Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations

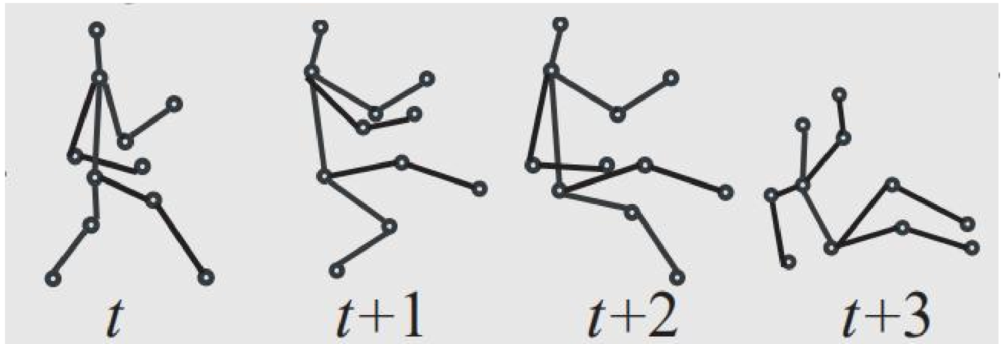
Wangxuan Institute of Computer Technology, Peking University

Jiahang Zhang Lilang Lin Jiaying Liu

2022.12.26

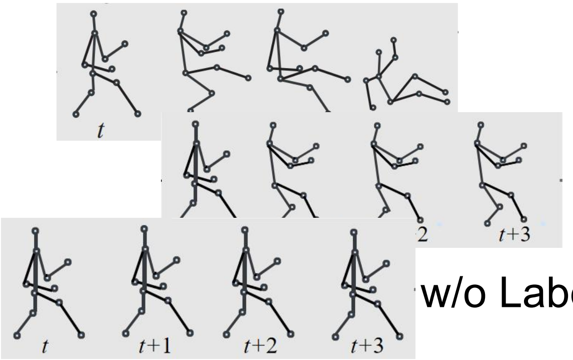
02 Aim and Challenge

Skeleton-Based Action Recognition:



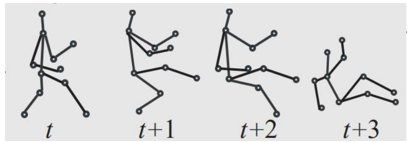
Action label:
Fall

Self-Supervised Learning:



