrastively, we treat the multi-modal optimization task as a pseudo multi-task learning paradigm (Fig. 6(c)), thus the sub-tasks that contain only one feature modality ($\boldsymbol{f}_p$ or $\boldsymbol{f}_v$) can provide optimization guide to the tangled feature modality $\boldsymbol{f}_c$. Specifically, each of the three predicted probabilities $\boldsymbol{p}^p$, $\boldsymbol{p}^v$, $\boldsymbol{p}^c$ produces a SMFL loss via,

$$\mathcal{L}_k = \sum_{i=1}^{C} y_i \left(log(e^m + (1-e^m)p_i^k) - (1-p_i^k)^{\gamma} log(p_i^k)\right), \tag{18}$$

where $k \in \{p, v, c\}$ is modality type, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_C)$ is the one-hot class label. With conventional separated learning policy in Fig. 6(c), the final supervised loss of the multi-task learning task is obtained as follows:

$$\mathcal{L}_{sup} = \mathcal{L}_p + \mathcal{L}_v + \mathcal{L}_c. \tag{19}$$

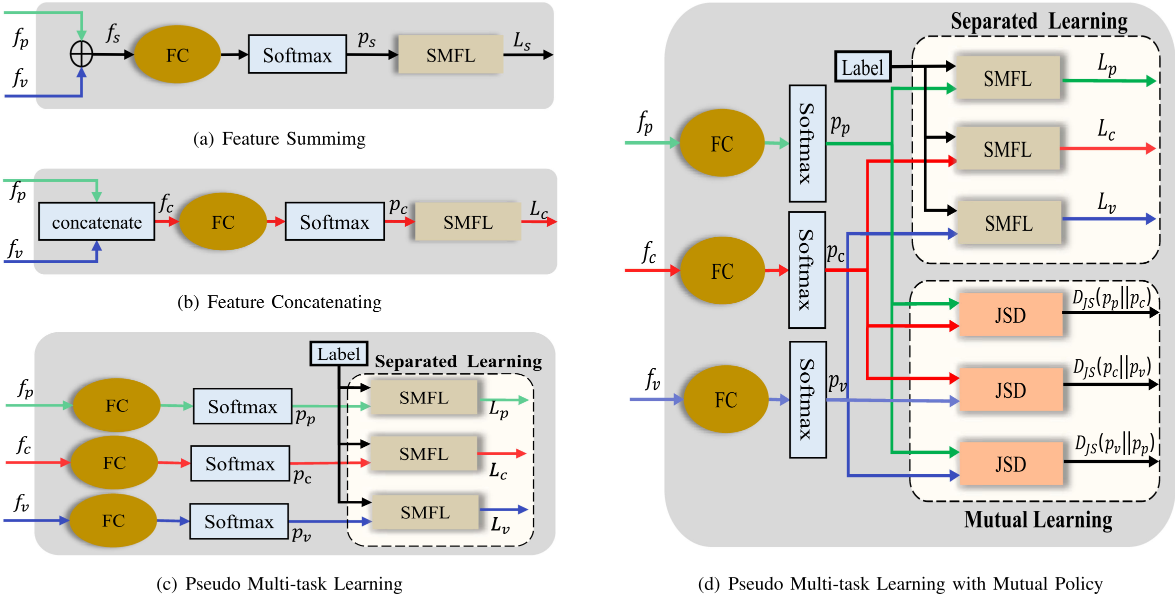
Besides, as shown in Fig. 6(c), the sub-networks with different feature modalities try to learn the same joint probability regarding video clips and action classes. This paradigm is somewhat similar as a situation in which a handful of persons are assigned to learn a common task. Motivated by the fact that persons can mutually teach each other, we proposed a mutual learning policy to further assist the training process of the pseudo multi-task learning task (Fig. 6(d)). Specifically, every feature branch $k \in \{p, c, v\}$ learns from another feature branch $j \in \{p, c, v\}$ by minimizing their Kullback-Leibler Divergence, thus the divergence punishment for branch k is,

$$\mathcal{D}_k = \frac{1}{2} \sum_{j\neq k} D_{KL}(\boldsymbol{p}_j \parallel \boldsymbol{p}_k). \tag{20}$$

Then a mimicry loss for the mutual learning policy can be formulated as

$$\begin{aligned}\mathcal{L}_{mim} &= \mathcal{D}_p + \mathcal{D}_c + \mathcal{D}_v\\ &= D_{JS}(\boldsymbol{p}_p \parallel \boldsymbol{p}_c) + D_{JS}(\boldsymbol{p}_c \parallel \boldsymbol{p}_v) + D_{JS}(\boldsymbol{p}_v \parallel \boldsymbol{p}_p),\end{aligned} \tag{21}$$

where the D_{JS} denotes Jensen-Shannon Divergence (JSD).



