

论 文 汇 报

Language Knowledge-Assisted Representation Learning for Skeleton-Based Action Recognition

(基于骨架的动作识别中的语言知识辅助表示学习)

► 汇报时间：2024.04.10

► 汇报人：梁琨

目 录

CONTENT

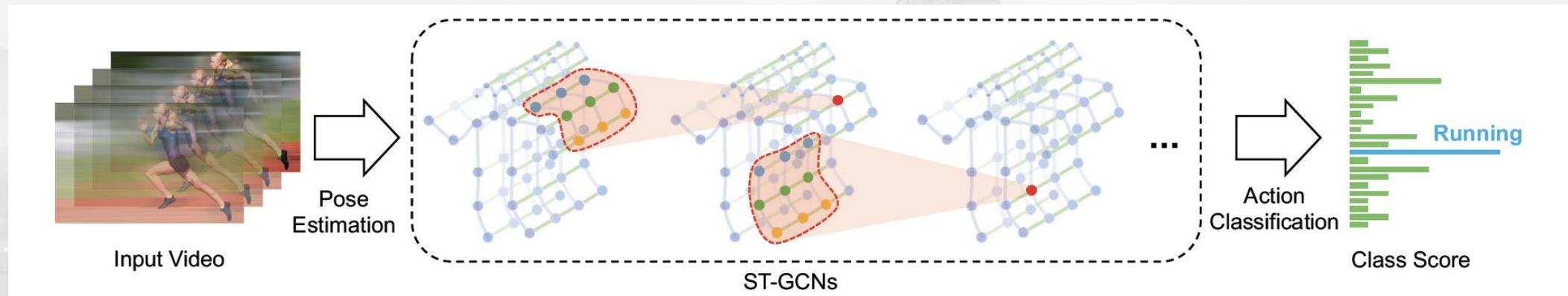
壹	<u>研究背景和意义</u>	03
貳	<u>论文创新点及原理</u>	06
參	<u>模型结构</u>	17
肆	<u>实验设计及结果</u>	32
伍	<u>论文总结</u>	37

1 *Part One*

研究背景和意义

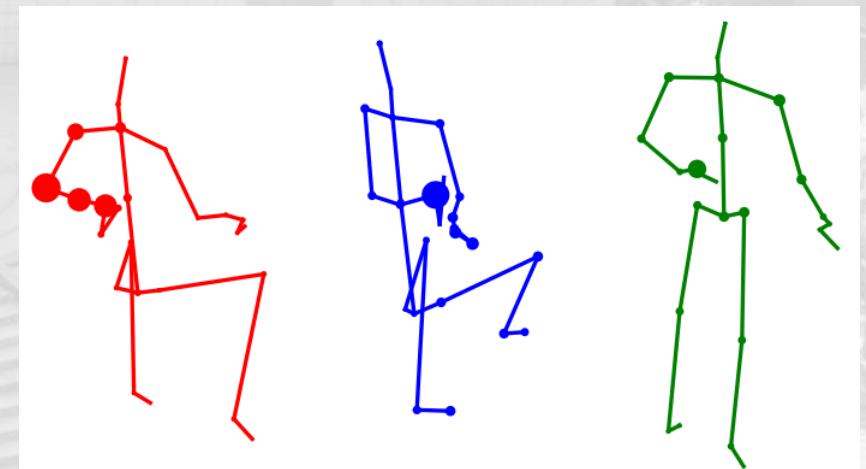
简要介绍基于骨骼的动作识别领域的现有的研究成果。
同时，介绍本文模型的baseline，以及实验过程中使用的实验方法和数据准备。

