Skeleton-Based Action Recognition with Synchronous Local and Non-local Spatio-temporal Learning and Frequency Attention

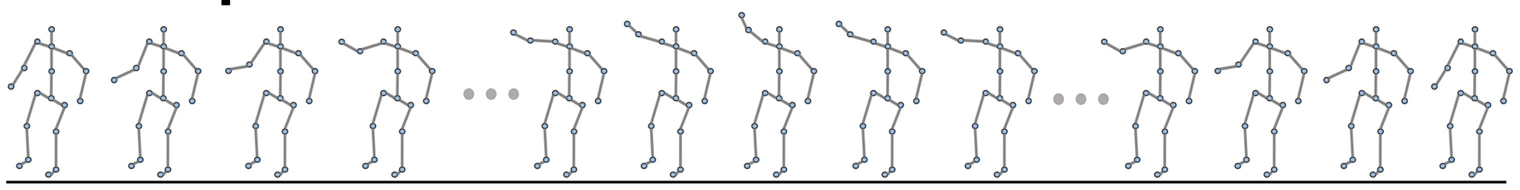
Guyue Hu, Bo Cui, Shan Yu
Institute of Automation, Chinese Academy of Sciences (CASIA)

Shanghai, China
July 8-12, 2019

1. Introduction

Task: Skeleton-based action recognition

- **Input:** Skeleton sequence
- **Output:** Action Class

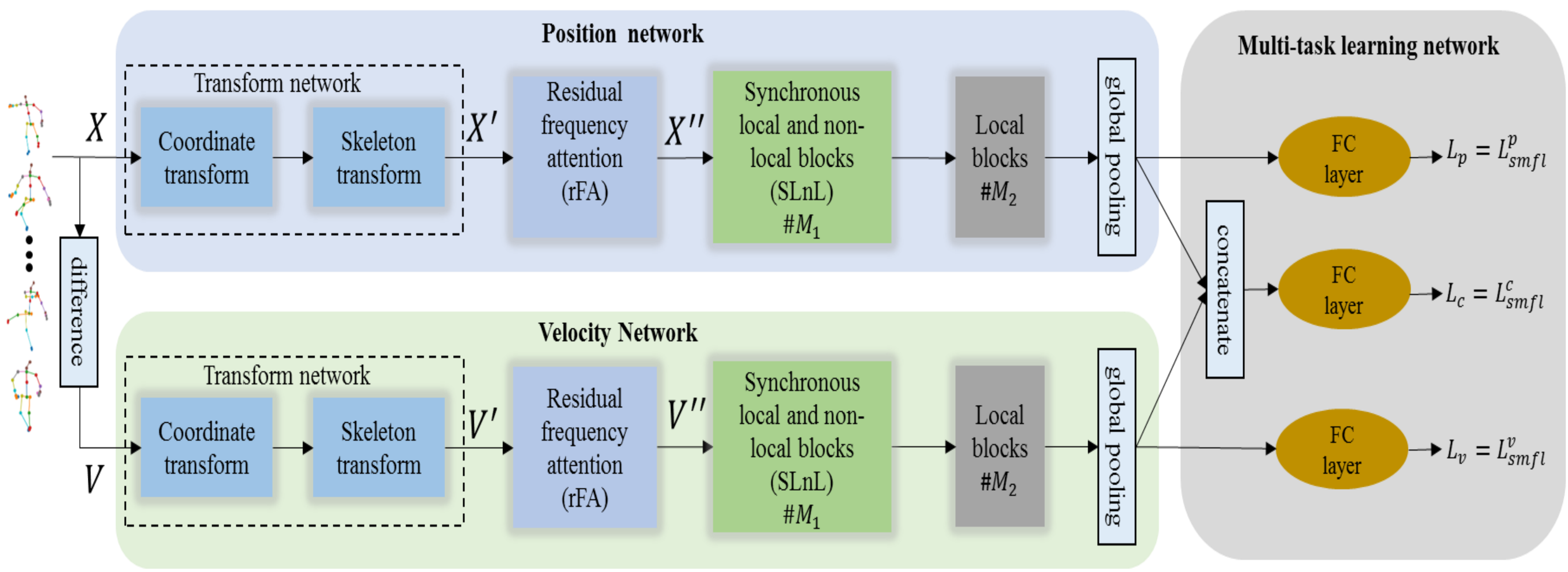


Label: Waving Right Hand

Motivation:

- Some actions (e.g, *clapping, brushing, shaking*) contain characteristic frequency patterns.
- The local detailed information and non-local semantic information are captured asynchronously in lower and higher layers of local networks like RNN, CNN,GCN).
- Classification: discrimination difficulties are different across samples and classes, why not conduct data selection and margin encouraging?

