Actionlet-Dependent Contrastive Learning for Unsupervised Skeleton-Based Action Recognition

Wangxuan Institute of Computer Technology, Peking University

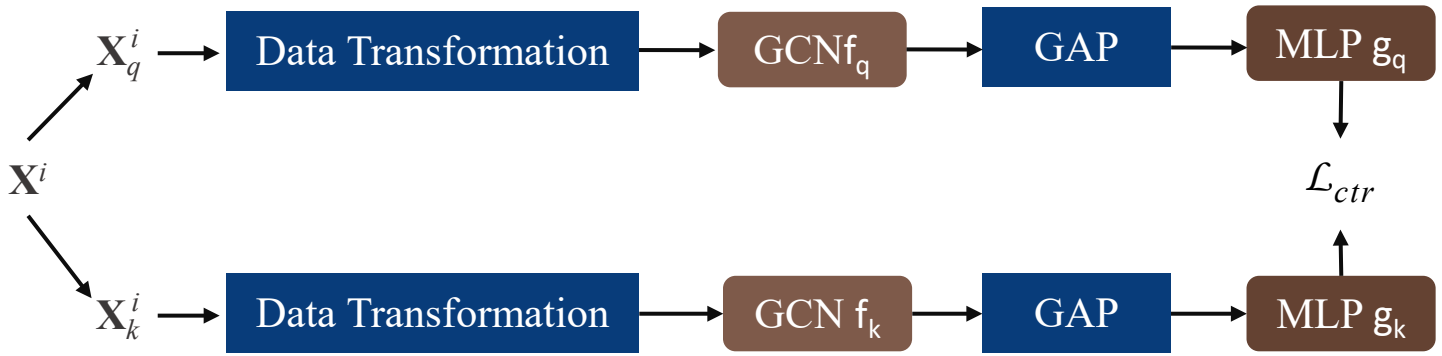
CVPR 2023 Highlight

Poster ID: TUE-AM-226

Lilang Lin Jiahang Zhang Jiaying Liu

Challenges:

- Uniform data transformation → degrade the motion information
- Global average pooling → make feature space indistinguishable

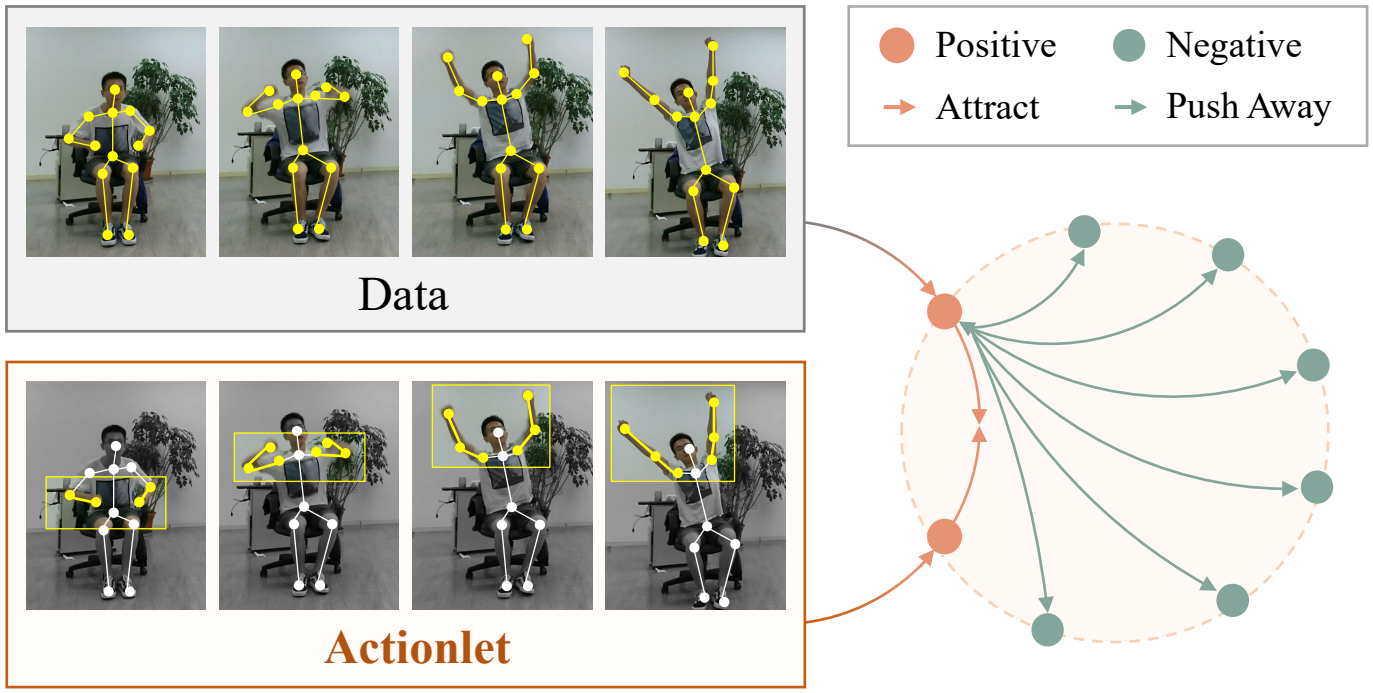


Challenges:

- Uniform data transformation → degrade the motion information
- Global average pooling → make feature space indistinguishable