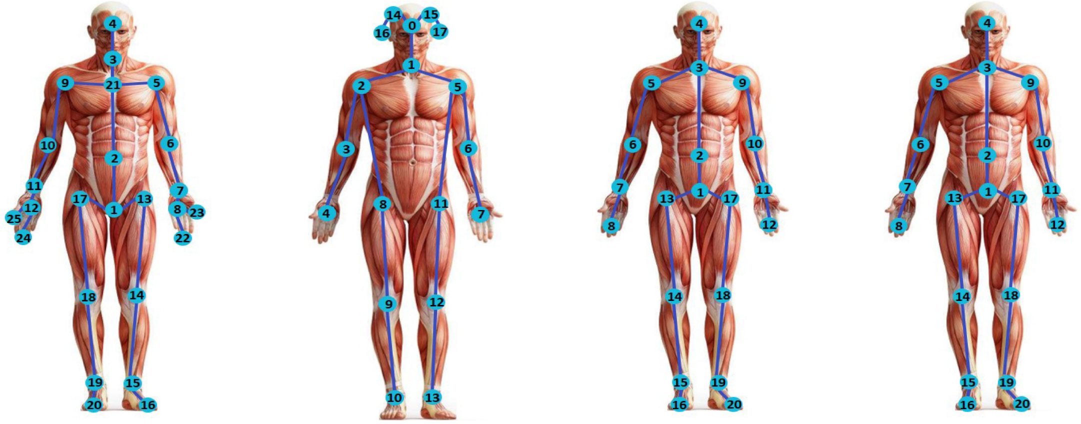
(a) Feature Summimg

(b) Feature Concatenating

(c) Pseudo Multi-task Learning

(d) Pseudo Multi-task Learning with Mutual Policy

Fig. 6. Comparing different processing policies for multi-modal features. Contrasting to the feature summimg (a) or concatenating (b) policy, we put the task under a pseudo multi-task learning (MTL) paradigm. Then, the MTL task is optimized only by the conventional separated learning policy (c) or further assisted with the proposed mutual learning policy (d). The "FC" denotes fully connected layer, the "SMFL" block is the proposed soft-margin focal loss, and the "JSD" denotes Jensen-Shannon Divergence.



(a) Joints in NTU-RGB+D

(b) Joints in Kinetics

(c) Joints in SYSU

(d) Joints in UCLA

Fig. 7. The sketches to illustrate joint locations in different skeleton-based human action recognition datasets. The NTU-RGB+D and Kinetics datasets captured 25 and 18 joints respectively while SYSU and UCLA datasets both captured 20 joints.

Finally, the multi-task learning task can not only learn from the supervised loss via separated learning policy, but also learn from the mimicry loss via mutual learning policy, i.e.,

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{mim}, \tag{22}$$

where λ is the parameter to control the influence of mutual learning policy. After trained with the total loss $\mathcal{L}$, the obtained probability $\boldsymbol{p}^c$ is eventually adopted to predict the action class during inference.

IV. Experiments

A. Datasets and Experimental Details

NTU RGB+D (NTU) dataset [10] is currently the largest in-door action recognition dataset. It contains 56,000 clips in 60 action categories performed by 40 subjects. Each clip consists of 25 joint locations with one or two persons, and the joint sketch is shown in Fig. 7(a). There are two evaluation protocols for this dataset, i.e., cross-subject (CS) and cross-view (CV). For the cross-subject evaluation, 40320 samples from 20 subjects are used for training and 16540 samples from the rest subjects are used for testing. For the cross-view evaluation, samples are split by camera views, with two views for training and the rest one for testing.

Kinetics dataset [50] is by far the largest unconstrained action recognition dataset, which contains 300,000 video clips in 400 classes retrieved from YouTube [14]. The skeleton dataset is estimated by Yan *et al.* from the raw RGB videos by the OpenPose toolbox [14]. Each joint consists of 2D coordinates (X, Y) in the pixel coordinate system and a confidence score

C, thus finally represented by a tuple of (X, Y, C). Each skeleton frame is recorded as an array of 18 joint tuples, and the joint sketch is shown in Fig. 7(b). For the multi-person cases, 2 people with the highest average joint confidence in each clip is selected. The dataset is split into training set and validation set with 240,000 and 20,000 clips, respectively. We use the released skeletal dataset to train our model, and evaluate the performance by the top-1 and top-5 accuracies as recommended by Key *et al.* [50]. To have fair comparisons with the previous works [14], [51], the confidence score is also treated as a channel of input in this paper.

SYSU 3D Human-Object Interaction (SYSU) dataset [52] is collected by Kinect camera. It contains 480 skeleton clips of 12 action categories performed by 40 subjects and each clip has 20 joints (Fig. 7(c)). There are two standard evaluation protocols for this dataset, i.e., cross-subject (CS) setting and same-subject (SS) setting [52]. Following [16], we use the 30-fold cross validation and report their mean accuracy for each setting. For the cross-subject setting, half of the subjects are used for training and the rest are for testing. For the same-subject setting, half of the samples from each activity are used for training and the rest are for testing.

Northwestern UCLA Multiview Action 3D (N-UCLA) dataset [53] is simultaneously captured by three Kinect cameras from a variety of viewpoints. It contains 1494 video clips covering 10 action categories performed by 10 different of subjects. Each person contains 20 joints, as shown in Fig. 7(d). Following the evaluation protocol in [53], samples of the first two cameras constitute the training set, and samples of the third camera constitute the testing dataset.

Implementation Details: During the data preparation, we firstly translate the origin of coordinate system to the body center of the first frame, then randomly crop a sub-sequence from the entire sequence. We randomly crop sequences with a ratio uniformly drawn from [0.5,1] for training, and centrally crop sequences with a fixed ratio of 0.95 for inference. We resize the sequences to a length of 64, 64, 64, 128 frames with bilinear interpolation for SYSU, N-UCLA, NTU and Kinetics, respectively. Finally, the obtained data are fed into a batch normalization layer to normalize the scale. During training, we apply Adam optimizer with weight decay of 0.0005. Learning rate is initialized as 0.001, followed by an exponential decay with a rate of 0.92, 0.95, 0.98, 0.95 per epoch for SYSU, N-UCLA, NTU and Kinetics, respectively. A dropout with ratio of 0.2 is applied to each block to alleviate overfitting for all datasets. The controlling parameter for mutual learning policy is empirically set as 0.1. The model is trained for in total 40, 60, 100, 300 epoches with a batch size of 16, 32, 32, 128 for SYSU, N-UCLA, NTU and Kinetics, respectively. All the experiments are conducted with the PyTorch framework.