w/o Label

Pretext Tasks

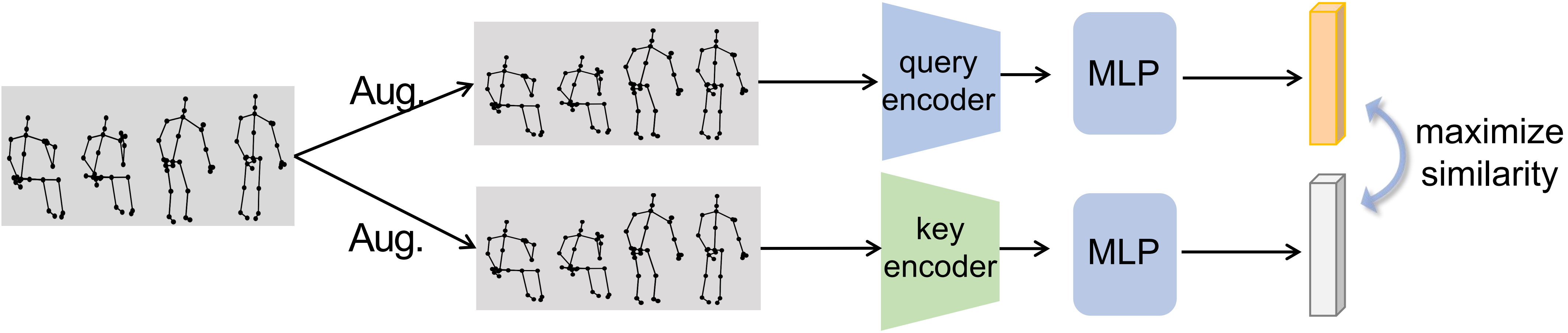


Fall

Self-Supervised Pretrain Stage

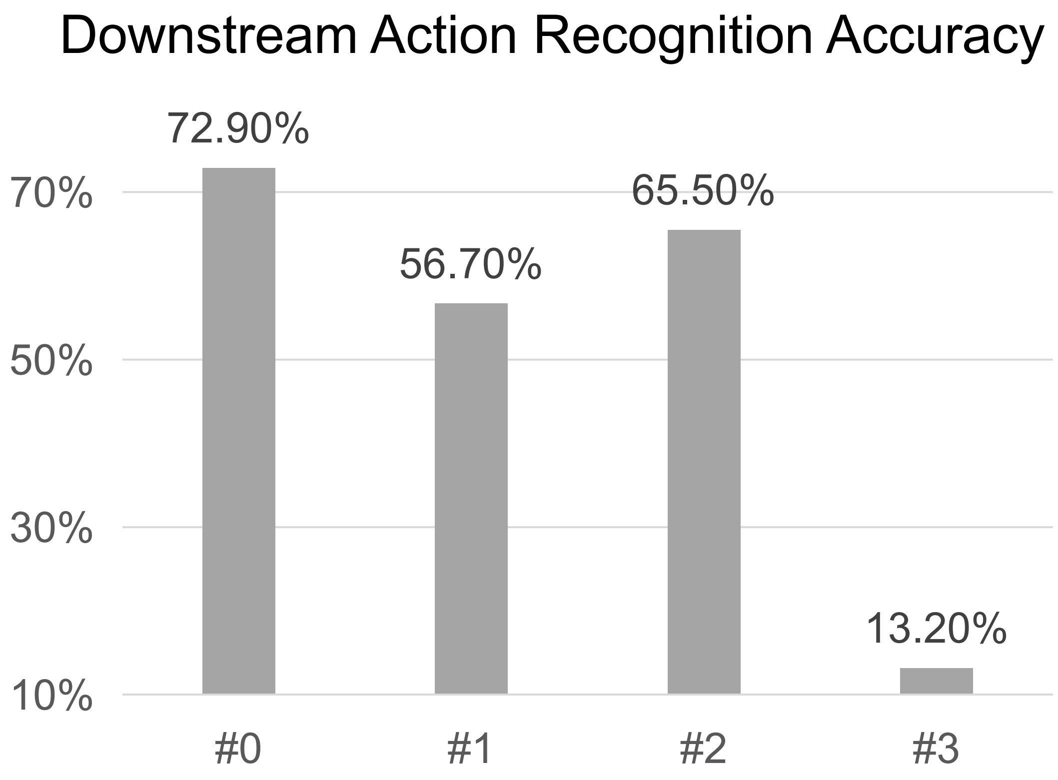
Supervised Finetune Stage

- **Contrastive Learning for Skeleton:**
  - Data augmentation module to generate positive pairs
  - Pull positive pairs
  - Push negative pairs



Challenges:

- Tranditional contrastive learning pipeline cannot benefit from the strong data augmentation.