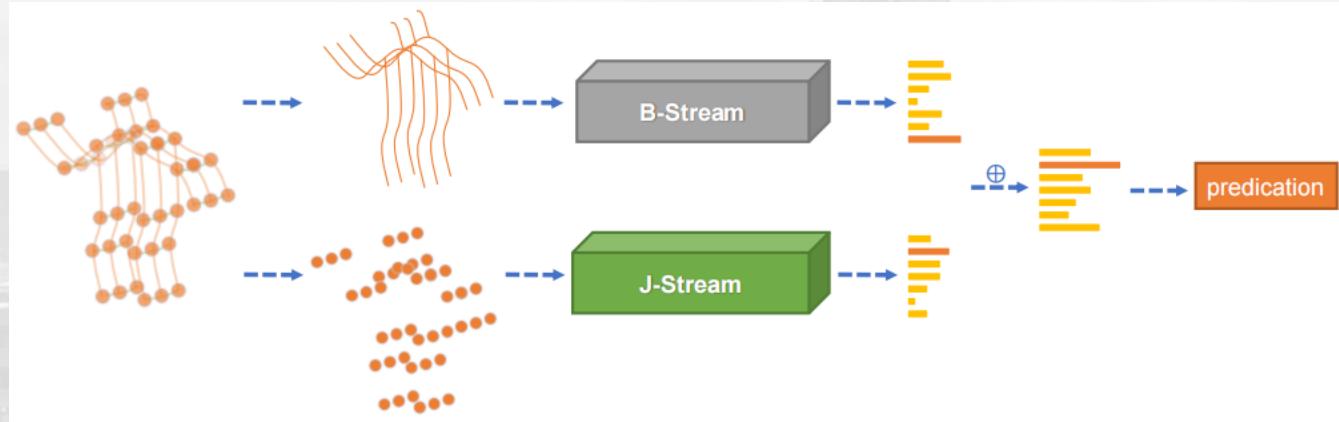
基于骨骼的动作识别的研究现状和背景



Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition (ST-GCN)[1]

人体动作识别是视频理解的核心任务之一，它是一种从第三人称视角观察和推断代理行为的分类任务。该任务可广泛应用于人机交互、视频监控、医疗保健和短娱乐视频。其中，基于骨架的动作识别由于其对视频中各种环境噪声的鲁棒性和便于模型聚焦的高紧致性而得到广泛应用。





Two-stream adaptive graph convolutional networks
for skeleton-based action recognition[2]

- ✓ 骨骼数据的**二阶信息**（骨骼的长度和方向）对于动作识别更有信息性，在这篇论文中，提出了一种新的双流自适应图卷积网络（2s-AGCN）用于基于骨架的动作识别。
- ✓ 模型中图的拓扑可以通过BP算法以端到端方式统一或单独学习。这种**数据驱动**的方法增加了模型构建图的灵活性，并为适应各种数据样本带来了更多的通用性。

2 Part Two

论文创新点及原理

介绍本篇论文的主要创新点，结合LLM的先验拓扑关系的生成，基于LLM的先验知识生成的一种新的骨架模态表示，PC-AC，K-Loop

创新点

Aid topological modeling using LLM

Multi-hop Attention

A new skeleton representations method

An auxiliary supervised module PC-AC with class information encoding

Class Prior Relation Graph(CPR)

Global Prior Relation Graph(GPR)