Each stream of model for SYSU, N-UCLA or NTU datasets is composed of totally 6 blocks in Fig. 4 with local kernels of 3 and channels of 64, 64, 128, 128, 256, 256 respectively, also max-pooling is applied every two blocks. For Kinetics, two additional blocks with channels of 512 are appended, also the local kernels of the first two blocks are changed into 5. The numbers of new coordinate systems K and new joints N' in the transform network are set as 10 and 64 for all datasets.

TABLE I
COMPARING WITH THE STATE-OF-THE-ART APPROACHES IN ACTION RECOGNITION ACCURACY (%) ON THE NTU-RGB+D DATASET

Methods	Year	NTU-CS	NTU-CV
Lie Group [22]	2014	50.1	62.8
H-RNN [54]	2015	59.1	64.0
PA-LSTM [10]	2016	62.9	70.3
ST-LSTM+TG [11]	2016	69.2	77.7
VA-LSTM [8]	2017	79.4	87.6
Synthesized+Pre-trained [55]	2017	80.0	87.2
TS-CNN [24]	2017	83.2	89.3
ST-GCN [14]	2018	81.5	88.3
HCN [13]	2018	86.5	91.1
SR-TSL [9]	2018	84.8	92.4
MANs [56]	2018	83.0	93.2
SGN [16]	2019	86.6	93.4
AS-GCN [15]	2019	86.8	94.2
SLnL-rFA [18] (ours)	2019	**89.1**	**94.9**
SLnL-rFA+ML (ours)	-	**89.7**	**95.4**

B. Experimental Results

To validate the effectiveness and generalization of the proposed SLnL-rFA in constrained and unconstrained environments, we conduct experiments on in total four datasets for skeleton-based action recognition, incuding the NTU RGB+D, Kinetics, SYSU, and N-UCLA datasets. In this section, we compare the performances of the proposed SLnL-rFA and its mutual learning version SLnL-rFA+ML against other state-of-the-art methods. Because there is no previous method with the capability of mining patterns in the frequency domain for skeleton-based action recognition, we only compare our method to the ones in the spatio-temporal domain.

On NTU RGB+D dataset, we compare with one hand-crafted features method [22], four RNN-based methods [8], [10], [11], [54], four CNN-based methods [13], [24], [55], [56], three graph convolutional methods [14]–[16], one graph and LSTM hybridized method [9]. As the local components of our SLnL are CNN-based while the non-local components are designed to learn the affinity degree between each target position (node) to every position (node) in the figure (graph), our SLnL-rFA can be treated as a variant of CNN and graph hybridized method. As shown in Table I, the deep learning methods outperform the hand-crafted method, the CNN-based methods are generally better than LSTM-based methods, and graph-based or graph-hybridized methods also perform well. Our preliminary method SLnl-rFA consistently outperforms the state-of-the-art approaches proposed at the same time by a large margin for both cross-subject (CS) and cross-view (CV) evaluation. Specifically, the SLnL-rFA outperforms the best CNN-based method at that time (HCN) by 2.6% (CS) and 3.8% (CV), also outperforms the best graph-related approach at that time (SR-TSL) by 4.3% (CS) and 2.5% (CV). Moreover, our preliminary SLnL-rFA is even superior to two recently reported works [15], [16]. Finally, through further exploiting the proposed mutual

TABLE II
COMPARING WITH THE STATE-OF-THE-ART APPROACHES IN ACTION RECOGNITION ACCURACY ON THE KINETICS DATASET. BOTH OF THE TOP-1 AND TOP-5 ACCURACIES ARE REPORTED

Methods	Year	top-1 (%)	top-5 (%)
Feature Encoding [57]	2015	14.9	25.8
Deep LSTM [10]	2016	16.4	35.3
Temporal ConvNet [58]	2017	20.3	40.0
ST-GCN [14]	2018	30.7	52.8
AS-GCN [15]	2019	34.8	56.5
SLnL-rFA [18] (ours)	2019	**36.6**	**59.1**
SLnL-rFA+ML (ours)	-	**37.5**	**60.3**

TABLE III
COMPARING WITH THE STATE-OF-THE-ART APPROACHES IN ACTION RECOGNITION ACCURACY (%) ON THE SYSU DATASET REGARDING THE CS AND SS PROTOCOLS

Methods	Year	SYSU-CS	SYSU-SS
VA-LSTM [8]	2017	77.5	76.9
ST-LSTM [59]	2018	76.5	-
DPRL+GCNN [60]	2018	76.9	-
GCA-LSTM [61]	2018	78.6	-
SR-TSL [9]	2018	81.9	80.7
ElAtt-GRU [62]	2018	85.7	85.7
MANs [56]	2018	87.6	-
SGN [16]	2019	86.9	86.5
SLnL-rFA [18] (ours)	2019	**87.7**	**87.4**
SLnL-rFA+ML (ours)	-	**88.3**	**88.1**

learning policy for collaboratively learning, our SLnL-rFA+ML achieves the best accuracy on both the evaluation protocols of CS and CV.

On Kinetics dataset, we compare with five characteristic methods, including hand-crafted features [57], deep LSTM network [10], temporal convolutional network [58], and graph convolutional networks [14], [16]. As shown in Table II, the deep models outperform the hand-crafted features method, and the CNN-based method works better than the LSTM-based method. Our preliminary SLnL-rFA outperforms the state-of-the-art approach proposed at the same time (ST-GCN) by large margins of 5.9% (top1) and 6.3% (top5) for recognition accuracies. The extended version SLnL-rFA+ML again achieves the state-of-the-art performance, which indicates the effectiveness of the proposed method.

