



哈爾濱工業大學  
HARBIN INSTITUTE OF TECHNOLOGY

智能检测与控制实验室 Intelligent Detection and Control Lab

# Int&Int: A Two-Pathway Network for Skeleton-Based Action Recognition

**Xiangyuan Qi\***, Zhen He, Qiang Wang

Department of Control Science and Engineering

Harbin Institute of Technology, Harbin, China

xiangyuan202103@163.com, hezhen@hit.edu.cn, wangqiang@hit.edu.cn

# CONTENTS

1

## INTRODUCTION

Research Status  
Our Framework  
Contributions

2

## *Int&Int* NETWORK

Poses Extraction  
Two Pathways  
Late Fusion

3

## MOTION INTENSITY

Joint-Defined  
Bone-Defined  
Angle-Defined  
JBA-Defined

4

## EXPERIMENTS

Datasets and Settings  
Comparisons with SOTA Methods  
Ablation Study

# 1

## Research Status

- Mainly focus on the local or global information **alone**.
- Cannot focus on both the intensity and the integrity.

### local dynamics

Two-Stream<sup>[1]</sup>

TDD<sup>[2]</sup>

C3D<sup>[3]</sup>

Res3D<sup>[4]</sup>

Slowfast<sup>[5]</sup>

### global information

TSN<sup>[6]</sup>

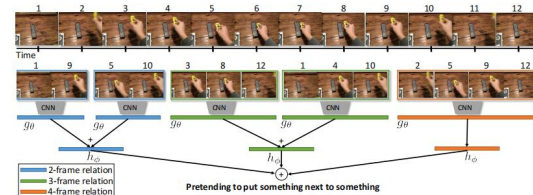
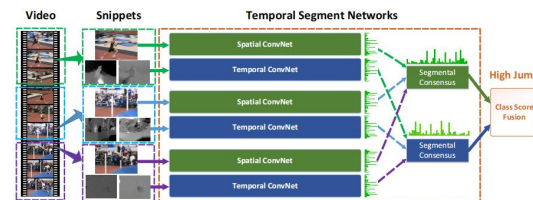
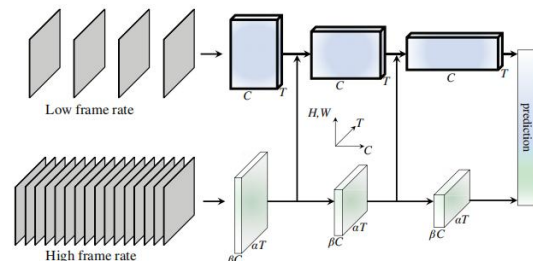
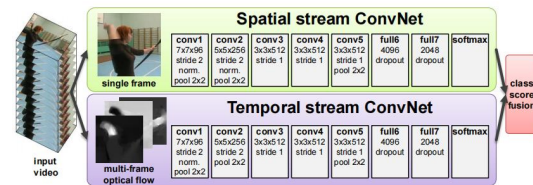
TRN<sup>[7]</sup>