#0： Baseline augmentation

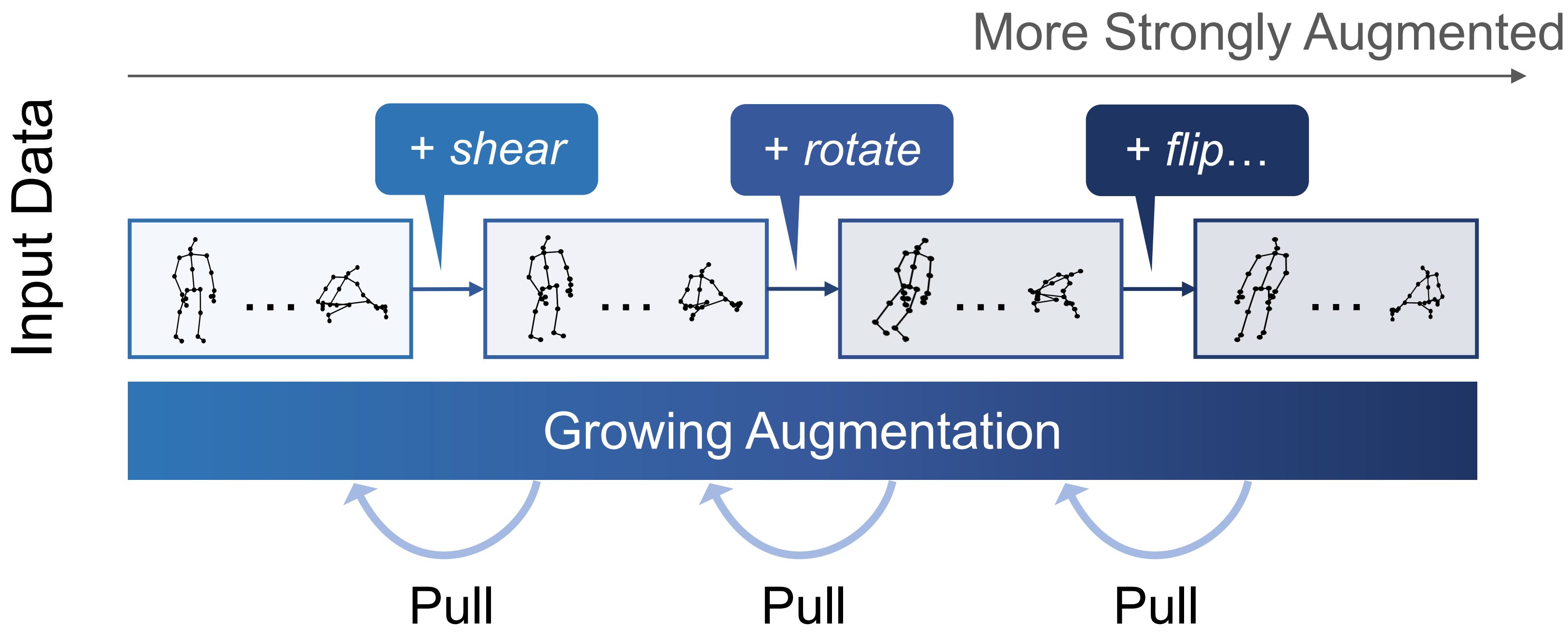
#1： *w/ Random Mask*

#2： *w/ Drop/Add Edges*

#3： *w/ Skeleton AdaIN*

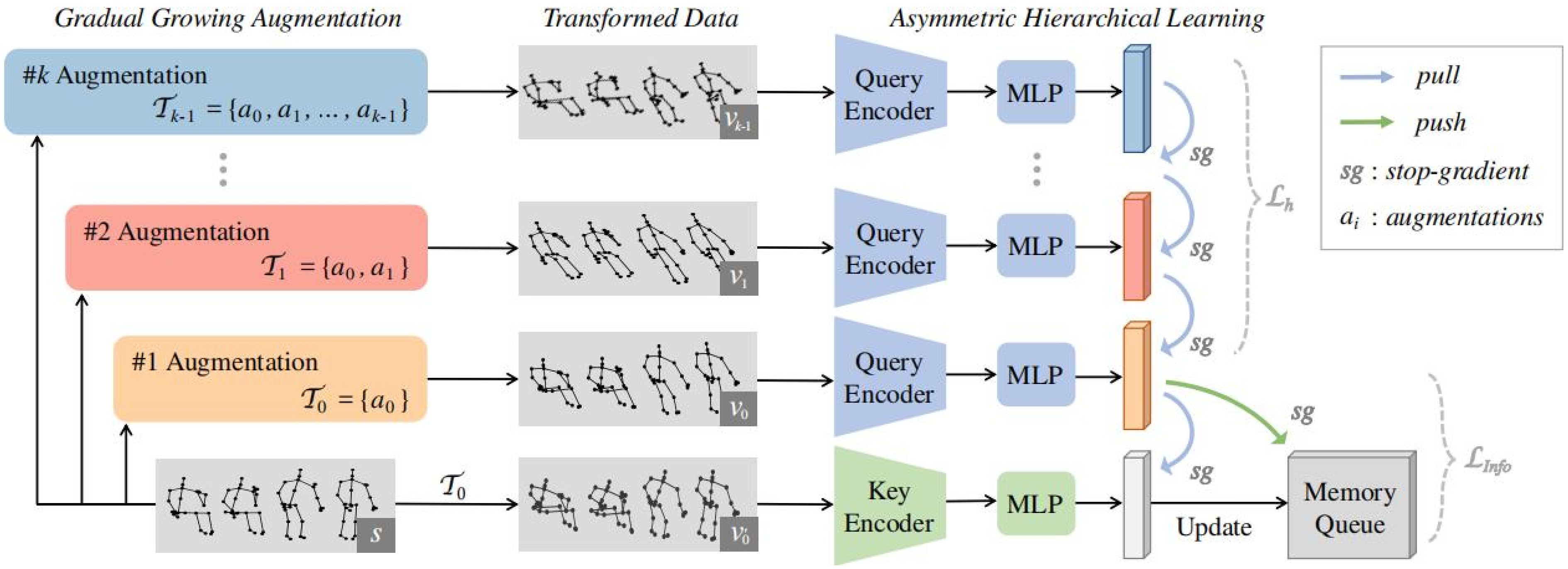
- **Challenges:**
  - Tranditional contrastive learning pipeline cannot benefit from the strong data augmentation.
  - Treating all augmentations equally cause sub-optimal representations.
- **Solution:**

- **Challenges:**
  - Tranditional contrastive learning pipeline cannot benefit from the strong data augmentation.
  - Treating all augmentations equally cause sub-optimal representations.
- **Solution:**