On SYSU dataset and N-UCLA dataset, the methods to be compared with also fall into the categories of hand-crafted features methods [22], [63], RNN-based methods [8], [54], [59], [61], [62], [65], CNN-based methods [55], [64], and graph-based methods [9], [16], [60]. The recognition results are shown in Table III and Table IV, respectively. Similarly, the CNN- or graph-based methods currently dominate this task and generally achieve better than hand-crafted features or LSTM-based methods in the early period. Our preliminary method SLnL-rFA and extended version SLnL-rFA+ML consistently outperform other state-of-the-art approaches on both of the two datasets.

TABLE IV
COMPARING WITH THE STATE-OF-THE-ART APPROACHES IN ACTION RECOGNITION ACCURACY ON THE N-UCLA DATASETS

Methods	Year	Accuracy (%)
Lie Group [22]	2014	74.2
Actionlet ensemble [63]	2014	76.0
HBRNN-L [54]	2015	78.5
Visualization CNN [64]	2017	86.1
Ensemble TS-LSTM [65]	2017	89.2
Synthesized+Pre-trained [55]	2017	92.6
ElAtt-GRU [62]	2018	90.7
SGN [16]	2019	92.5
SLnL-rFA (ours) [18]	2019	**93.1**
SLnL-rFA+ML (ours)	-	**93.5**

TABLE V
COMPARISONS OF DIFFERENT TRANSFORM METHODS IN ACCURACY (%)

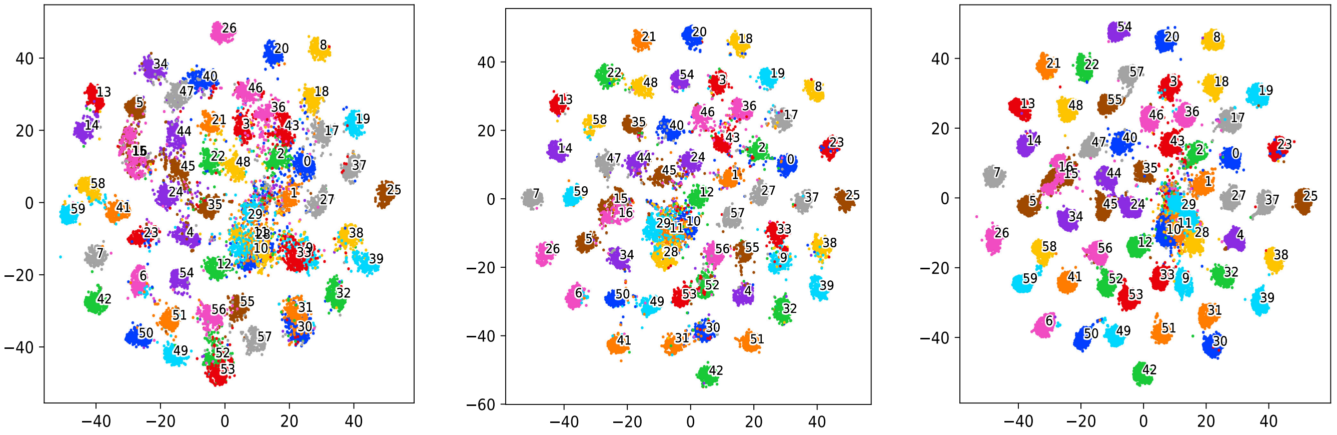
Transform Methods	NTU-CS	NTU-CV	N-UCLA
No Trans. (Baseline$_0$)	84.5	90.6	89.6
CNN Variant	84.7	90.7	89.8
Coordinate Trans.	85.0	91.1	90.2
Skeleton Trans.	85.1	90.9	90.0
Coordinate and Skeleton Trans.	**85.5**	**91.3**	**90.5**

C. Ablation Studies and the Influence of Parameters

To analyze the effectiveness of every component, extensive ablation studies are conducted on the N-UCLA dataset, the NTU RGB+D dataset with cross-view protocol (NTU-CV), and the NTU RGB+D dataset with cross-subject protocol (NTU-CS).

1) Raw Data vs. Transformed Data: The baseline model (Baseline$_0$) of this section contains only local blocks in Fig. 4(b). The baseline network inputs raw data without any transform (No Trans.), and it is optimized with conventional cross entropy loss and separated learning policy. The coordinate and skeleton transform, the coordinate transform, the skeleton transform, and a CNN variant with the same depth are respectively applied to transform the raw data. As shown in Table V, the performances obtained with transformed data consistently outperform that with the raw data. The improvement of coordinate transform indicates that representing action in adaptive multiple oblique coordinate systems is better than the original coordinate system. Also the improvement of skeleton tranform indicates the augmented and rearranged data encode more structure information than the original structureless data. Even with the same depth, the improvement of CNN variant is insignificant, indicating that our improvement is not induced by adding depth. Finally, the coordinate and skeleton transform preforms the best, indicating that the coordinate transform and the skeleton transform are complementary to each other.