TLE<sup>[8]</sup>

ActionVLAD<sup>[9]</sup>

LTC<sup>[10]</sup>

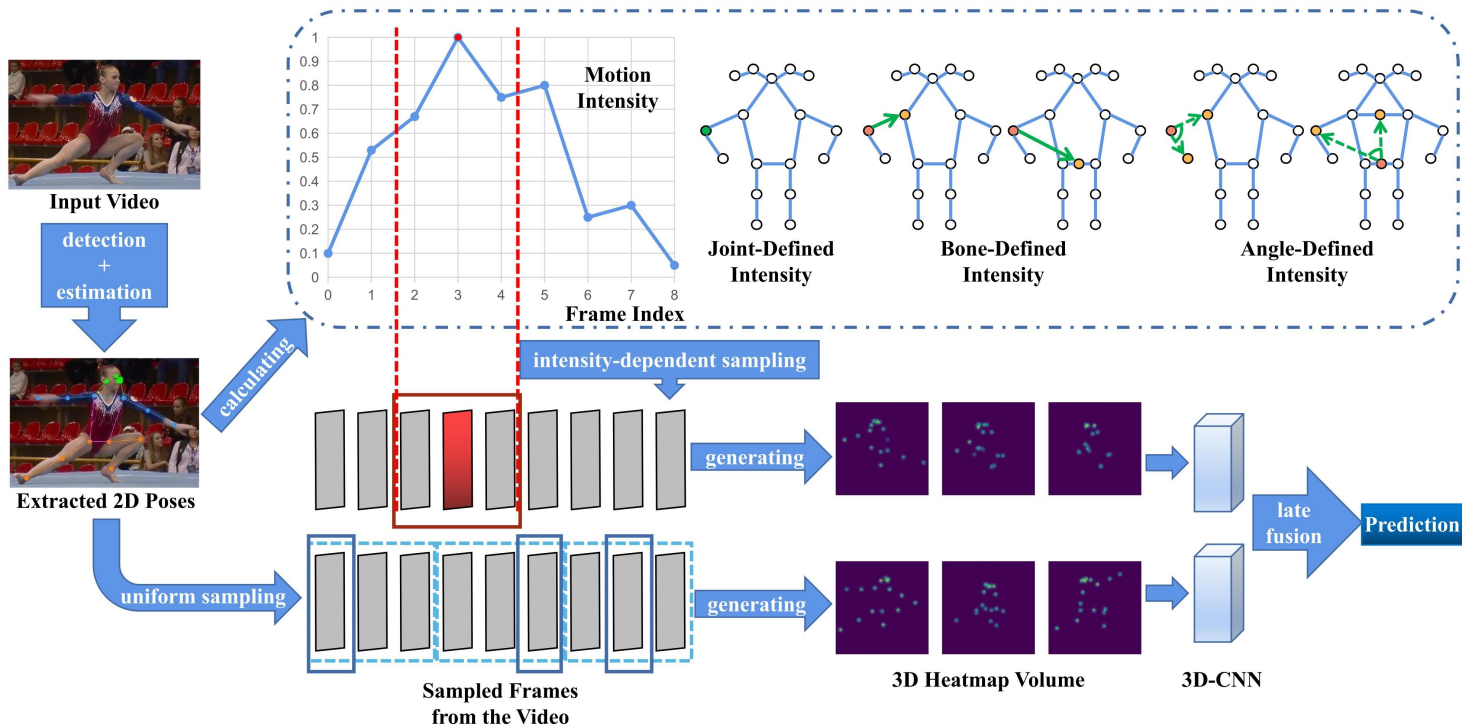
Timeception<sup>[11]</sup>



# 1

## Our Framework

We propose a two-pathway *Int&Int* network (*Intensity&Integrity*) for skeleton-based