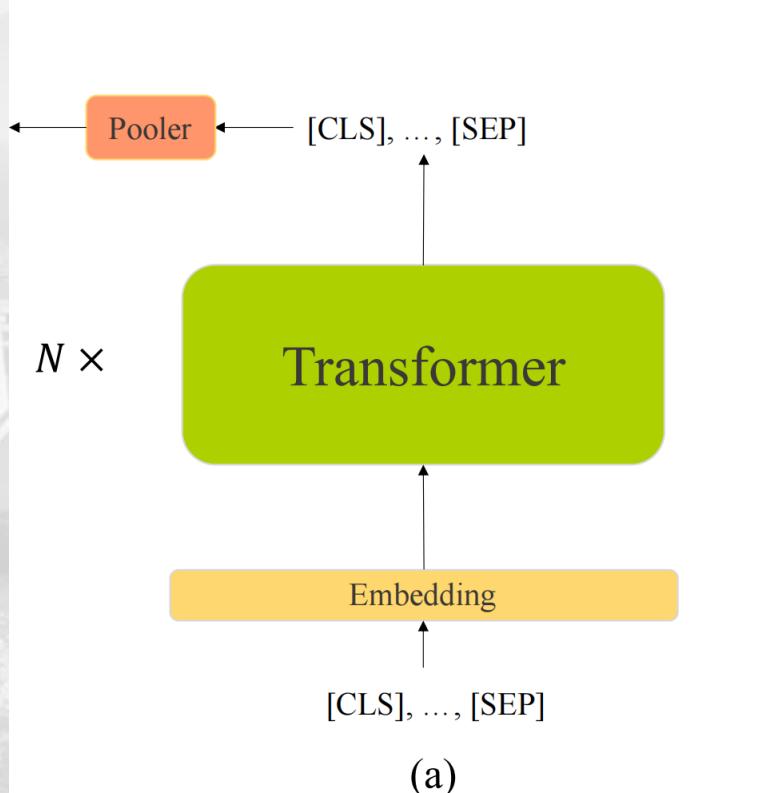
GPR

CPR

A Global Prior Relation Graph Generated by LLM

GPR+CPR

XDU



使用BERT来提取每一类动作标签和所有关节的文本特征。在预训练过程中BERT的损失值包括填空和输入两句话来预测后者是否是前一句的句子。模型的最后一层的输出用于完成填空，Pooler部分的输出用于下一个句子的预测。由于Pooler之后的特征包含了整个句子的语义，因此它们可以在执行一般的文本分类等任务时直接使用。在本文中，也选择了Pooler后的特征作为我们的骨架节点的特征。

A Global Prior Relation Graph Generated by LLM

GPR+CPR

XDU

给定M个动作类别和N个人类的关节点，对每个节点提取特征。

输入：包含**动作类别 + 骨骼点名称**

例如：

p1: [J] function in [C].

p2: What happens to [J] when a person is [C]?

p3: What will [J] act like when [C]?

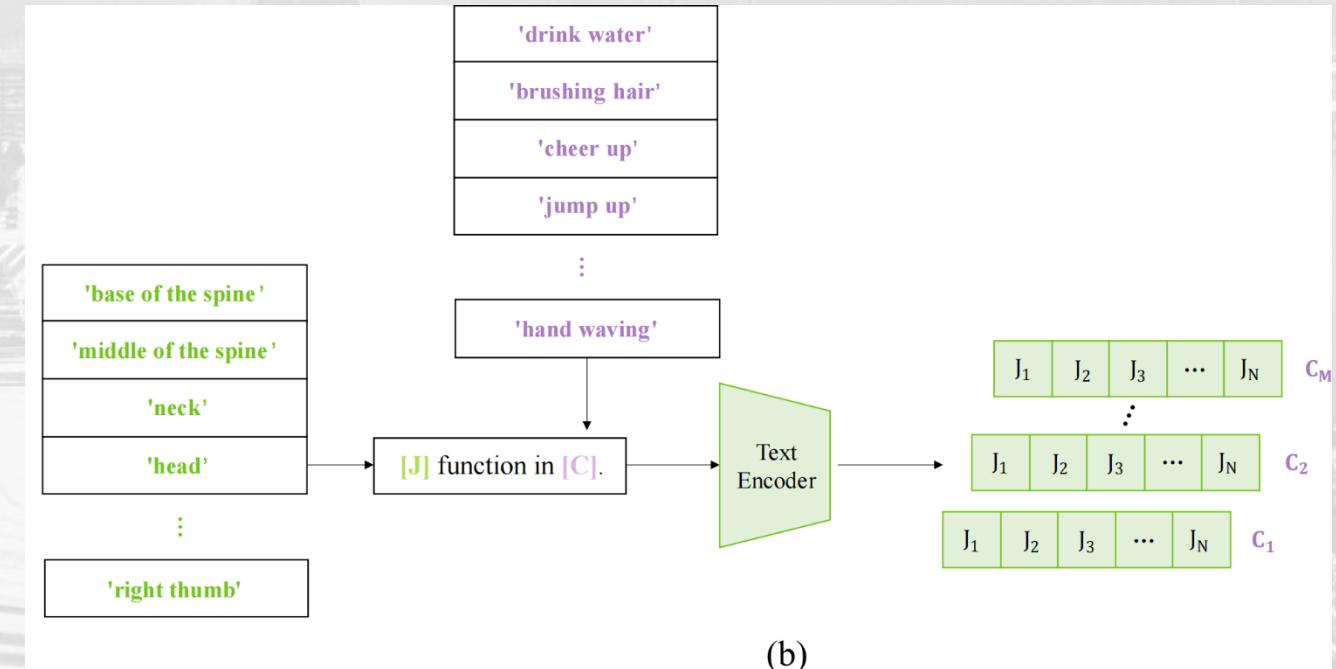
p4: When [C][J] of human body.

p5: When [C] what will [J] act like?

p6: When a person is [C], [J] is in motion.