2) Comparisons on Loss Function: We firstly further reform the Baseline$_0$ into Baseline$_1$ by adding the above coordinate and skeleton transform network for this section. Then the model is optimized with the cross entropy loss (CE), focal loss (FL), soft-margin cross entropy loss (SMCE), and soft-margin focal loss (SMFL), respectively. To save space, at most two best parameters for each loss are listed in Table VI. Due to the adaptive data selection, the FL performs better than the CE. Benefiting



(a) CE (b) SMCE (c) SMFL

Fig. 8. The 2-dimensional t-SNE visualization of obtained features according to different loss functions on the NTU-CV dataset. (a), (b) and (c) are obtained with the cross entropy loss (CE), soft-margin cross entropy loss (SMCE), and soft-margin cross focal loss (SMFL), respectively. The numbers denote the action labels. Comparing (a) with (b), we can see that the SMCE has larger inter-class distances, indicating that the proposed soft-margin loss term is beneficial to encourage margin between positive samples and negative samples. However, some hard samples in (b) are still seriously confused, such as "10" vs "11," "52" vs "53". Fortunately, our SMFL (c) could alleviate this issue by further combining the advantage of hard sample mining from the focal loss.

TABLE VI
COMPARISONS OF DIFFERENT LOSS FUNCTIONS IN ACCURACY. THE FOCUSING PARAMETER γ AND THE MARGIN PARAMETER m OF LOSSES ARE EXPRESSED AS (γ, m)

Loss Types	NTU-CS (%)	NTU-CV (%)	N-UCLA (%)
CE (Baseline$_1$)	85.5	91.3	90.5
FL (2,)	85.8	91.9	90.9
FL (3,)	85.6	91.8	90.6
SMCE (,0.4)	86.4	92.0	90.8
SMCE (,0.6)	86.2	92.3	91.2
SMFL (2,0.4)	**86.9**	92.5	91.0
SMFL (2,0.6)	86.5	**92.6**	**91.4**

TABLE VII
PERFORMANCE COMPARISONS OF DIFFERENT FREQUENCY ATTENTION METHODS IN HUMAN ACTION RECOGNITION ACCURACY (%)

Frequency Attention Methods	NTU-CS	NTU-CV	N-UCLA
No FA (Baseline$_2$)	86.9	92.6	91.4
Amplitude FA	84.7	89.8	90.1
Sinusoidal FA	87.2	92.8	91.6
Cosine FA	87.1	92.8	91.7
Shared FA	87.3	92.9	91.7
Dependent FA	87.5	93.2	92.0
Residual FA (rFA)	**87.7**	**93.6**	**92.2**
Spatio-temporal Attention	87.3	93.1	91.8

from the encouraged margins between the positive and negative samples, both the SMCE and the SMFL perform better than their original versions CE and FL, respectively. Finally, our SMFL achieves the best for its advantages from adaptive data selection and intrinsic margin encouraging.

To further intuitively understand the advantages of the proposed SMFL, we apply t-SNE [66] to visualize the learned features of different losses, including the losses of CE, SMCE, and SMFL. Comparing Fig. 8(a) with Fig. 8(b), we can see that the feature obtained by our SMCE performs better, and it has larger inter-class and smaller intra-class distances. The result indicates that the proposed soft-margin loss term is beneficial to encourage margin between positive samples and negative samples. However, Fig. 8(b) also shows that some hard samples from several class pairs are still seriously confused, such as "10" vs "11," "52" vs "53," and "30" vs "31". Fortunately, the results in Fig. 8(c) indicate that the proposed soft-margin focal loss could alleviate this issue by further combining the advantage of hard sample mining from the focal loss. As a result, the network optimized with our soft-margin focal loss learned the most discriminative features.

3) How to Select Discriminative Frequency Patterns: We firstly reform the Baseline$_1$ into Baseline$_2$ (No FA) for this section by adding the SMFL. To validate the effectiveness of proposed rFA, we compare it with several variants. The amplitude frequency attention (aFA) is built on frequency spectrum instead of sinusoidal and cosine components. The sinusoidal FA (cosine FA) that uses only sinusoidal (cosine) component. The shared frequency attention (sFA) learns shared attention parameters for sinusoidal and cosine components, while the dependent frequency attention (dFA) learns two set of parameters independently. The rfA is constructed by applying the residual learning trick to dFA in the spatio-temporal domain (Fig. 3). The spatio-temporal attention applies residual attention to original spatio-temporal feature X' in Fig. 3 directly. In Table VII, we can see that the aFA is harmful since the phase angle information is completely missing when only using the frequency spectrum. The dFA outperforms the sFA because that it has more parameters to model the frequency patterns. Besides, we can see that the residual spatio-temporal attention in the spatio-temporal domain can also bring some improvements over Baseline$_2$. In addition, the rFA achieves the best because that its residual trick

TABLE VIII
COMPARISONS OF IMPROVEMENT DIFFERENCES ON THE KINETICS DATASET AND KINETICS-FREQUENCY DATASET. THE TOP-1 ACTION RECOGNITION ACCURACIES (%) ARE REPORTED

Datasets	w/o Atten.	w/ rFA	w/ S.T. Atten.
Kinetics	36.1 (+0)	37.5 (+1.4)	36.7 (+0.6)
Kinetics-Frequency	45.4 (+0)	49.7 (+4.3)	46.5 (+1.1)

can strengthen key frequency patterns without destroying information in the spatio-temporal domain severely. Finally, the rFA outperforms the Baseline$_2$ with a large margin, indicating that the frequency information is effective for skeleton-based action recognition task.