action recognition to satisfy both aspects.



Research Status

► Our Framework

Contributions

# 1

## Contributions

Research Status

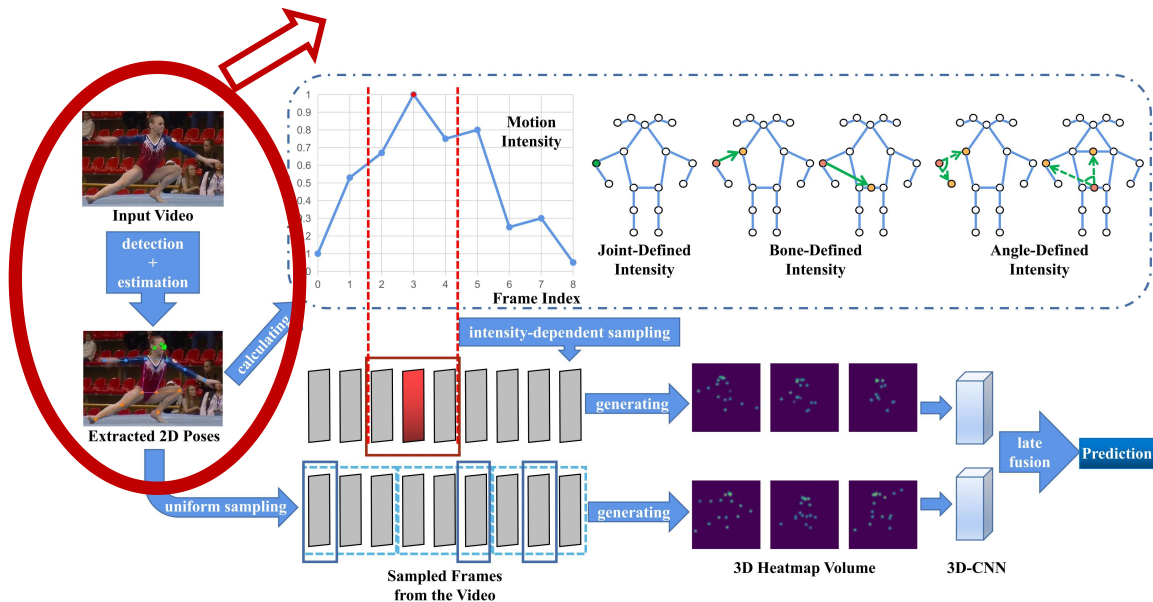
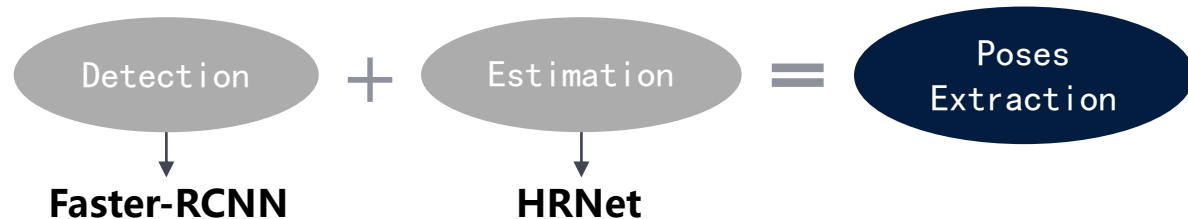
Our Framework