2. Overview



- The rFA selects discriminative frequency patterns in the frequency domain.
- SLnL simultaneously extracts local details and non-local semantics in the spatio-temporal domain.
- Soft-margin focal loss (SMFL) selects data during training and encourages soft-margins in classifiers
- Enrolling the multi-feature branches network in a pseudo multi-task learning paradigm.

3. Preliminary

- *Coordinate Transform*

Adaptively augment the number of coordinate system:*1* to *K*.

$X \in R^{3 \times T \times N} \dashrightarrow R^{3K \times T \times N}$

- *Skeleton Transform*

Adaptively augment the number of joints: *N* to *N'*.

$R^{3K \times T \times N} \dashrightarrow X' \in R^{3K \times T \times N'}$

4. Residual Frequency Attention

- *Fast Fourier Transform*

$Y' = fft2(X') = F_{sin} + jF_{cos}$

- *Attention*

$M_{sin} = f_{attention}(F_{sin})$

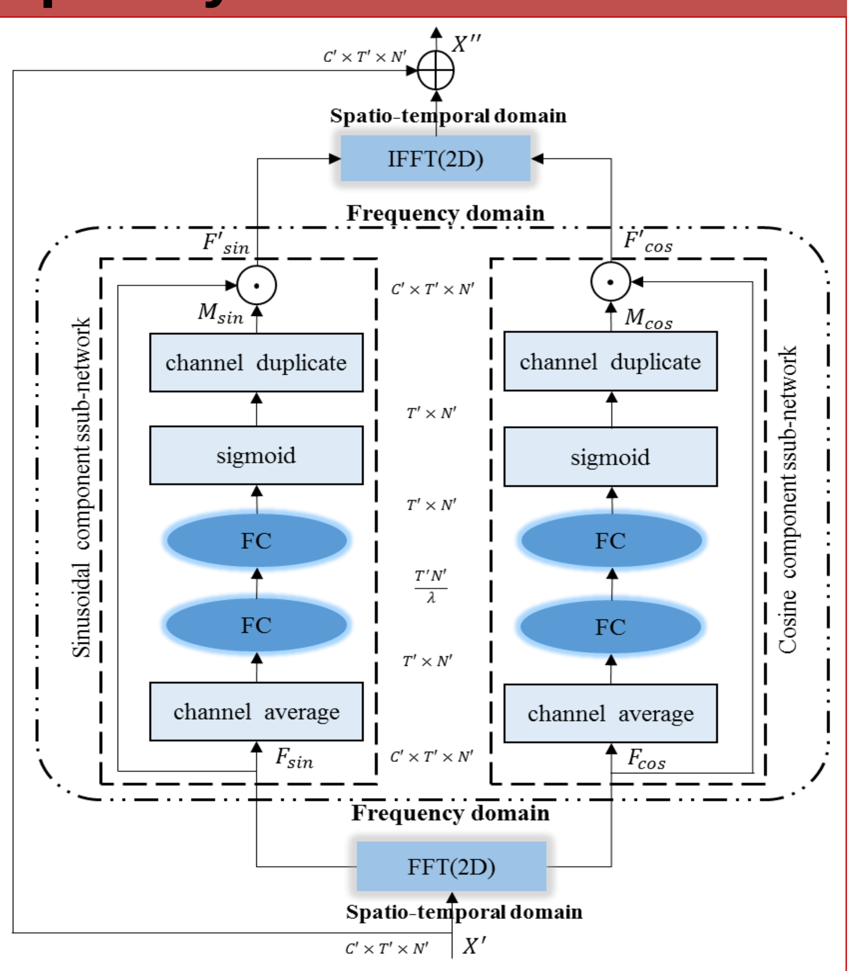
$M_{cos} = f_{attention}(F_{cos})$