Solution:

- Decouple **motion** and **static regions** in the data sequences



Video in Big Data Era

- **Videos in Internet**
  - Over a billion users on YouTube
  - A billion hours of videos each day

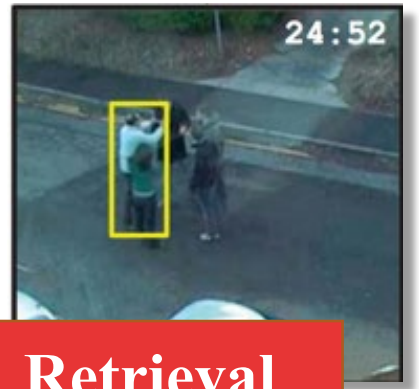
- **Surveillance Videos**
  - 176 million in China in 2017
  - Expected 626 million by 2020

Huge number of videos contains **Human Action**

→ **Video Action Analytics**

Various Applications

Surveillance



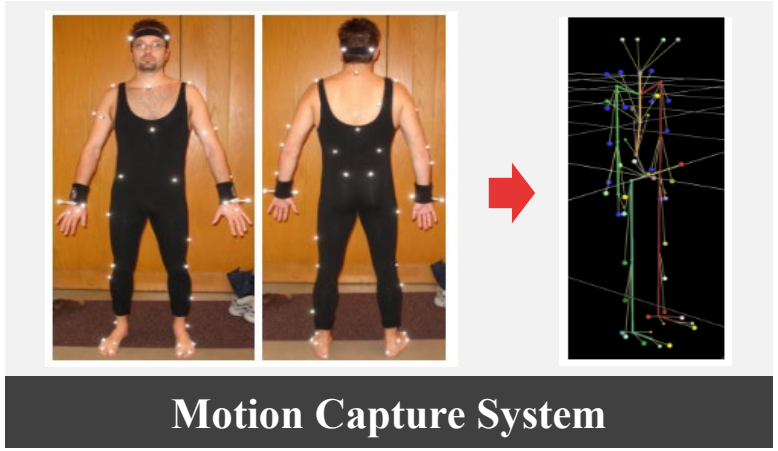
Retrieval

HCI

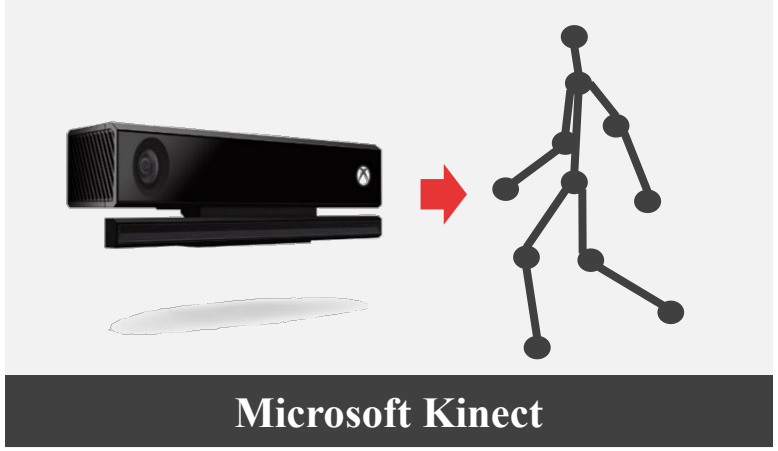
Home Care

Skeleton Data

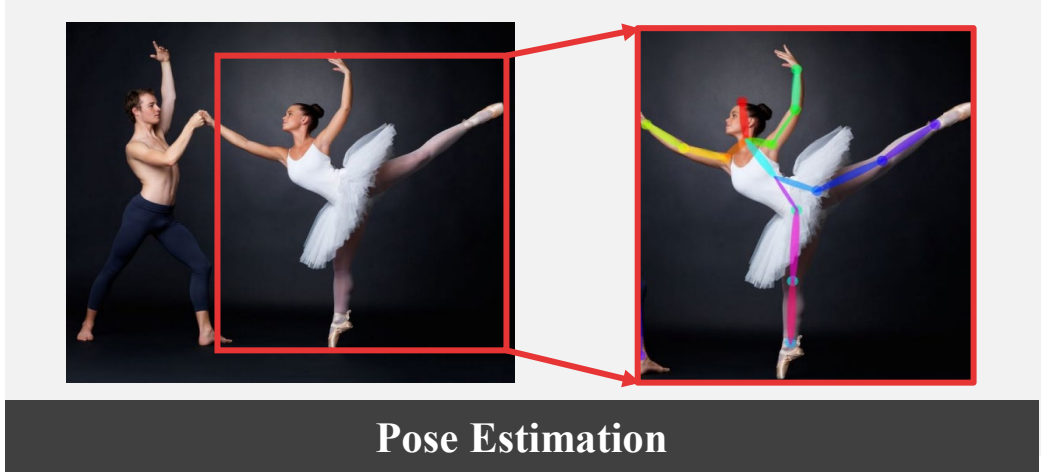
Data Access



Motion Capture System



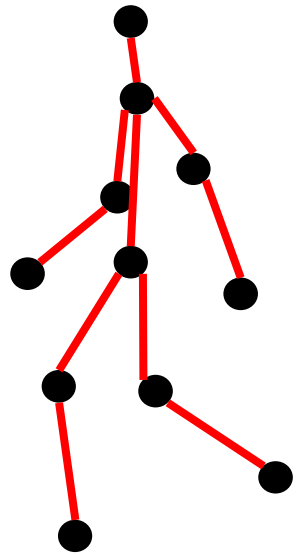