► Contributions

- 1) A two-pathway *Int&Int* framework is proposed which concentrates on the local and global information of the action, where there is great complementarity in the two pathways.
- 2) A clip whose position depends on the most intense frame is sampled as for *Intensity* pathway, where there is a clear physical meaning in the sampling process.
- 3) The motion intensity of each frame is defined in 3 + 2 ways containing different semantic information.

# 2

## Poses Extraction



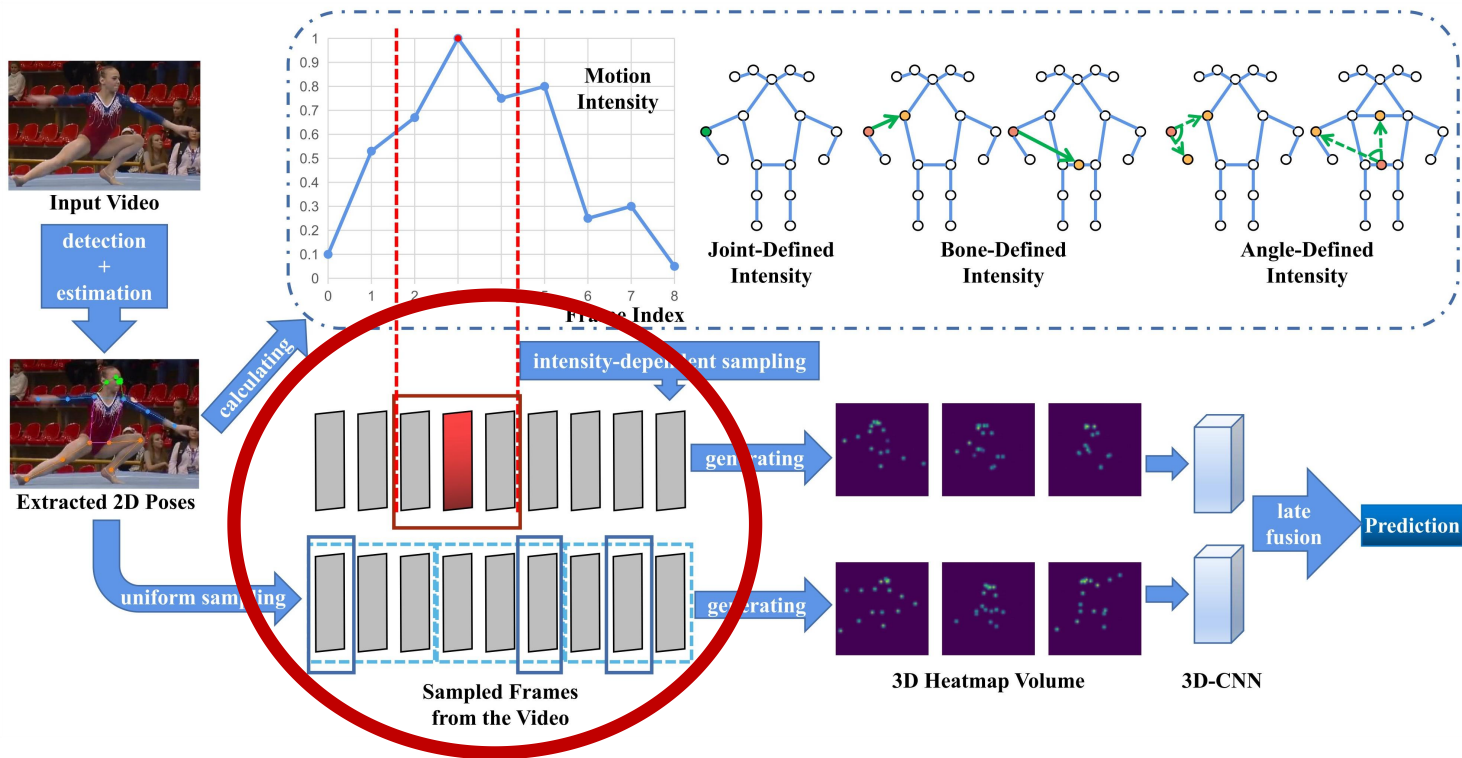
# 2

## Two Pathways

Poses Extraction

Two Pathways

Late Fusion



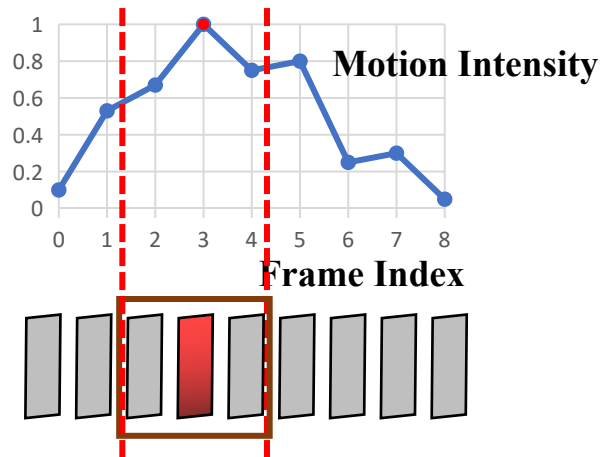
# 2

Poses Extraction

► Two Pathways

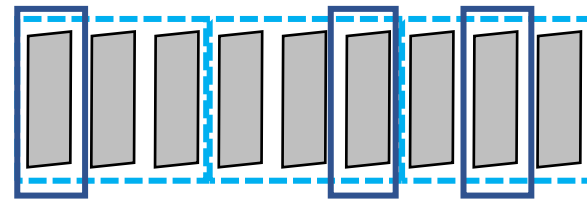
Late Fusion

## Two Pathways



### Intensity-dependent sampling in Intensity pathway

- Motion features are **non-uniformly** distributed along the time axis.
- Motions with large amplitude and quick changes have more action-specific semantic information for identifying the action intuitively.
- Intensity-dependent sampling can avoid missing the salient features:
  - 1) extracted 2D poses --> motion intensity of each frame in the video
  - 2) select the largest element corresponding to the most intense frame
  - 3) sample a clip composed of consecutive frames around the most intense frame



Sampled Frames  
from the Video

### Uniform sampling in Integrity pathway

- Maintain the global information of video:
  - 1) one input video is divided into n segments of equal length (n frames to sample)
  - 2) one frame is randomly selected from each segment