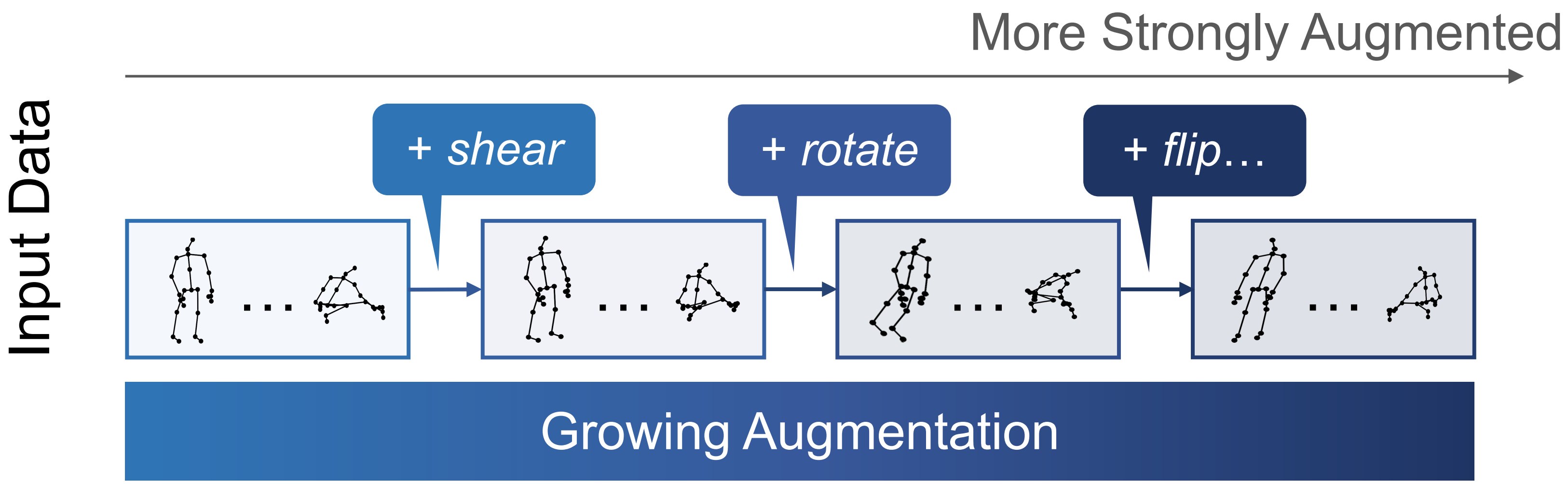
Our Framework Overview:

- Gradual Growing Augmentation
- Asymmetric Hierarchical Learning

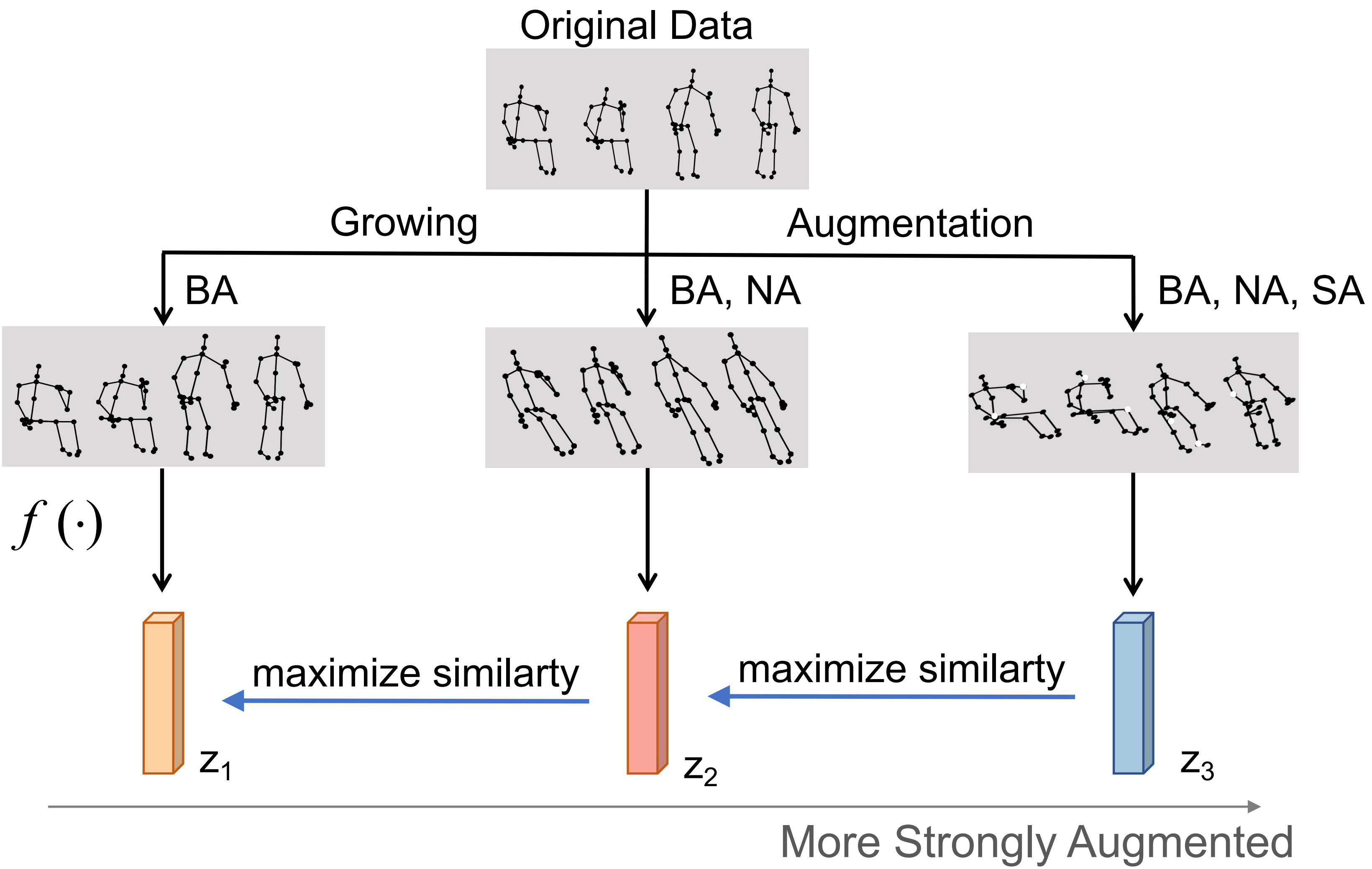


- **Gradual Growing Augmentation**
  - Divide the all augmentations into different sets.
    - Basic Augmentation Set (BA)
      - *Shear, Temporal Crop*
    - Normal Augmentation Set (NA)
      - *Flip, Rotate, Gaussion noise, ...*
    - Strong Augmentation Set (SA)
      - *Random Mask*
      - *Drop/Add Edges (DAE)*
      - *Skeleton AdaIN*

- **Gradual Growing Augmentation**
  - Divide the all augmentations into different sets.
  - Generate multiple positive pairs by applying these augmentation sets progressively.



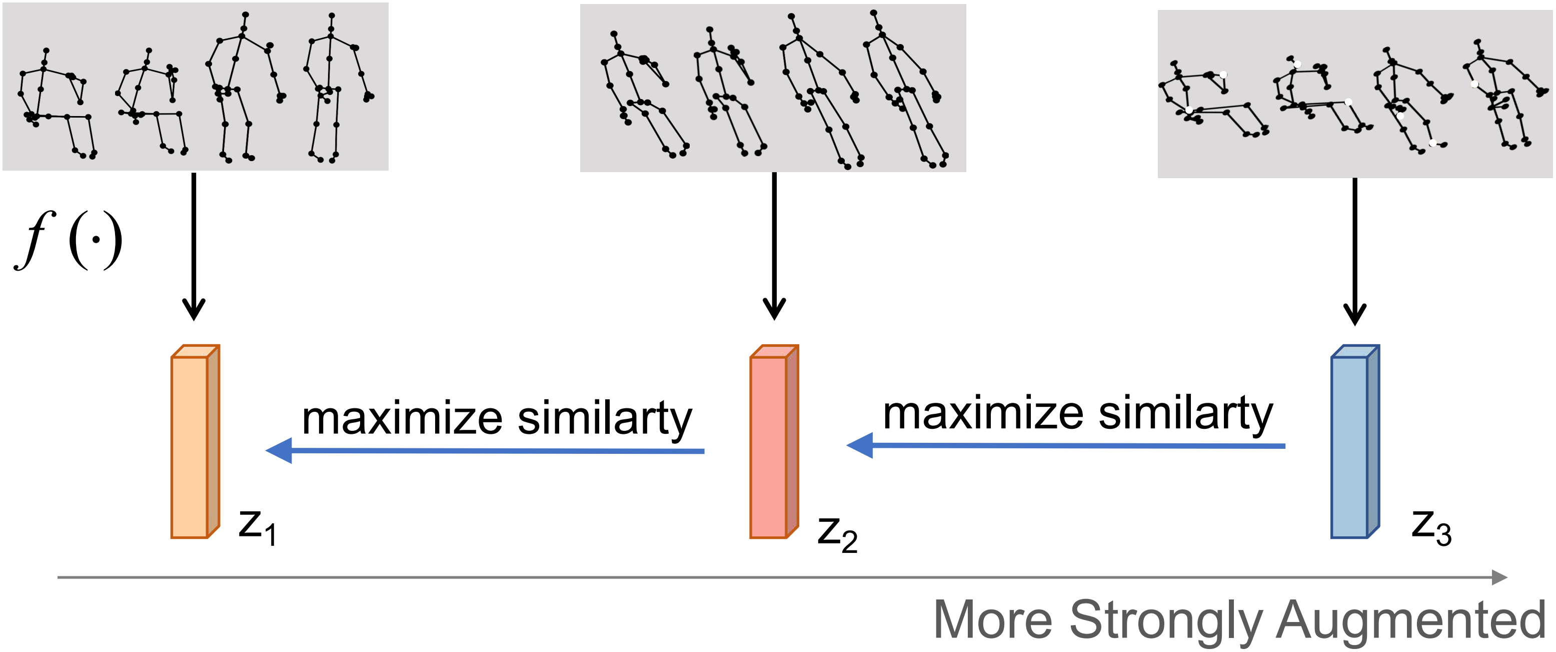
Asymmetric Hierarchical Learning



Asymmetric Hierarchical Learning

- Hierarchical self-supervised loss

$$\mathcal{L}_h = \sum_{i=1}^{k-1} sim\left(z_i, \mathrm{stopgrad}\left(z_{i-1}\right)\right)$$



Asymmetric Hierarchical Learning

- Hierarchical self-supervised loss

$$\mathcal{L}_h = \sum_{i=1}^{k-1} sim\left(z_i, \text{stopgrad}\left(z_{i-1}\right)\right)$$