对于动作 C_i ，LLM的文本编码器得到相应的输出特征

$$C_i \in \mathbb{R}^{N \times C} (i=1, 2, \dots, M)$$



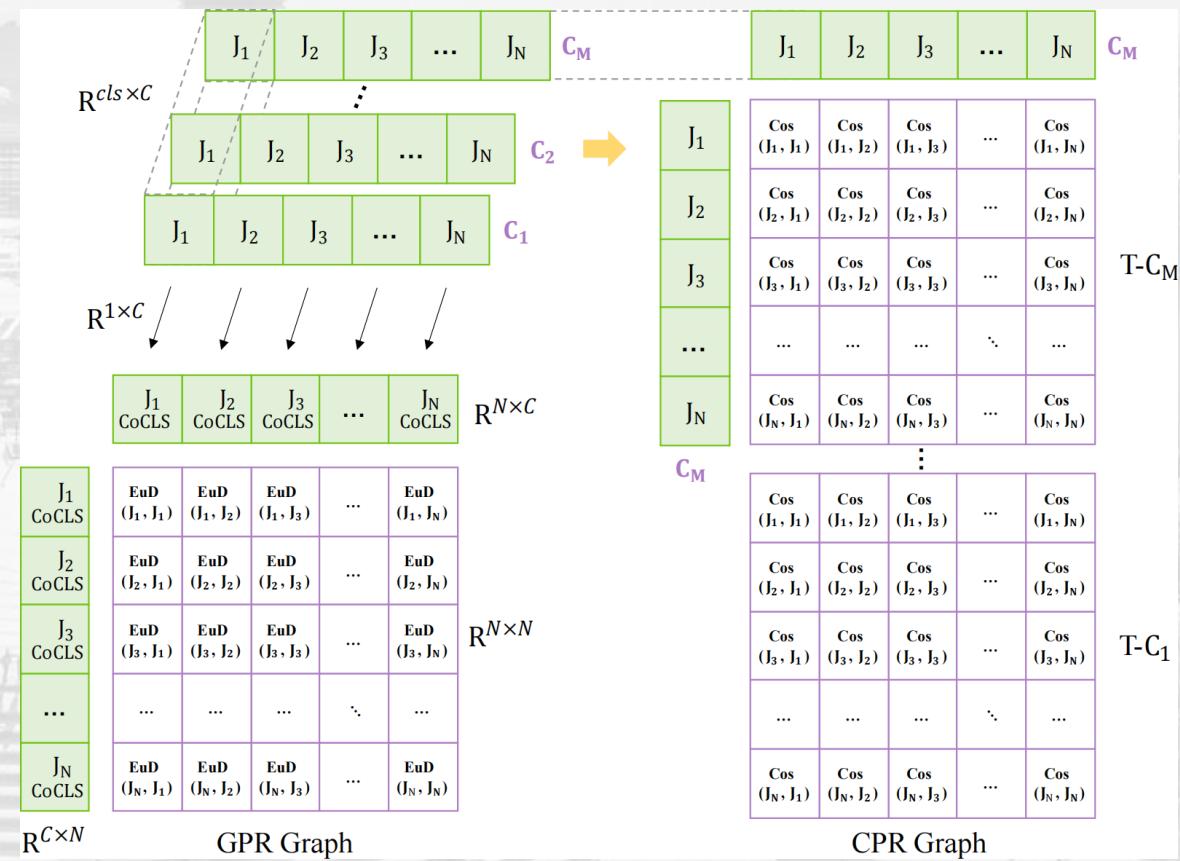
CPR的计算方法

XDU

A class relationship topology graph containing a priori node relationships

构造由文本特征生成的类别拓扑图。对于动作 $C_i (i = 1, 2, \dots, M)$ ，
设节点 $k (k = 1, 2, \dots, N)$ 的文本特征为 $J_k \in \mathbb{R}^{1 \times C}$
将两两文本特征逐个组合起来计算余弦相似度

对于每个动作 C_i ，都可以生成唯一一个 $T - C_i$ 拓扑图与之对应，该图主要包含了任意两个关节点关于动作 C_i 的相似和相关程度。
我们将 $T - C_i$ 称为一个类拓扑范例，并假设它们包含应该存在于动作识别操作中的节点关系