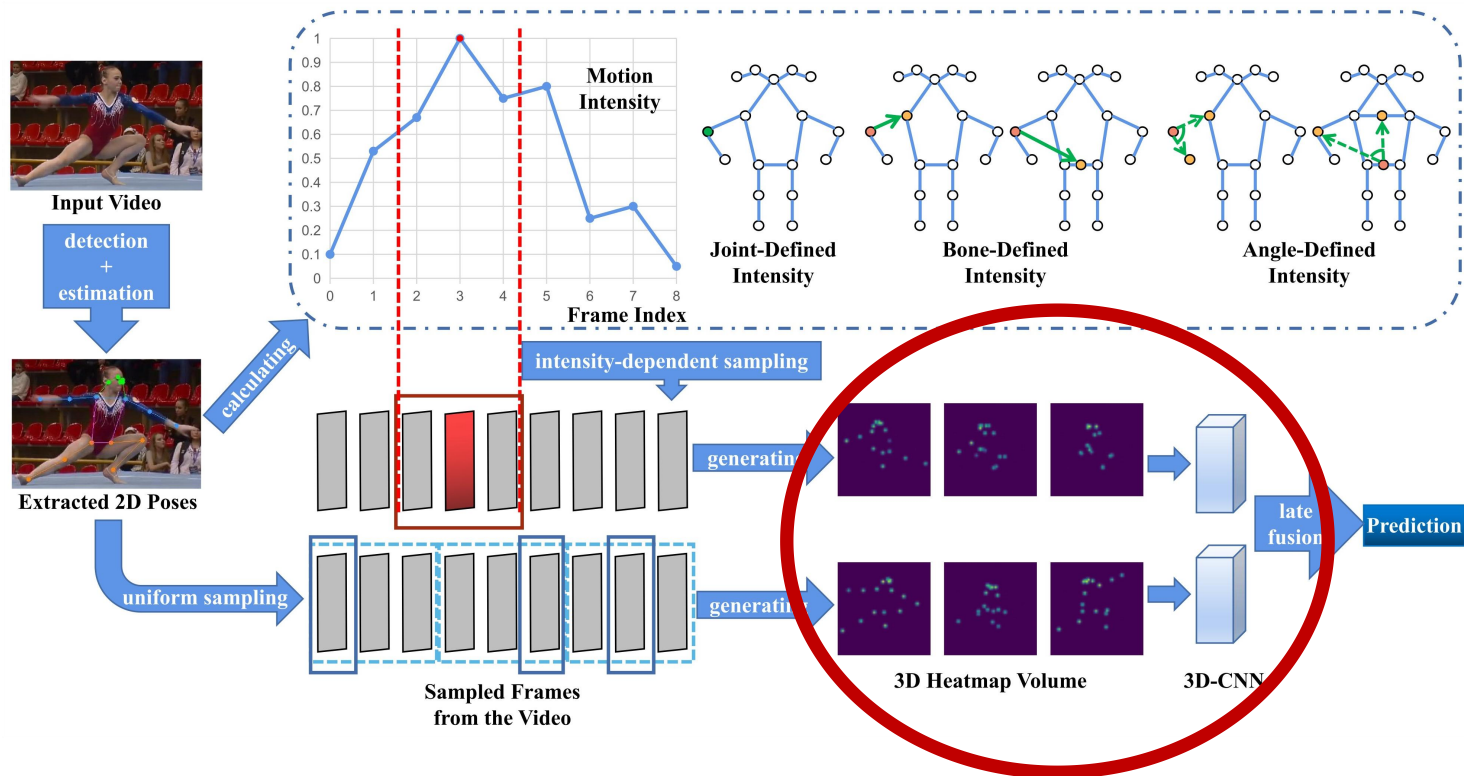
# 2

## Two Pathways

Poses Extraction

Two Pathways

Late Fusion



# 2

## Two Pathways

### 3D heatmap volumes

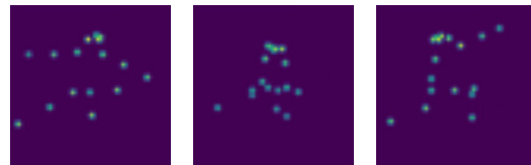
- We generate a joint heatmap  $J$  by composing  $K$  gaussian maps centered at every joint, which are based on joint coordinates and confidence:

$$J_{kij} = e^{-\frac{(i-x_k)^2 + (j-y_k)^2}{2\sigma^2}} * C_k$$

- In multi-person case, the maps of persons are directly accumulated.
- For all sampled  $T$  frames, we stack all heatmaps along the temporal dimension to obtain a 3D heatmap volume of size  $K \times T \times H \times W$  ( $K$ : number of joints,  $H$  and  $W$ : height and width)
- Subjects-centered cropping: reduce the redundancy of volumes.

### 3D-CNN of two pathways

- We use the pseudo heatmap volumes as the input.
- $T \times S^2$ ,  $C$  denote the dimensions of kernels for temporal, spatial, channel sizes.
- We choose SlowOnly, obtained by inflating the ResNet layers in the last two stages from 2D to 3D, to instantiate the backbone.
- Both pathways have the same architecture, and they train the individual losses respectively.



3D Heatmap Volume

Stage	Pathway	Output Size $T \times S^2$
Data Layer	$32, 4^2$	$32 \times 56^2$
Stem Layer	Conv $1 \times 7^2, 32$ Stride 1, $1^2$	$32 \times 56^2$
ResNet3	$\begin{bmatrix} 1 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 4$	$32 \times 28^2$
ResNet4	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 6$	$32 \times 14^2$
ResNet5	$\begin{bmatrix} 3 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 3$	$32 \times 7^2$
GAP	GlobalAveragePooling	#Classes