Microsoft Kinect



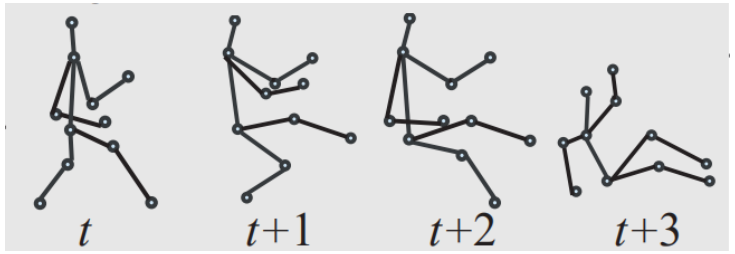
Pose Estimation

Skeleton Data

- **Pros**
  - High-level human representation
  - Robust to illumination and clustered background
  - Additional depth information
  - Real-time online performance
- **Cons**
  - Missing visual information
  - Not reliable due to noise and occlusion

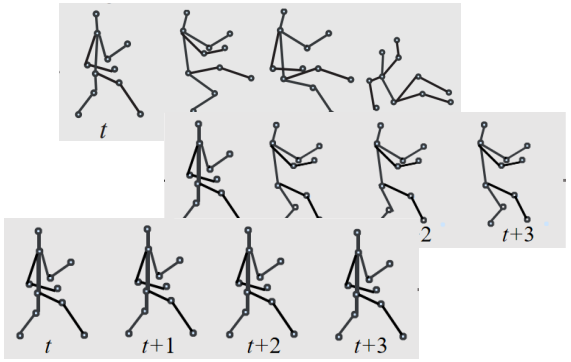


Skeleton-Based Action Recognition:



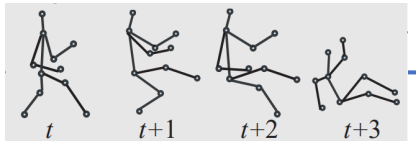
Action label:
Fall

Self-Supervised Learning:



No Label!

Pretext Tasks



Action label:
Fall

Self-Supervised Pretrain

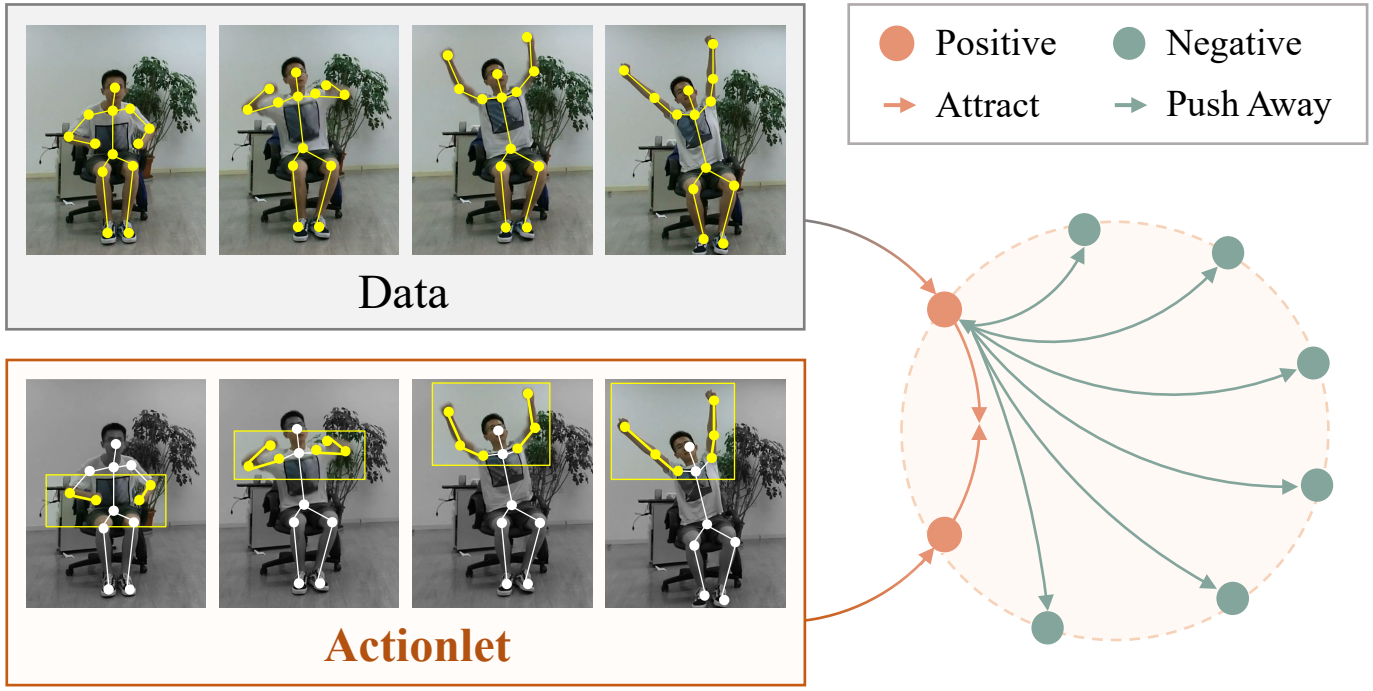
Supervised Finetune

Challenges:

- Uniform data transformation → degrade the motion information
- Global average pooling → make feature space indistinguishable

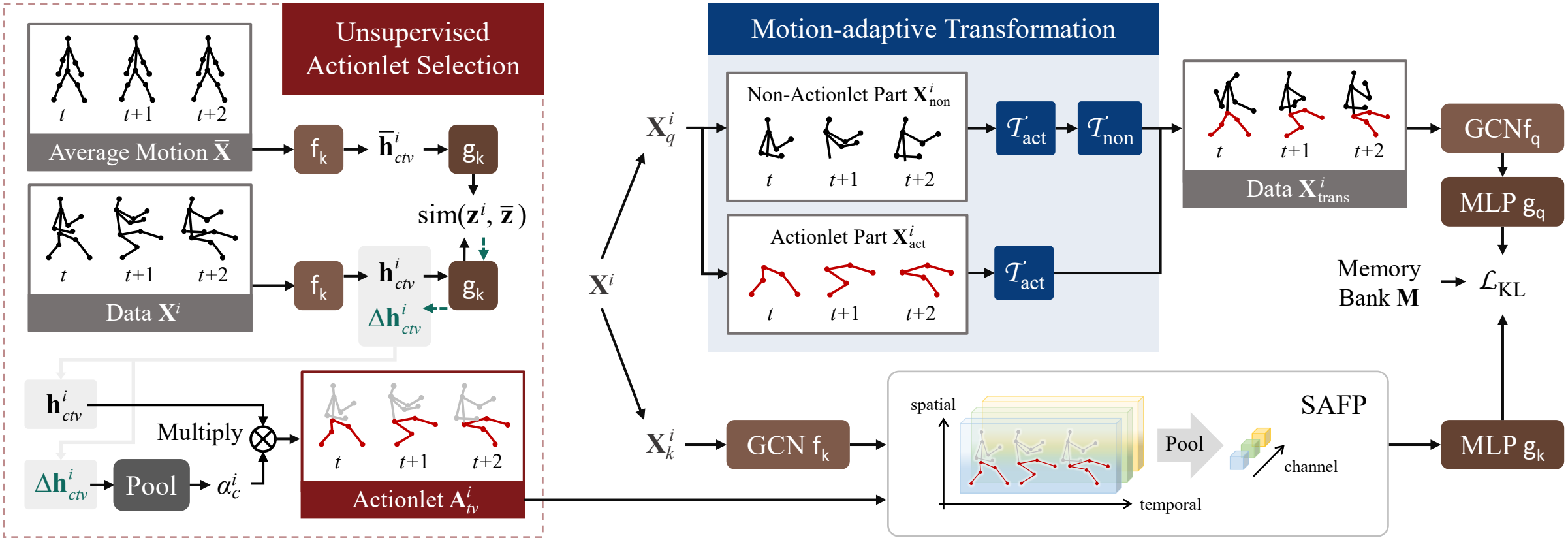
Solution:

- Decouple **motion** and **static regions** in the data sequences



Overall Network Architecture

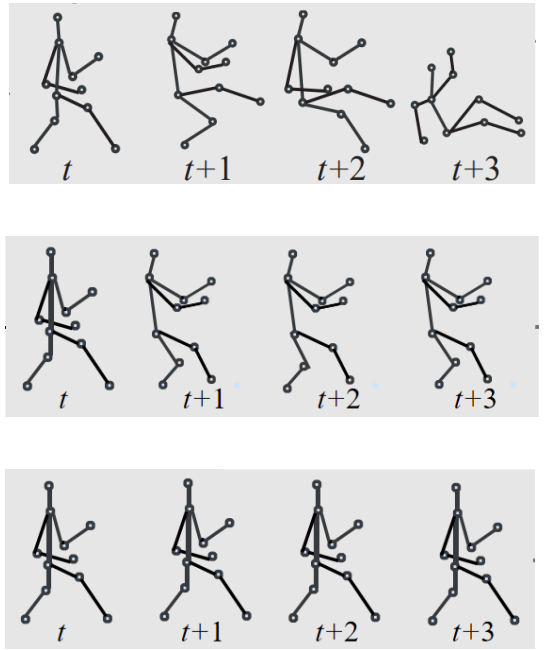
- Unsupervised Actionlet Selection
- Actionlet-Guided Contrastive Learning