GPR的计算方法

XDU

A global inter-node relationship topology graph GPR-Graph

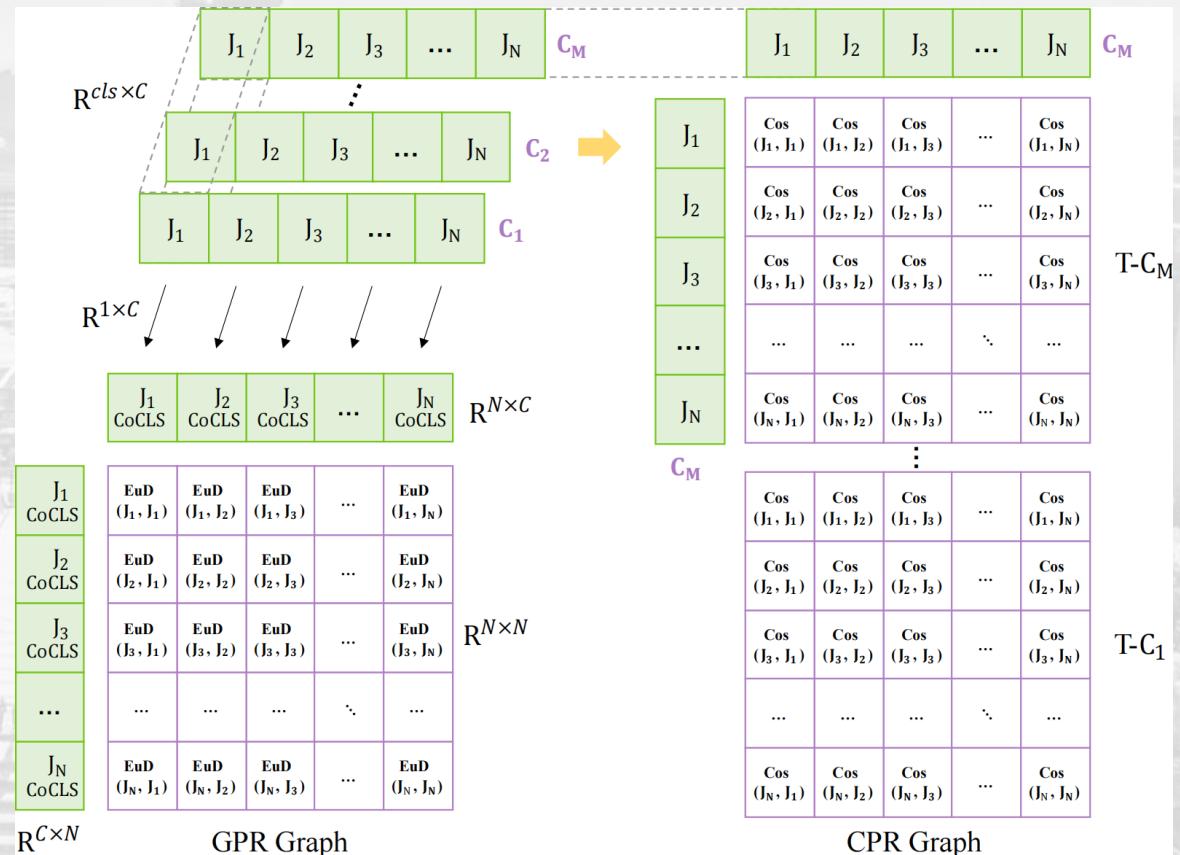
GPR包含LLM中语义知识的全局节点间的关系，通过GPR-Graph来指导骨架表示的生成

对于节点 $J_k (k = 1, 2, \dots, N)$, 将其在 M 个动作类别上的文本特征进行平均，得到该节点的类中心向量：

$$J_k^{CoCLS} = \frac{1}{M} \sum_{j=1}^M J_k^{C_j}$$

$$J^{CoCLS} \in \mathbb{R}^{N \times C}$$

计算任意两个节点的类中心向量之间的相关性，计算方法采取欧氏距离



A Global Prior Relation Graph Generated by LLM

GPR+CPR

XDU

生成先验拓扑图的方法：

GPR图：

通过计算关节的类中心特征相关性；得到全局先验信息，对应于每个节点的动作类中心之间的距离

CPR图：

计算每个动作的节点特征之间的相关性

LLM → BERT → 文本特征 → 特征相关性 →

GPR：先验骨架模态表示
对输入骨架序列进行加权表示

CPR：先验一致性辅助分类
PC-AC模块
额外的分支协同监督

先验的骨架多模态表示

A Priori Skeleton Modal Representation

XDU

人体骨架的
多模态表示

关节：原始关节坐标

骨骼：通过对具有物理连接的关节坐标进行差分得到的向量

关节运动：关节在时间维度上的差分

骨骼运动：骨骼在时间维度上的差分

对于关节 i , 坐标可以表示为 (x_i, y_i, z_i)