Poses Extraction

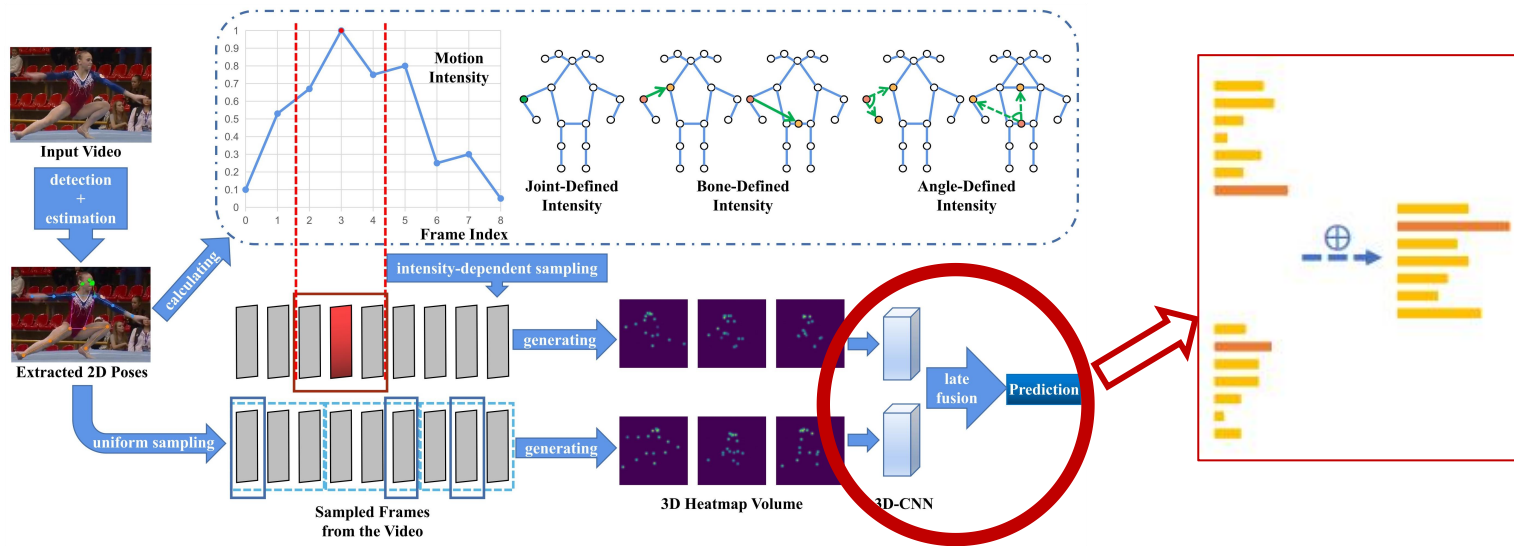
Two Pathways

Late Fusion

# 2

## Late Fusion

The softmax scores of the two pathways are added to obtain the fused score and predict the corresponding action category.



Poses Extraction

Two Pathways

► Late Fusion

# 3

## Joint-Defined

The final value of joint-defined motion intensity is the arithmetic average of values calculated and normalized in two ways respectively.

### 1) Using corresponding position in the next frame

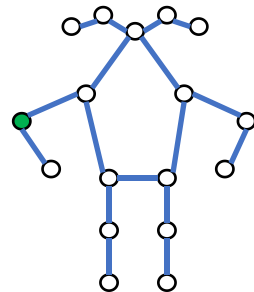
- calculate the distance between corresponding positions of adjacent frames
- $J_{ij}^t$  means the position of the  $i$ -th keypoint of the  $j$ -th person in the  $t$ -th frame.  $K$  is the number of keypoints of the skeleton.  $M$  is the number of people in the frame:

$$SV_{Jnf}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} |J_{ij}^{t+1} - J_{ij}^t|}{M}$$

### 2) Using average position across the frames

- calculate the distance between the  $i$ -th keypoint in the  $t$ -th frame and the average position of the keypoint
- $\bar{J}_{ij}$  means the average position of the  $i$ -th keypoint of the  $j$ -th person across all frames:

$$SV_{Jap}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} |J_{ij}^t - \bar{J}_{ij}|}{M}$$



# 3

## Bone-Defined

The final value of bone-defined motion intensity is the arithmetic average of  $2 \times 2$  values (2 ways to calculate  $\times 2$  types of bone features) which have been normalized.

### 1) Using corresponding bone-vector in the next frame

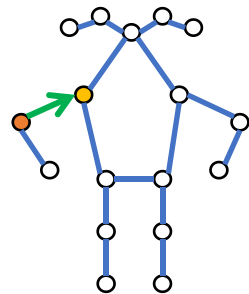
- calculate the angle between bone-vector  $B^t = \overrightarrow{J^t v^t}$  and  $B^{t+1} = \overrightarrow{J^{t+1} v^{t+1}}$
- $B_{ij}^t$  means the  $i$ -th bone-vector of the  $j$ -th person in the  $t$ -th frame:

$$SV_{Bnf}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} \left( 1 - \frac{B_{ij}^t \cdot B_{ij}^{t+1}}{|B_{ij}^t| |B_{ij}^{t+1}|} \right)}{M}$$

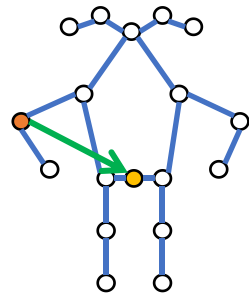
### 2) Using average bone-vector across the frames

- calculate the angle between bone-vector  $B^t = \overrightarrow{J^t v^t}$  and  $\bar{B} = \overrightarrow{J \bar{v}}$
- $\bar{B}_{ij}$  means the  $i$ -th average bone-vector of the  $j$ -th person across all frames:

$$SV_{Bap}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} \left( 1 - \frac{B_{ij}^t \cdot \bar{B}_{ij}}{|B_{ij}^t| |\bar{B}_{ij}|} \right)}{M}$$



Locally-Defined



Center-Oriented

Joint-Defined

► Bone-Defined

Angle-Defined

JBA-Defined

# 3

## Angle-Defined

The final value of angle-defined motion intensity is the arithmetic average of  $2 \times 2$  values (2 ways to calculate  $\times 2$  types of angular features) which have been normalized.