$F'_{sin} = F_{sin} \odot M_{sin}$

$F'_{cos} = F_{cos} \odot M_{cos}$

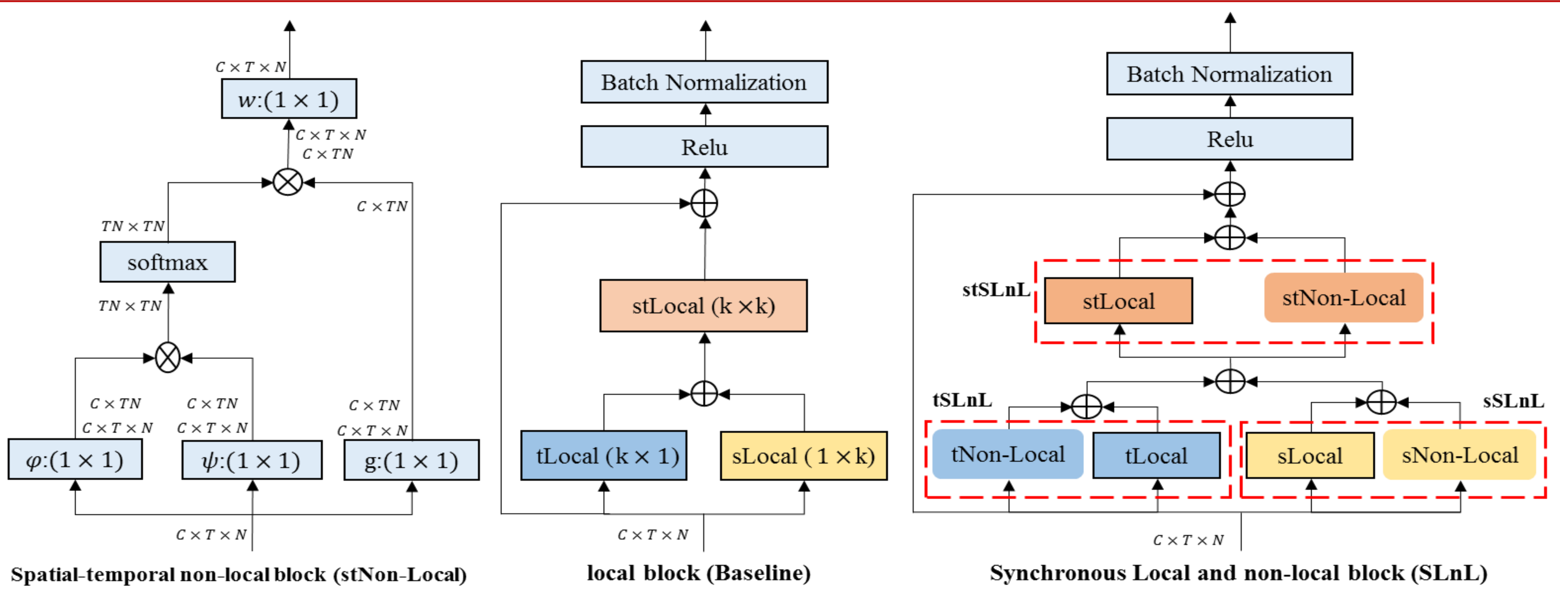
- *Residual trick & Inverse Fast Fourier Transform*

$X'' = X' + ifft2(F'_{sin}, F'_{cos})$



Strengthening key frequency patterns without severely destroying information in the spatio-temporal domain.

5. Synchronous Local and Non-Local Block



- *Non-local module*[1]

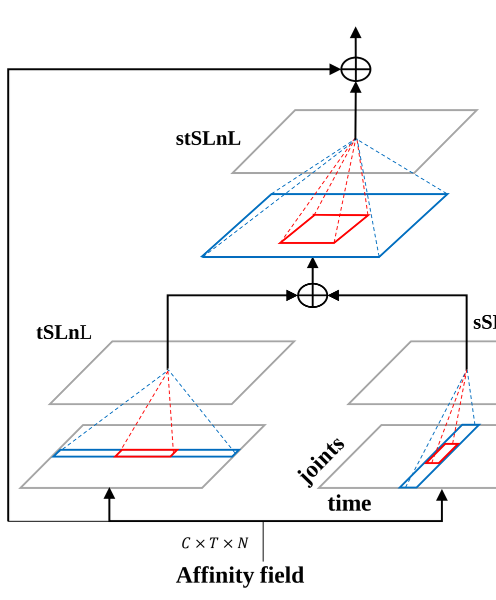
$$y_i = \frac{1}{Z_i(X)} \sum_{j \in \Omega} \phi(x_i, x_j) g(x_j)$$

- *Local module* (*CNN in this paper*)

$$y_i = \frac{1}{Z_i(X)} \sum_{j \in \delta_i} \phi(x_i, x_j) g(x_j)$$

- *SLnL module*

Non-local module and Local module operate in parallel.



- *SLnL Block (tSLnL + sSLnL + stSLnL)*

3SLnL modules along temporal, spatial and spatio-temporal dimensions, respectively.

Contrasting to conventional local networks or non-local networks, SLnL module can extract local details & non-local semantics synchronously.