- KL divergence as sim(·) function

$$D_{KL}\left(\text{stopgrad}(p\left(z|z_{i-1}\right)), p\left(z|z_i\right)\right)$$

$$p\left(z|z_i\right) = \frac{\exp(z \cdot z_i/\tau)}{\exp(z_0' \cdot z_i/\tau) + \sum_{i=1}^{M} \exp(m_i \cdot z_i/\tau)}$$

Full Model

- Optimization Objective
  - InfoNCE loss between the basic positive pair

$$\mathcal{L}_{Info} = -\log \frac{\exp(z \cdot z'/\tau)}{\exp(z \cdot z'/\tau) + \sum_{i=1}^{M} \exp(z \cdot m_i/\tau)}$$

  - The proposed hierarchical self-supervised loss

$$\mathcal{L}_h = \sum_{i=1}^{k-1} sim\left(z_i, \mathrm{stopgrad}\left(z_{i-1}\right)\right)$$

  - Overall loss

$$\mathcal{L} = \mathcal{L}_{Info} + \lambda_h \mathcal{L}_h$$

Full Model

Optimization Objective

- InfoNCE loss between the basic positive pair

$$\mathcal{L}_{Info} = -\log \frac{\exp(z \cdot z'/\tau)}{\exp(z \cdot z'/\tau) + \sum_{i=1}^{M} \exp(z \cdot m_i/\tau)}$$

- The proposed hierarchical self-supervised loss

Training process

Self-supervised pretrain for the encoder

$$\mathcal{L} = \mathcal{L}_{Info} + \lambda_h \mathcal{L}_h$$

Supervised finetune for the classifier $\mathcal{L}_{cls}$

Experiment Settings

- Unsupervised approaches

  - Train the classifier with pretrained encoder fixed.

- Semi-supervised approaches

  - Jointly train classifier and encoder with partial labeled data.

- Supervised approaches

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

Datasets

- NTU RGB+D 60 Dataset (NTU 60)[1]
- NTU RGB+D 120 Dataset (NTU 120)[2]
- PKU Multi-Modality Dataset (PKUMMD)[3]
  - PKUMMD part I (Part I)
  - PKUMMD part II (Part II)

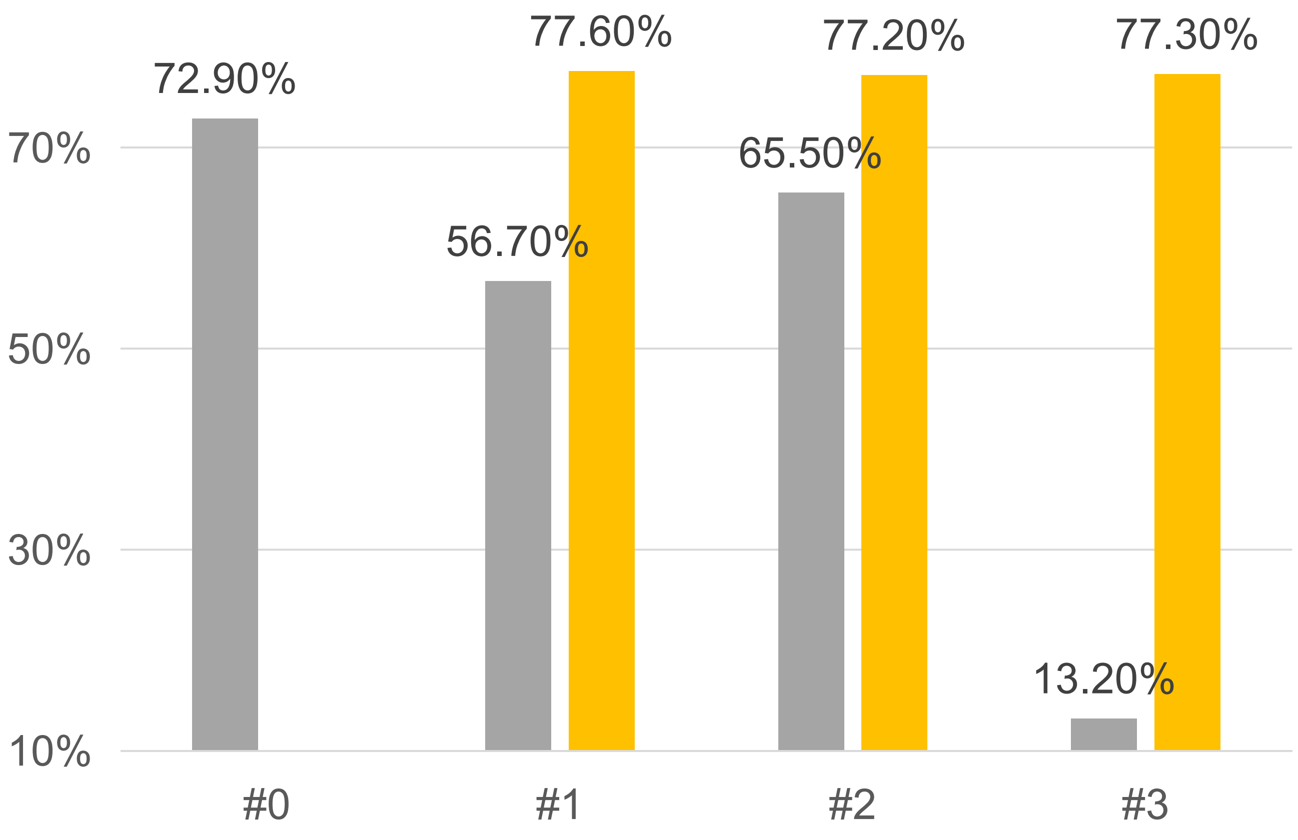
[1] Shahroudy et al. NTU RGB+ D: A large scale dataset for 3D human activity analysis. CVPR 2016.