4) Does the rFA Indeed Improve the Performance on Frequency-Related Actions: Inspired by the "Kinetics-Motion" in [14], we select a subset of 30 (out of 400) action classes from the Kinetics dataset that have characteristic frequency patterns (referred as Kinetics-Frequency), which includes actions like "shaking head," "playing drums," and "filling eyebrows". Then, two variants of our SLnL+rFA+ML with rFA (w/ rFA) are constructed by replacing the rFA block with a residual spatio-temporal attention (w/ S.T. Atten.) and directly removing all attention (w/o Atten.), respectively. The results obtained from the original Kinetics dataset and the selected Kinetics-Frequency dataset are shown in Table VIII. We can see that the improvements from rFA are larger than these from residual spatio-temporal attention on both datasets, indicating the effectiveness of the proposed rFA. Moreover, on the selected frequency-related subset, the increment from our rFA is much more significant than that from the spatio-temporal attention, indicating that our rFA indeed improves the recognition performance of frequency-related actions.

5) Comparisons of Methods with Different Affinity Fields: We further reform the Baseline$_2$ into Baseline$_3$ with a rFA block for this section. Although non-local dependencies can be captured in higher layers of hierarchical local networks, it appears to be brutal and with low efficiency. We argue that synchronously explore and fuse non-local information in early stages is more preferable. We merge one temporal non-local module (tSLnL), spatial non-local module (sSLnL), or spatial-temporal non-local block (SLnL) into Baseline$_3$ to examine their effectiveness. As shown in Table IX, both the non-local information from the temporal and spatial dimensions during early stages are helpful. In addition, benefiting from the synchronous fusion of the local details and non-local semantics, the proposed SLnL blocks boosts up the recognition performance w.r.t Baseline$_3$ by 1.1% (NTU-CV), 1.4% (NTU-CS) and 0.9% (N-UCLA), respectively.

To further investigate the properties of deeper SLnL blocks, we replace M_1 local blocks in Baseline$_3$ with SLnL block. Table IX shows more SLnL blocks in lower layers generally lead to better results, but the improvements of higher layers is relatively small because the affinity field of local operations is also increasing with layers. The results clearly show that synchronously extracting local details and non-local semantics is vital for modeling the spatio-temporal dynamics of actions.

TABLE IX
COMPARISONS OF METHODS WITH VARIOUS AFFINITY FIELDS IN ACCURACY (%). M_1 AND M_2 DENOTES THE NUMBERS OF SLNL AND LOCAL BLOCKS IN FIG. 2, RESPECTIVELY. $M_1 + M_2 = 6$

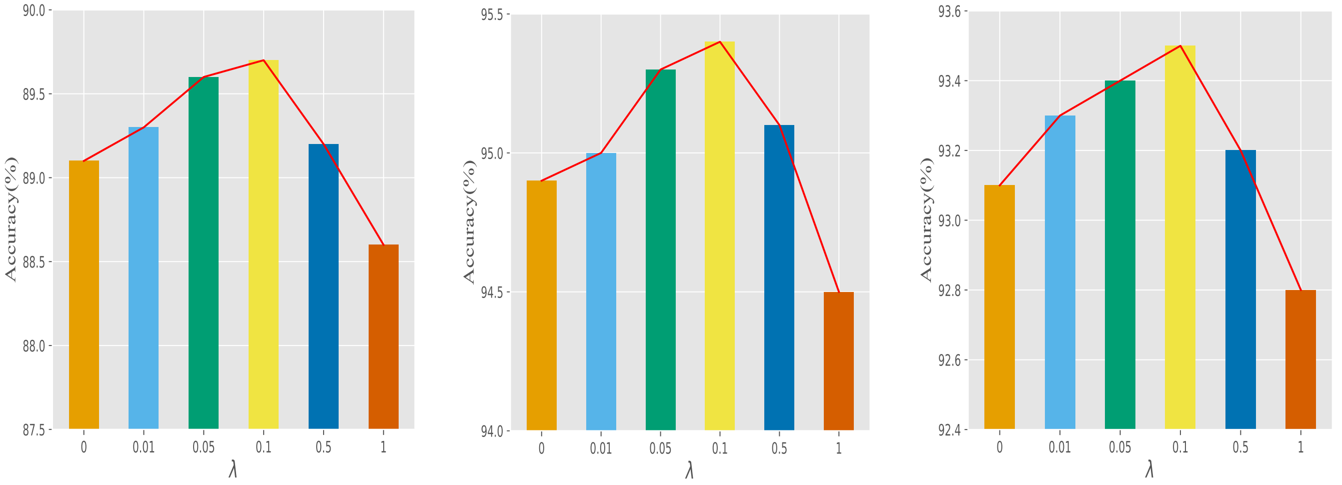
Affinity Fields	NTU-CS	NTU-CV	N-UCLA
Local (Baseline$_3$)	87.7	93.6	92.2
tSLnL (M_1 = 1, M_2 = 5)	88.1	93.9	92.4
sSLnL (M_1 = 1, M_2 = 5)	88.0	94.1	92.3
SLnL (M_1 = 1, M_2 = 5)	88.3	94.3	92.6
SLnL (M_1 = 2, M_2 = 4)	88.6	94.6	92.8
SLnL (M_1 = 3, M_2 = 3)	88.8	**94.9**	**93.1**
SLnL (M_1 = 4, M_2 = 2)	88.9	94.8	92.9
SLnL (M_1 = 5, M_2 = 1)	**89.1**	94.7	93.0
SLnL (M_1 = 6, M_2 = 0)	88.8	94.7	92.9

TABLE X
PERFORMANCE COMPARISONS OF DIFFERENT MULTI-MODAL PROCESSING POLICIES IN HUMAN ACTION RECOGNITION ACCURACY (%)

Processing Policies	NTU-CS	NTU-CV	N-UCLA
Position Feature	87.7	93.8	92.1
Velocity Feature	87.4	93.6	91.9
Feature Summing	88.1	94.0	92.3
Feature Concatenating	88.5	94.3	92.7
MTL w/o ML	89.1	94.9	93.1
MTL w/ ML	**89.7**	**95.4**	**93.5**