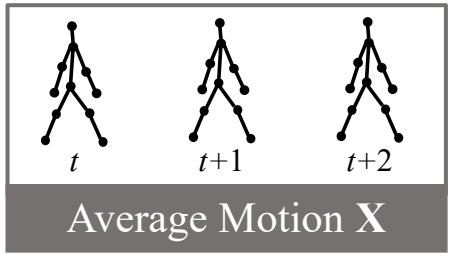


Overall Network Architecture

Unsupervised Actionlet Selection

Average Motion as Static Anchor



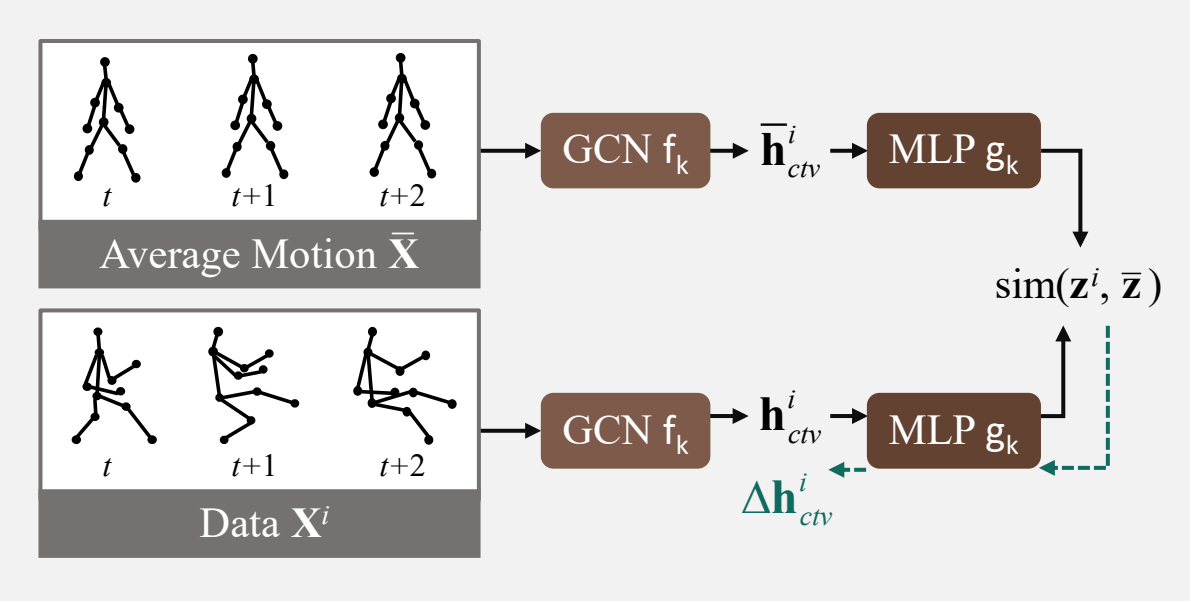


$$\bar{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}^i)$$

Overall Network Architecture

Unsupervised Actionlet Selection

Difference Activation Mapping for Actionlet Localization



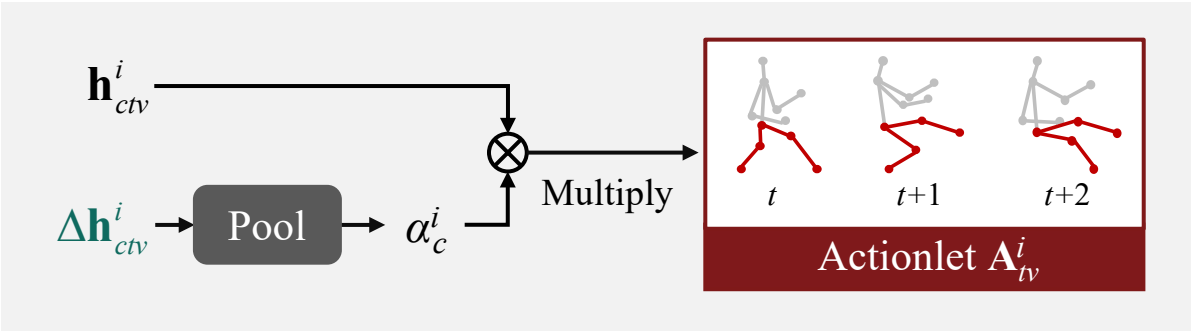
$$\Delta \mathbf{h}^i_{ctv} = \frac{\partial(-\mathrm{sim}(\mathbf{z}^i, \bar{\mathbf{z}}))}{\partial \mathbf{h}^i_{ctv}},$$

$$\alpha^i_c = \frac{1}{T \times V} \sum_{t=1}^{T} \sum_{v=1}^{V} \sigma(\Delta \mathbf{h}^i_{ctv}),$$