### 1) Using corresponding angle in the next frame

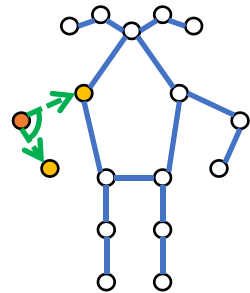
- calculate the difference between corresponding angles of adjacent frames
- $A_{ij}^t = 1 - \cos \theta_{ij}^t$  where  $\theta_{ij}^t$  means the  $i$ -th angle of the  $j$ -th person in the  $t$ -th frame:

$$SV_{Anf}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} |A_{ij}^{t+1} - A_{ij}^t|}{M}$$

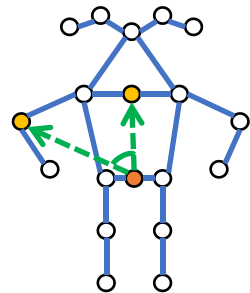
### 2) Using average angle across the frames

- calculate the difference between the  $i$ -th angle in the  $t$ -th frame and average angle
- $\bar{A}_{ij} = 1 - \cos \bar{\theta}_{ij}$  where  $\bar{\theta}_{ij}$  means the  $i$ -th average angle of the  $j$ -th person across all frames:

$$SV_{Aap}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} |A_{ij}^t - \bar{A}_{ij}|}{M}$$



**Locally-Defined**



**Center-Oriented**

Joint-Defined

Bone-Defined

► Angle-Defined

JBA-Defined

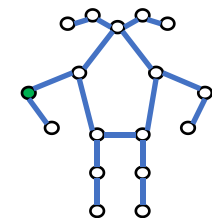
# 3

## JBA-Defined

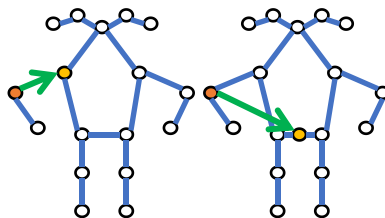
- The values of motion intensity calculated in three definitions need to be normalized because three definitions have different dimensions.
- Two kinds of methods to ensemble them:

1) **Arithmetic average**

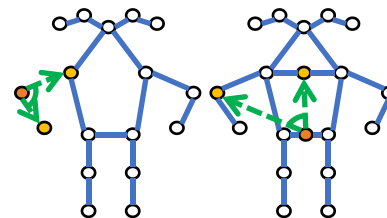
2) **Weighted average: utilize the normalized two-pathway accuracy we get from each definition as the weight of the corresponding definition**



**Joint-Defined  
Intensity**



**Bone-Defined  
Intensity**



**Angle-Defined  
Intensity**

Joint-Defined

Bone-Defined

Angle-Defined

► JBA-Defined

# 4

## ► Datasets & Settings

Comparisons with  
SOTA Methods

Ablation Study

### Datasets

#### 1) FineGYM-99:

- 29K videos of 99 fine-grained gymnastic action classes

#### 2) HMDB-51:

- 51 distinct action categories, each containing at least 101 clips for a total of 6766 video clips extracted from a wide range of sources

### Settings

#### 1) Pytorch framework, 4 NVIDIA GTX 3090 GPUs

#### 2) For FineGYM-99:

- learning rate: 0.4, weight decay: 0.0003, learning rate adjustment: cosine annealing strategy, number of epochs: 80, model: trained from scratch, ResNet blocks setting is (4,6,3)

#### 3) For HMDB-51:

- learning rate: 0.01, weight decay: 0.0001, learning rate adjustment: step strategy, number of epochs: 12, model: pretrained on Kinetics400, ResNet blocks setting is (3,4,6)

#### 4) For both datasets:

- batch size: 16, optimizer: SGD, momentum: 0.9, the length of sampled clip: 48

The code has been shown at <https://github.com/SarahQi666/Int-and-Int>.



# 4

Datasets&Settings

► Comparisons with  
SOTA Methods

Ablation Study

## Comparisons with SOTA Methods

- Although these prior works all use at least videos as the input which contain much more information than our skeleton-based method, our results achieve superior performance and outperform the results of SOTA methods.

FineGYM-99

Method	Accuracy
TSN(RGB) [6] [19]	74.8%
TSN(RGB+Flow) [6] [19]	86.0%
TRN(RGB) [7] [19]	79.9%
TRN(RGB+Flow) [7] [19]	87.4%
ActionVLAD [9] [19]	69.5%
Int&Int(Ours)	<b>95.6%</b>

HMDB-51

Method	Accuracy
Two-Stream(fusion by SVM) [1]	59.4%
Two-Stream(fusion by averaging) [1]	58.0%
TDD [2]	63.2%
C3D [3] [8]	56.8%
TSN(RGB+Flow) [6]	68.5%
TSN(RGB+Flow+Warped Flow) [6]	69.4%
TLE(FC-Pooling) [8]	68.8%
ActionVLAD [9]	66.9%
LTC(RGB+Flow) [10]	64.8%
Int&Int(Ours)	<b>69.6%</b>

- For FineGYM-99, we report the Top-1 accuracy. For HMDB-51, we report average accuracy over three splits.