6) Comparisons of Different Multi-Modal Processing Policies: In order to be identical to our preliminary work [18], the ablation analyses above are all obtained by optimizing the pseudo multi-task learning task with the conventional separated learning policy (MTL w/o ML) in Fig. 6(c). In this section, we will compare it with other multi-modal processing policies, including feature summing, feature concatenating, and the proposed multi-task learning paradigm with mutual learning policy (MTL w/ ML), also the variants that contain either position or velocity feature are compared. Table X shows all multi-modal fusion policies are superior to single feature methods because they contain more input information. And our pseudo multi-task learning policies perform better than conventional feature summing and feature concatenating policies because concatenated feature branch can obtain additional guide from single-feature branches in the MTL paradigm. Benefiting from the collaborative learning between different feature branches, the proposed multi-task learning method with mutual learning policy achieves the best performance on both datasets. It should be noted that the proposed pseudo multi-task learning policy with mutual learning is not specific for this task, it is potential to be generalized to process other tasks that with multiple feature modalities.

7) Influence of Mutual Controlling Parameter: We explore the effect of key parameter λ in Eq.22, which controls the influence of mutual policy during the multi-task learning. The results in Fig. 9 show the optimal performances are achieved at a middle value of mutual controlling parameter ($\lambda = 0.1$). When the parameter λ is small, the performance is relatively low because individual branches in the network obtained limited collaborative guiding from other branches. Besides, when the parameter λ is too large, it is even harmful for the multi-task learning.



(a) NTU-CV (b) NTU-CS (c) N-UCLA

Fig. 9. The influence of mutual controlling parameter λ towards action recognition accuracies on the NTU RGB+D dataset and the N-UCLA dataset.

TABLE XI
COMPARISONS OF THE MODEL COMPLEXITY, INFERENCE EFFICIENCY, AND RECOGNITION ACCURACY ON THE NTU-CV DATASET

Methods	Year	#Params	#FLOPs	Accuracy
HCN [13][†]	2018	2.648M	0.393G	91.1%
ST-GCN [14]*	2018	3.099M	3.492G	88.3%
Js-AGCN [51]*	2019	3.451M	3.994G	93.7%
SLnL+rFA_pos (ours)	-	4.723M	3.889G	93.8%
2s-AGCN [51]*	2019	6.902M	7.987G	95.1%
SLnL+rFA+ML (ours)	-	9.461M	7.778G	95.4%

[†]The #Params and #FLOPs are calculated from our reproduction (https://github.com/huguyuehuhu/HCN-pytorch).
[*]The #Params and #FLOPs are calculated from the official released code.

In Fig. 9, we can see that the performances of $\lambda = 1$ are worse than conventional separated learning policy ($\lambda = 0$). It is probably because a large mutual loss term also has side effect to force the branches to produce similar features that will reduce diversity. Therefore, at a middle controlling parameter the model can achieve a proper balance between the mutual learning and separated learning.

D. Model Complexity and Inference Efficiency

In this section, we study the model complexity and inference efficiency of the proposed framework. We report two metrics on the NTU-CV dataset for comparing, including the number of network parameters (#Params) and the number of forward floating-point operations (#FLOPs). Since most of the previous approaches didn't contain the analysis of model complexity and inference efficiency, we calculate the two metrics of recent methods [13], [14], [51] that have publicly available codes. The results are shown in Table XI. Note that a 64-frames action video is used to calculate #FLOPs in all methods for fair comparing the inference efficiency. For the recognition performance, we adopt the accuracies reported in original papers which were obtained with 300-frames [14], [51] or 32-frames [13] action videos. For one-stream framework, our SLnL+rFA_pos using only position information of joints has comparable mount of parameters with the state-of-the-art methods Js-AGCN [51] and is slightly faster than it. Although our SLnL+rFA_pos is a little more complex and slower than HCN [13] and ST-CGN [14], its recognition accuracy is much better than them. Finally, compared to the state-of-art two-stream recognition model 2s-AGCN [51], our SLnL+rFA+ML achieves better performance with comparable model parameters and slightly lower computation burden.

V. CONCLUSION

In this work, we proposed a novel model SLnL-rFA to extract synchronous detailed and semantic information from multi-domains for skeleton-based action recognition. The SLnL synchronously extracts local details and non-local semantics in the spatio-temporal domain. The rFA adaptively selects discriminative frequency patterns, which sheds a new light to exploit information in the frequency domain for skeleton-based action recognition. In addition, we proposed a novel soft-margin focal loss, which can encourage intrinsic margins in classifiers and conduct adaptive data selection. Furthermore, we also proposed to put the multi-modal features processing under a pseudo multi-task learning paradigm and proposed a mutual learning policy to optimize the sub-tasks collaboratively. Extensive experimental results on four widely used datasets have shown the superiority of the proposed approach in comparison with other state-of-the-arts methods.

REFERENCES

[1] D. Li, T. Yao, L. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 416–428, Feb. 2019.
[2] N. E. El-Madany, Y. He, and L. Guan, "Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1317–1331, May 2019.

[3] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 20–36.

[4] Y. Zhao *et al.*, "Temporal action detection with structured segment networks," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2933–2942.

[5] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention-based multiview re-observation fusion network for skeletal action recognition," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 363–374, Feb. 2019.

[6] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1038–1050, May 2018.

[7] G. Hu, B. Cui, Y. He, and S. Yu, "Progressive relation learning for group activity recognition," 2019, *arXiv:1908.02948*.