Overall Network Architecture

Unsupervised Actionlet Selection

Difference Activation Mapping for Actionlet Localization

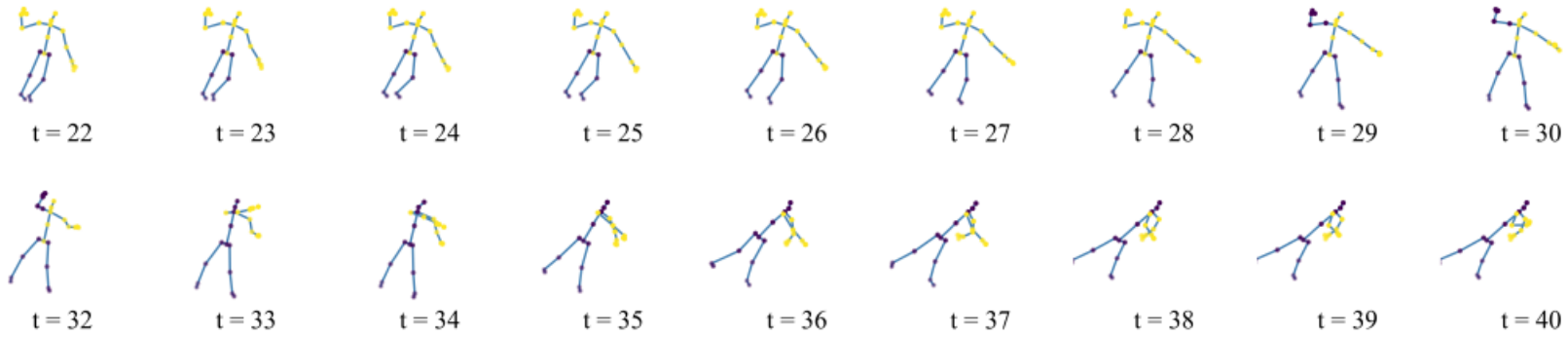


$$\mathbf{A}^i_{tv} = \sigma\left(\sum_{c=1}^{C} \alpha^i_c \mathbf{h}^i_{ctv}\right)\mathbf{G}_{vv}$$