## Ablation Study

## 1) Single-pathway vs. Two-pathway

- The two-pathway architecture almost consistently outperforms the single-pathway architectures.

Sampling	HMDB-51											
	3 Splits			Split 1			Split 2			Split 3		
	Int&Int	Intensity Pathway	Integrity Pathway	Int&Int	Intensity Pathway	Integrity Pathway	Int&Int	Intensity Pathway	Integrity Pathway	Int&Int	Intensity Pathway	Integrity Pathway
Random	<b>69.3%</b>	63.7%	68.6%	<b>70.2%</b>	64.4%	69.2%	<b>69.0%</b>	63.0%	68.6%	<b>68.6%</b>	63.7%	68.0%
Joint-Defined	<b>69.6%</b>	63.9%	68.6%	<b>70.5%</b>	64.5%	69.2%	<b>69.3%</b>	63.5%	68.6%	<b>69.0%</b>	63.5%	68.0%
Bone-Defined	<b>69.0%</b>	64.2%	68.6%	69.1%	63.9%	<b>69.2%</b>	<b>68.7%</b>	63.7%	68.6%	<b>69.2%</b>	65.0%	68.0%
Angle-Defined	<b>69.5%</b>	63.4%	68.6%	<b>69.3%</b>	63.9%	69.2%	<b>70.3%</b>	62.6%	68.6%	<b>68.8%</b>	63.6%	68.0%
JBA(Arithmetic)	<b>69.3%</b>	63.6%	68.6%	68.6%	63.4%	<b>69.2%</b>	<b>70.1%</b>	63.2%	68.6%	<b>69.2%</b>	64.3%	68.0%
JBA(Weighted)	<b>69.2%</b>	63.9%	68.6%	<b>69.6%</b>	64.0%	69.2%	<b>69.7%</b>	64.5%	68.6%	<b>68.2%</b>	63.1%	68.0%

## 2) Randomly sampling vs. Intensity-dependent sampling

- Randomly sampling: the central frame in the sampled clip is not the most intense frame of the video but a frame randomly selected.
- Intensity-dependent sampling consistently outperforms sampling from a random temporal window.

Sampling	FineGYM-99	HMDB-51			
		3 Splits	Split 1	Split 2	Split 3
Random	95.3%	69.3%	70.2%	69.0%	68.6%
Intensity	<b>95.6%</b>	<b>69.6%</b>	<b>70.5%</b>	<b>70.3%</b>	<b>69.2%</b>

### 3) Intensity defined through motion prior knowledge

- Various definitions place emphasis on different semantic information:
  - a) The positions of joints carry the most direct and obvious information of motion.
  - b) The lengths and directions of bones contain higher-order information, which have been proven to be effective in 2s-AGCN.
  - c) Angles can maintain invariance against different human body sizes and discriminate actions sharing similar motion trajectories such as taking off glasses and taking off headphones.

Sampling	FineGYM-99	HMDB-51			
		<i>3 Splits</i>	<i>Split 1</i>	<i>Split 2</i>	<i>Split 3</i>
Joint-Defined	95.0%	<b>69.6%</b>	<b>70.5%</b>	69.3%	69.0%
Bone-Defined	95.2%	69.0%	69.1%	68.7%	<b>69.2%</b>
Angle-Defined	95.5%	69.5%	69.3%	<b>70.3%</b>	68.8%
JBA(Arithmetic)	95.5%	69.3%	68.6%	70.1%	<b>69.2%</b>
JBA(Weighted)	<b>95.6%</b>	69.2%	69.6%	69.7%	68.2%

- For the professional and flexible gymnastics, it is necessary to combine the information from three aspects because of the richness of fine-grained action information.
- For the actions in HMDB-51, which are generally simpler, the information from only one aspect is enough to capture the feature.

# CONCLUSION

- In this work, we propose a two-pathway *Int&Int* network for skeleton-based action recognition. The final model achieves superior performance on both of the datasets.
- In the future:
  - a) reduce the computational complexity of our network and further improve the performance
  - b) conduct research on the applications such as security, human-computer interaction, physical training, medical rehabilitation, and entertainment



哈爾濱工業大學  
HARBIN INSTITUTE OF TECHNOLOGY

智能检测与控制实验室 Intelligent Detection and Control Lab

# Thanks for your attention!

**Xiangyuan Qi\***, Zhen He, Qiang Wang

Department of Control Science and Engineering

Harbin Institute of Technology, Harbin, China

xiangyuan202103@163.com, hezhen@hit.edu.cn, wangqiang@hit.edu.cn