[8] P. Zhang *et al.*, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2136–2145.

[9] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. Eur. Conf. Comput. Vision*, 2018. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/conf/eccv/SiJWWT18

[10] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1010–1019.

[11] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3d human action recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 816–833.

[12] Q. Ke, M. Bennamoun, S. An, F. A. Sohel, and F. Boussaïd, "A new representation of skeleton sequences for 3d action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4570–4579.

[13] 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 *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 786–792.

[14] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Assoc. Advancement Artif. Intell.*, 2018. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/conf/aaai/YanXL18

[15] M. Li *et al.*, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3595–3603.

[16] P. Zhang, C. Lan, W. Zeng, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," 2019, *arXiv:1904.01189*.

[17] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017.

[18] G. Hu, B. Cui, and S. Yu, "Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1216–1221.

[19] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 20–27.

[20] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 1290–1297.

[21] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.

[22] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 588–595.

[23] S. Zhang *et al.*, "Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2330–2343, Sep. 2018.

[24] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, 2017, pp. 597–600.

[25] M. Li and H. Leung, "Multiview skeletal interaction recognition using active joint interaction graph," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2293–2302, Nov. 2016.

[26] S. Kim, K. Yun, J. Park, and J. Y. Choi, "Skeleton-based action recognition of people handling objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2019, pp. 61–70.

[27] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. Boston, MA, USA: Cengage Learning, 2014.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.

[31] C. Beaudry, R. Péteri, and L. Mascarilla, "Action recognition in videos using frequency analysis of critical point trajectories," in *Proc. Int. Conf. Image Process.*, 2014, pp. 1445–1449.

[32] A. C. Cruz and B. Street, "Frequency divergence image: A novel method for action recognition," in *Proc. 14th IEEE Int. Symp. Biomed. Imag.*, 2017, pp. 1160–1164.

[33] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 5619–5628.

[34] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2005, pp. 60–65.

[35] K. Dabov, A. Foi, V. Katkovnik, and K. O. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[36] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. Int. Conf. Comput. Vision*, 2009, pp. 349–356.

[37] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24-1–24-11, 2009.

[38] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/conf/cvpr/Lefkimmiatis17

[39] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/conf/nips/VaswaniSPUJGKP17

[40] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," 2018. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/conf/nips/LiuWFLH18

[41] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 507–516.

[42] X. Wang *et al.*, "Ensemble soft-margin softmax loss for image classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 992–998.

[43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 499–515.

[44] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 815–823.

[45] Y. Fan, F. Tian, T. Qin, J. Bian, and T. Liu, "Learning what data to learn," in *Proc. Int. Conf. Mach. Learn. Workshops*, 2017. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/journals/corr/FanTQBL17

[46] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[47] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.

[48] I. Loshchilov and F. Hutter, "Online batch selection for faster training of neural networks," in *Proc. Int. Conf. Mach. Learn. Workshops*, 2015. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/journals/corr/LoshchilovH15

[49] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2999–3007.

[50] W. Kay *et al.*, "The kinetics human action video dataset," 2017. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/journals/corr/KayCSZHVVGBNSZ17

[51] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 12 026–12 035.

[52] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, Jun. 7–12, 2015, pp. 5344–5352.

[53] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. Zhu, "Cross-view action modeling, learning, and recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Columbus, OH, USA, Jun. 23–28, 2014, pp. 2649–2656.

[54] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1110–1118.

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

[56] C. Xie *et al.*, "Memory attention networks for skeleton-based action recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1639–1645.

[57] B. Fernando, E. Gavves, J. O. M., A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 5378–5387.

[58] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proc. Conf. Comput. Vision Pattern Recognit. Workshops*, 2017, pp. 1623–1631.

[59] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[60] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 5323–5332.

[61] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

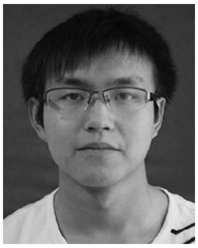
[62] P. Zhang *et al.*, "Adding attentiveness to the neurons in recurrent neural networks," in *Proc. 15th Eur. Conf. Comput. Vision.*, Munich, Germany, September 8–14, 2018, pp. 136–152.

[63] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.

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

[65] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 22–29, 2017, pp. 1012–1020.

[66] L. v. d. Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.



Guyue Hu received the B.E. degree from the Hefei University of Technology, Hefei, China, in 2016. He is currently working toward the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, pattern recognition, and computational neuroscience, especially in video understanding and human activity analysis.

Bo Cui received the M.E degree in information and communication engineering from the Tianjin University, Tianjin, China, in 2015. He is currently working toward the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, pattern recognition, and machine learning.

Shan Yu received the B.S. and Ph.D. degrees in biology from the University of Science and Technology of China, Hefei, China, in 2000 and 2005, respectively. From 2005 to 2014, he conducted Postdoctoral Research with the Max-Planck Institute of Brain Research, Germany (2005–2008), and the National Institute of Mental Health, USA (2008–2014). In 2014, he was with the Institute of Automation, Chinese Academy of Sciences (CASIA), as a recipient of the "One Hundred Talents" program of CAS. He is a Professor with the Brainnetome Center and National Laboratory of Pattern Recognition (NLPR), CASIA. Since 2018, he has been the Deputy Director with the NLPR. He has authored and coauthored more than 30 peer-reviewed papers in neuroscience and other interdisciplinary fields at leading international journals such as the *Nature Machine Intelligence*, the *Journal of Neuroscience, eLife*, etc. His current research interests include neuronal information processing, brain-inspired computing, and artificial intelligence.