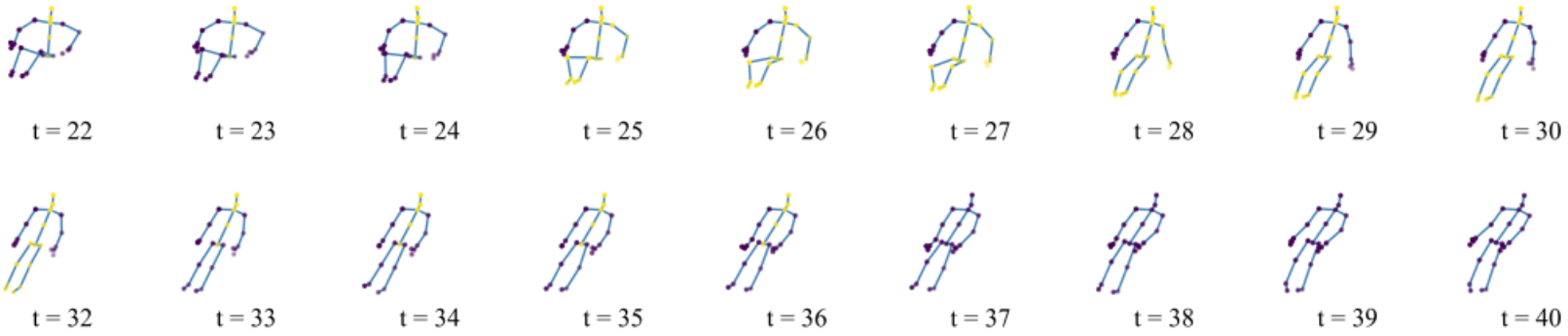
Overall Network Architecture

Unsupervised Actionlet Selection

Visualization of Actionlet



(a) Throw

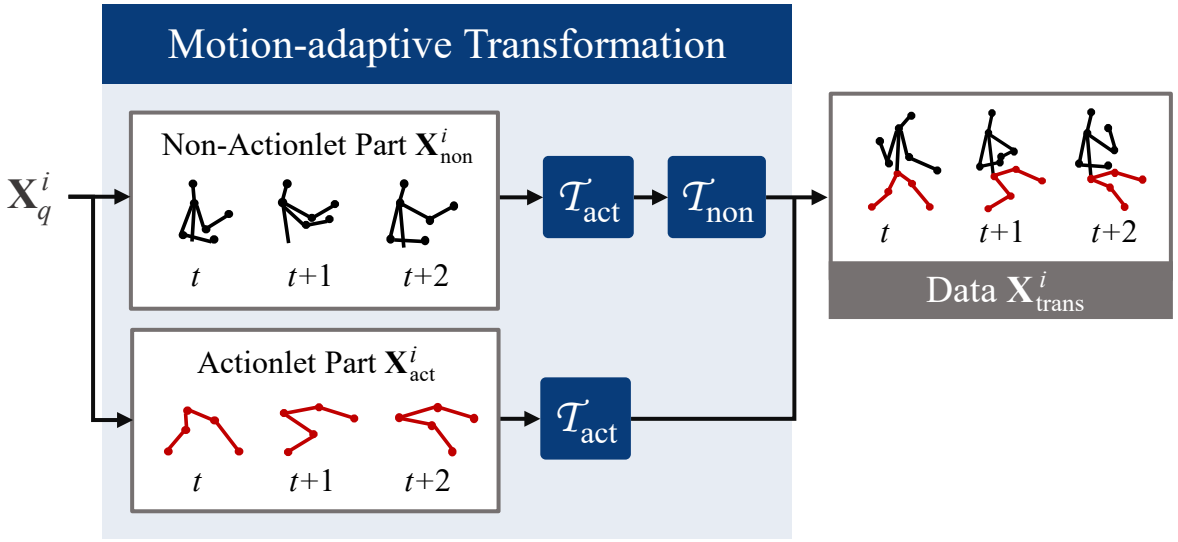


(b) Standup

Overall Network Architecture

Actionlet-Guided Contrastive Learning

Motion-Adaptive Transformation Strategy

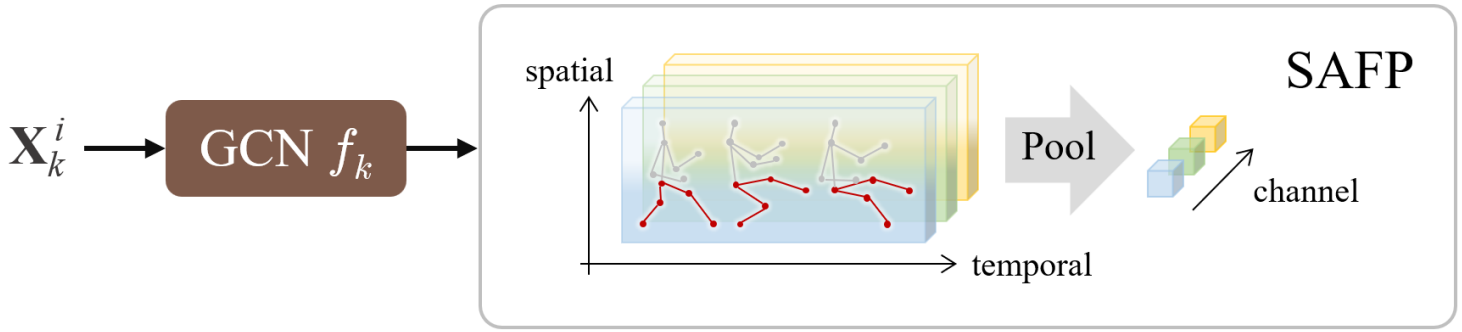


- Actionlet Transformation
  - *Shear, Spatial Flip, Rotate, Axis Mask*
  - *Crop, Temporal Flip*
  - *Gaussian Noise, Gaussian Blur*
  - *Skeleton AdaIN*
- Non-Actionlet Transformation
  - *Random Noise, Skeleton Mix*

Overall Network Architecture

Actionlet-Guided Contrastive Learning

Semantic-Aware Feature Pooling

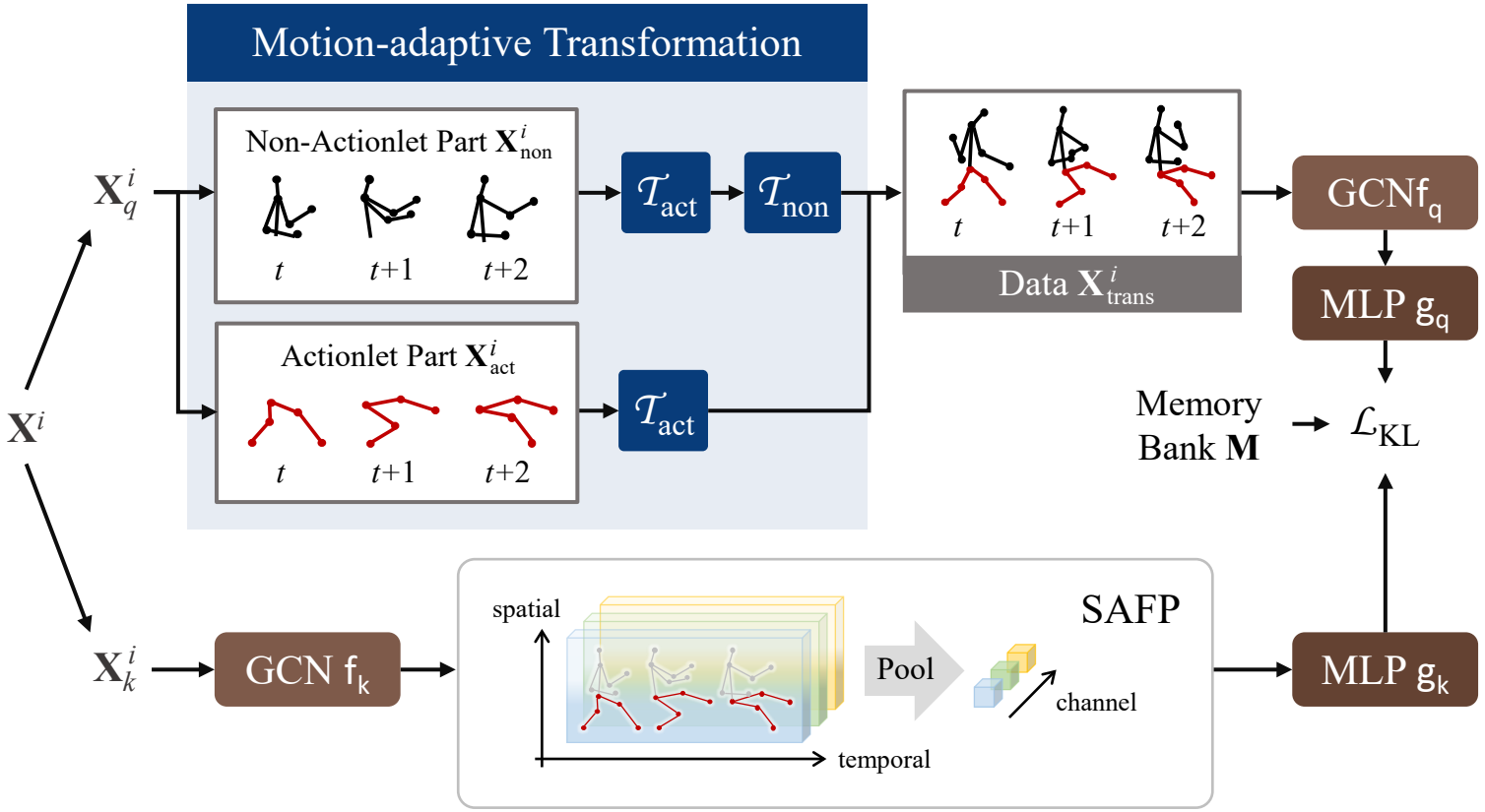


$$\mathrm{SAFP}(\mathbf{h}_{ctv}^i) = \sum_{t=1}^{T}\sum_{v=1}^{V} \mathbf{h}_{ctv}^i \left(\frac{\mathbf{A}_{tv}^i}{\sum_{t=1}^{T}\sum_{v=1}^{V} \mathbf{A}_{tv}^i} \right)$$