6. Soft-margin focal Loss

- *Soft-margin term(SM)*

$L_{sm}(p_t) = \log(e^m + (1 - e^m)p_t)$

Larger punishment for samples with smaller posterior

- *SM cross entropy(SMCE)*

$L_{smce}(p_t) = L_{sm} + L_{ce}$

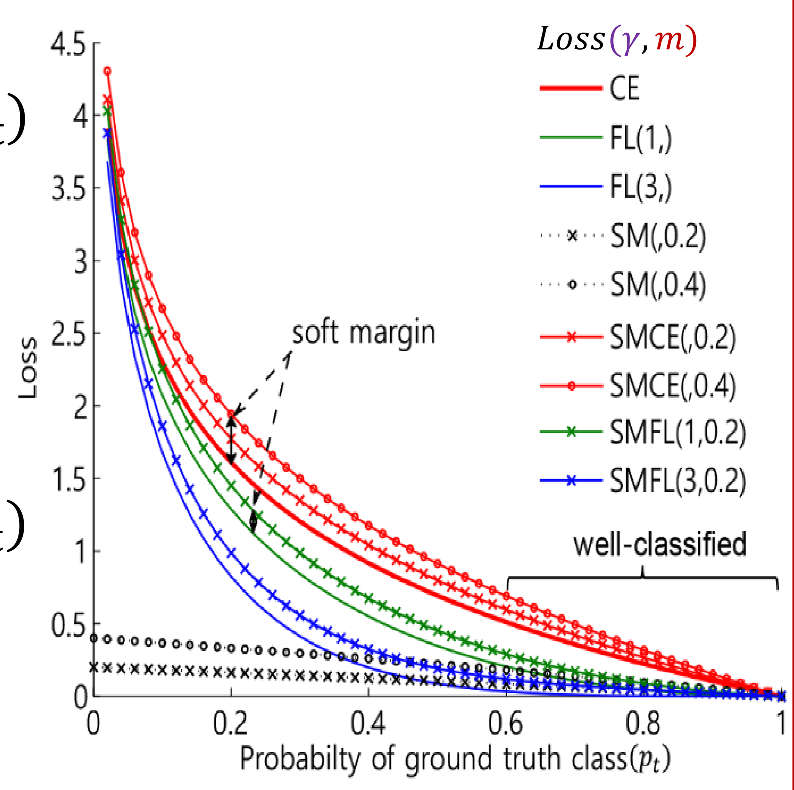
$= \log(e^m + (1 - e^m)p_t) - \log(p_t)$

$= -\log(\frac{e^{w_t x - m}}{e^{w_t x - m} + \sum_{c \neq t}^{C} e^{w_c x}})$

- *SM Focal loss*[2] *(SMFL)*

$L_{smfl}(p_t) = L_{sm} + L_{fl} = \log(e^m + (1 - e^m)p_t) - (1 - p_t)^{\gamma} \log(p_t)$

Margin-encouraging term   Data-selecting term



- *Pseudo multi-task learning with proposed SMFL*

$$L = L_{smfl}^{p} + L_{smfl}^{v} + L_{smfl}^{c}$$

Conduct margin encouraging and data selection within loss without destroying epoch-based training process.

7. Results

- *The comparisons on the NTU-RGB+D & Kinetics*

Methods	CS	CV
VA-LSTM (2017)	79.4	87.6
ST-GCN(2018)	81.5	88.3
HCN(2018)	86.5	91.1
SR-TSL (2018)	84.8	92.4
SLnL-rFA	**89.1**	**94.9**

Methods	Top1	top5
Feature Enc. (2015)	14.9	25.8
Deep LSTM (2016)	16.4	35.3
Tem. Conv1Net (2017)	20.3	40.0
ST-GCN (2018)	30.7	52.8
SLnL-rFA	**36.6**	**59.1**

- *Ablation studies on the NTU-RGB+D dataset*

Loss Types	CS	CV
CE (Baseline1)	85.5	91.3
FL (2,)	85.8	91.9
FL (3,)	85.6	91.8
SMCE (, 0.4)	86.4	92.0
SMCE (, 0.6)	86.2	92.3
SMFL (2 , 0.4)	**86.9**	92.5
SMFL (2 , 0.4)	86.5	**92.6**

Affinity Field	CS	CV
Local (Baseline3)	87.7	93.6
tSLnL (M1=1, M2=5)	88.1	93.9
sSLnL (M1=1, M2=5)	88.0	94.1
SLnL (M1=1, M2=5)	88.3	94.3
SLnL (M1=2, M2=4)	88.6	94.6
SLnL (M1=3, M2=3)	88.8	**94.9**
SLnL (M1=4, M2=2	88.9	94.8
SLnL (M1=5, M2=1)	**89.1**	94.7
SLnL (M1=6, M2=0)	88.8	94.7

Attention Types	CS	CV
No FA (Baseline2)	86.9	92.6
Amplitude FA	84.7	89.8
Shared FA	87.3	92.9
Dependent FA	87.5	93.2
Residual FA (rFA)	**87.7**	**93.6**

- *Tips*

FA: frequency attention;
Local: local CNN block;
M1: SLnL block number;
M2: local block number.

8. References

[1] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2017.

[2] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.