[2] Liu et al. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. TPAMI 2019.

[3] Liu et al. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. Proc. of the Workshop on Visual Analysis in Smart and Connected Communities 2017.

Results on Strong Data Augmentations

Unsupervised Action Recognition Accuracy on NTU 60



#0： Baseline augmentation

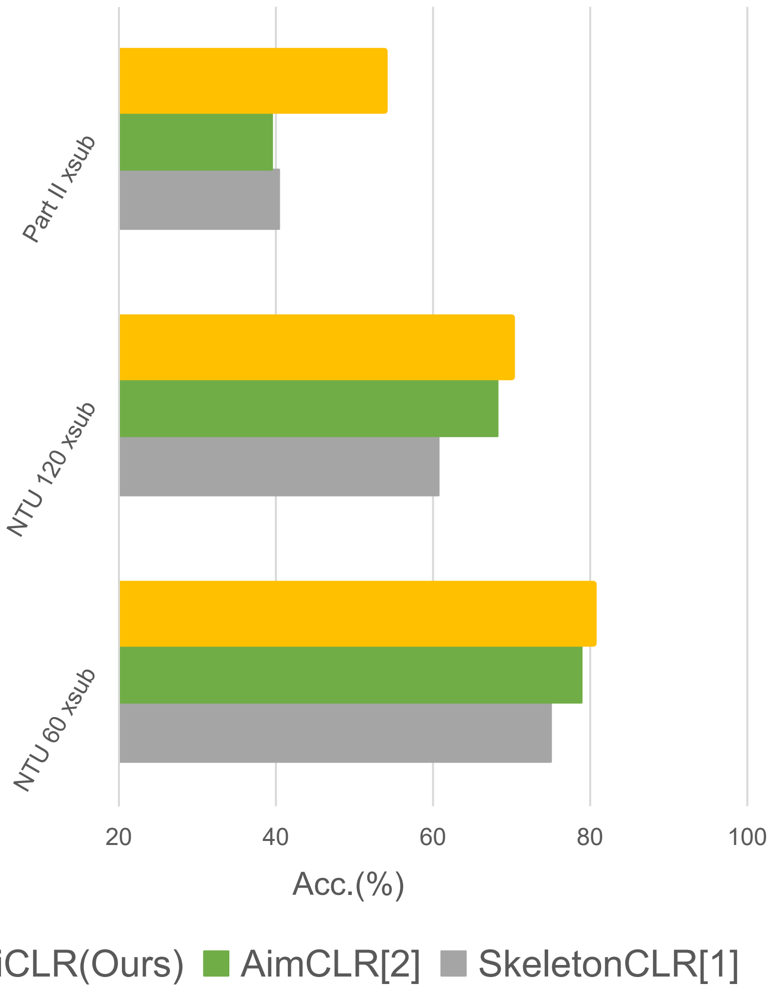
#1： *w/ Random Mask*

#2： *w/ Drop/Add Edges*

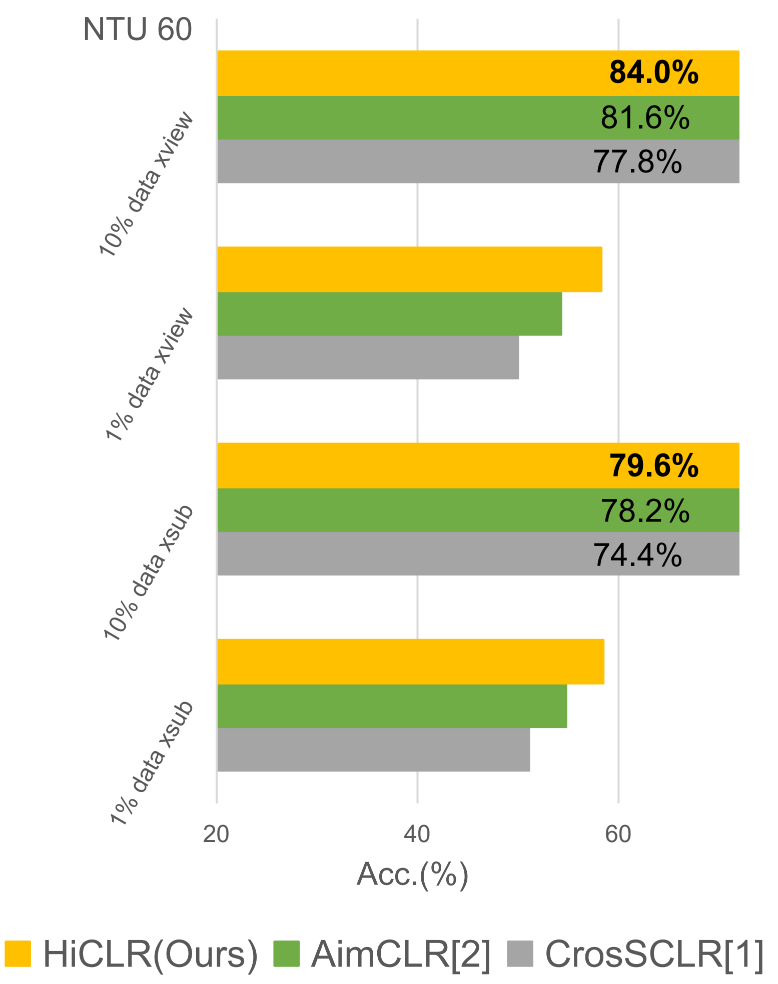
#3： *w/ Skeleton AdaIN*

Experiment Results

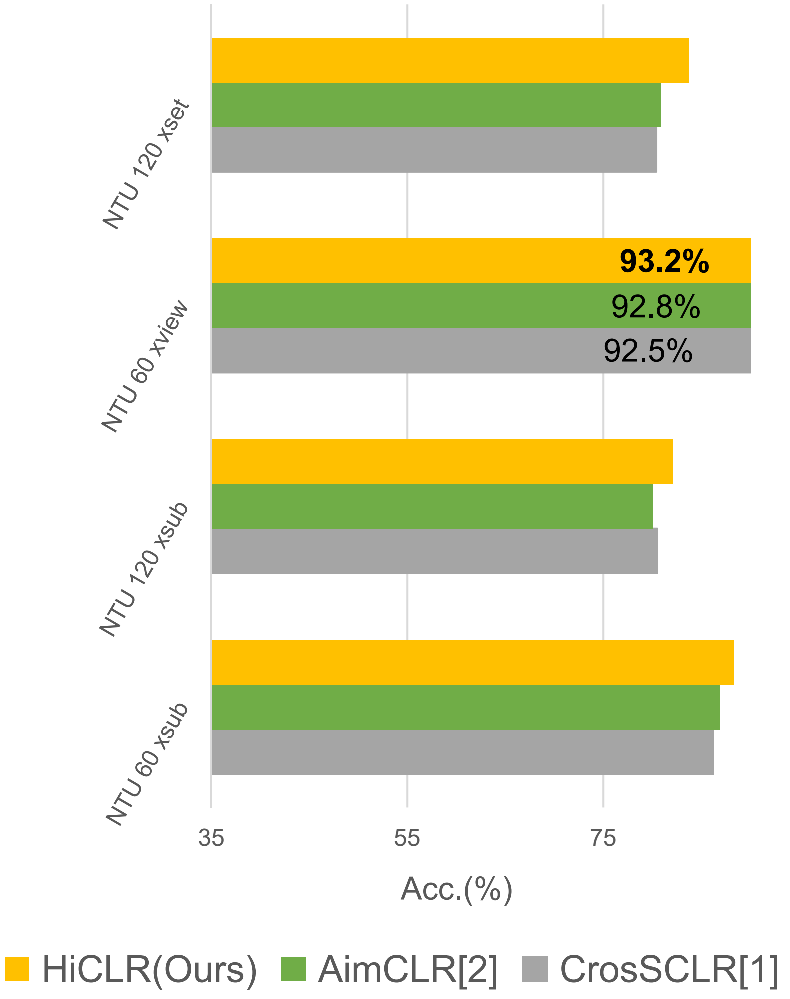
Unsupervised Approaches



Semi-supervised Approaches



Supervised Approaches



[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.

Results on Augmentation Arrangement

Augmentation Arrangement	Acc. (%)
k=1, BA	68.3
k=2, BA,NA	76.8
k=3, BA,NA,Mask	**77.6**
k=3, BA,NA,AdaIN	77.3
k=3, BA,NA,Drop/Add Edges	77.2
k=4, BA,NA,Drop/Add Edges,Mask	77.4
k=4, BA,NA,Drop/Add Edges,AdaIN	77.5

BA: Basic Aug. Set

NA: Normal Aug. Set

SA: Strong Aug. Set

k: branch number

Conclusion

- **Skeleton-Based Action Recognition**
  - Gradual Growing Augmentation
  - Asymmetric Hierarchical Learning
- **Experimental Results**
  - Impressive results compared with other methods
  - Generalizable in different settings

Jiahang Zhang (张佳航)
zjh2020@pku.edu.cn

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

Project