Overall Network Architecture

Actionlet-Guided Contrastive Learning

Training Overview



Experiment Configurations

- Unsupervised approaches

  - Train the classifier with pretrained encoder fixed.

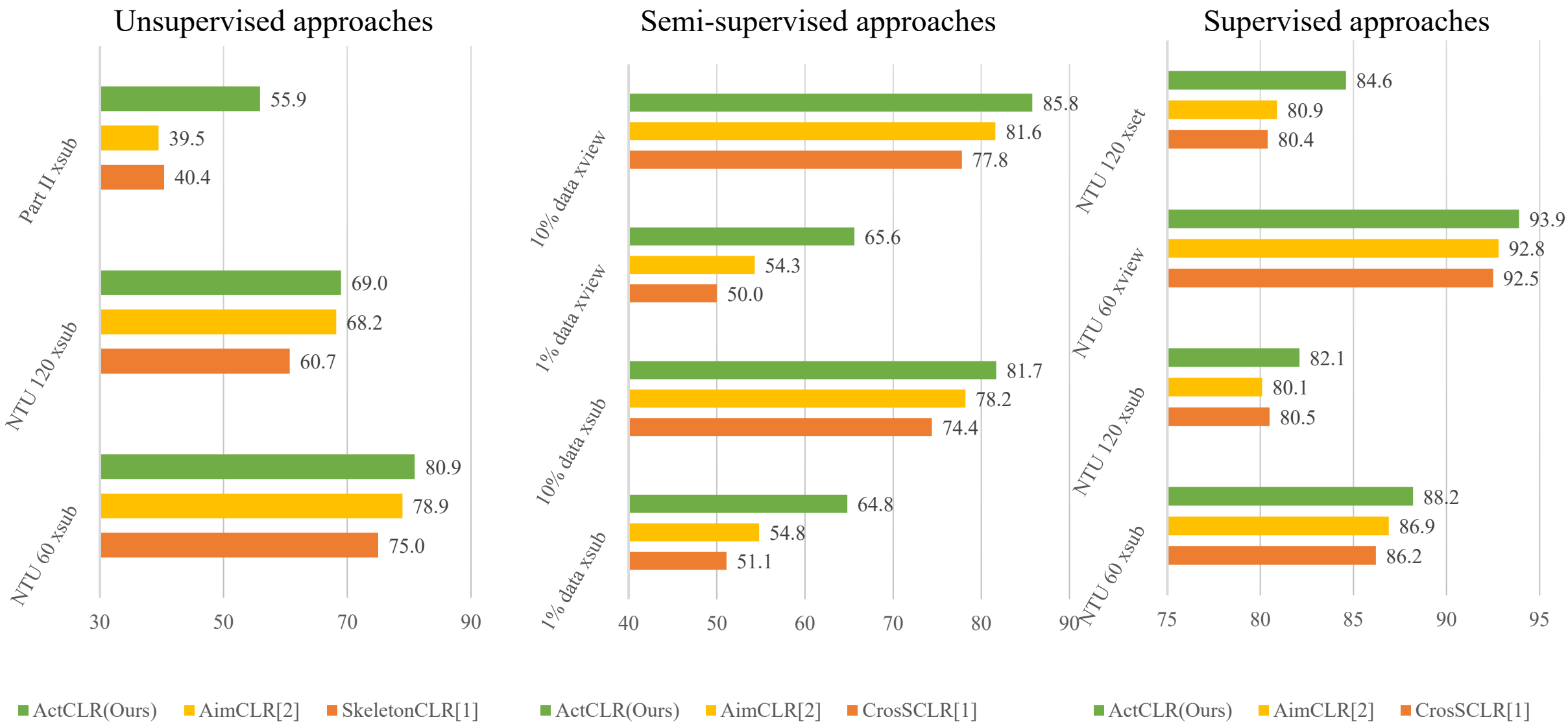
- Semi-supervised approaches

  - Jointly train the classifier and encoder with partial labeled data.

- Supervised approaches

  - Jointly train the classifier and encoder with full labeled data.

Experiment Results



[1] Li et al. 3D human action representation learning via cross-view consistency pursuit. CVPR 2021.
[2] Guo et al. Contrastive learning from extremely augmented skeleton sequences self-supervised action recognition. AAAI 2022.

- **Skeleton Based Action Recognition**
  - Unsupervised Actionlet Selection
  - Actionlet-Guided Contrastive Learning
- **Experimental Results**
  - Impressive results compared with other methods
  - Generalizable in different settings

Project

Lilang Lin (林里浪)
linlilang@pku.edu.cn

STRUCT: www.wict.pku.edu.cn/struct/