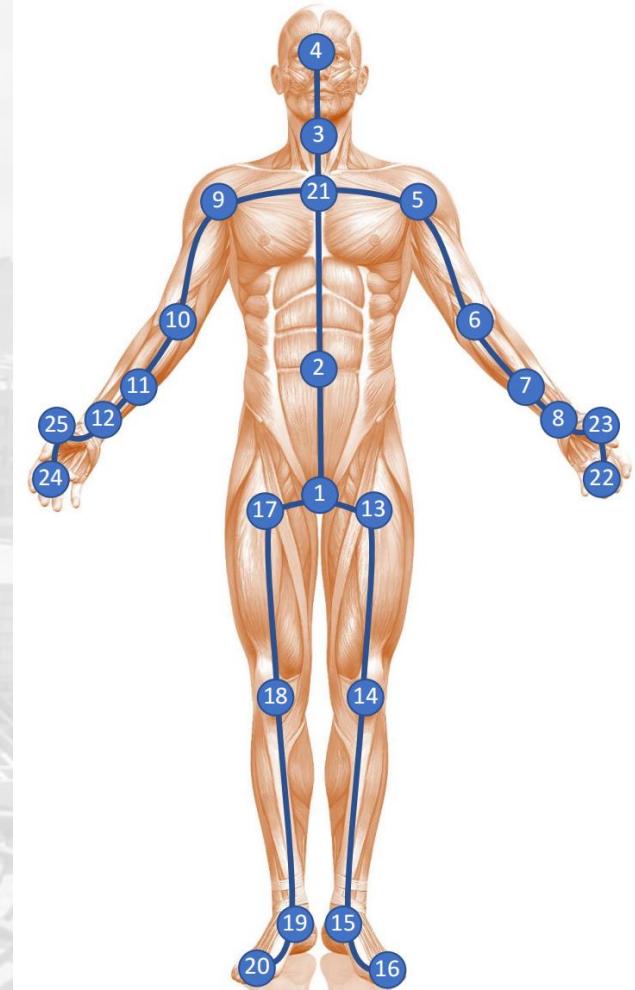
对于骨骼，其可以看做是连接两个相邻的关节之间的通道，因此，对于关节 i, j , 骨骼可以表示为

$$(x_i - x_j, y_i - y_j, z_i - z_j)$$

矩阵形式表示为 $\tilde{X}_t = (I - B)X_t$

其中， X 为 t 时刻的关节特征， \tilde{X}_t 为 t 时刻的骨骼特征

B 为骨矩阵即关节的邻接矩阵



NTU RGB+D的骨架结构

关节、骨骼、关节运动和骨骼运动这四种模式可以被视为多模态表示，它们各自提供了关于人体动作的不同视角和信息。

1. 关节：关节模式通常指的是人体关键关节的位置信息，这些信息可以用来表示人体在空间中的姿态。
2. 骨骼：骨骼模式则关注于连接关节的骨骼长度和方向，它提供了人体结构的几何信息。
3. 关节运动：关节运动模式关注于关节随时间的变化，即关节的角度变化或速度变化，这有助于捕捉动作的动态特征。
4. 骨骼运动：骨骼运动模式则关注于骨骼整体的运动，包括平移和旋转，这有助于理解动作的整体运动轨迹。

虽然这些模式都来源于骨骼数据，但它们各自强调了不同的方面，可以被视为不同的“模态”。在实际应用中，这些信息可以单独使用，也可以组合使用，以提供更全面的动作识别。例如，结合关节的位置信息和关节的运动信息，可以更准确地识别出细微的动作变化。

在多模态学习中，这些信息可以通过各种方式融合，比如在特征层面进行融合，或者在决策层面进行融合，以提高动作识别的准确性和鲁棒性。

我们将节点之间的差异视为附加表示的“骨”数据。例如，NTU数据集包含25个节点，如果我们将任何两个节点之间的差向量视为“骨”，就有 $C_{25}^2 = 300$ 个“骨骼”。

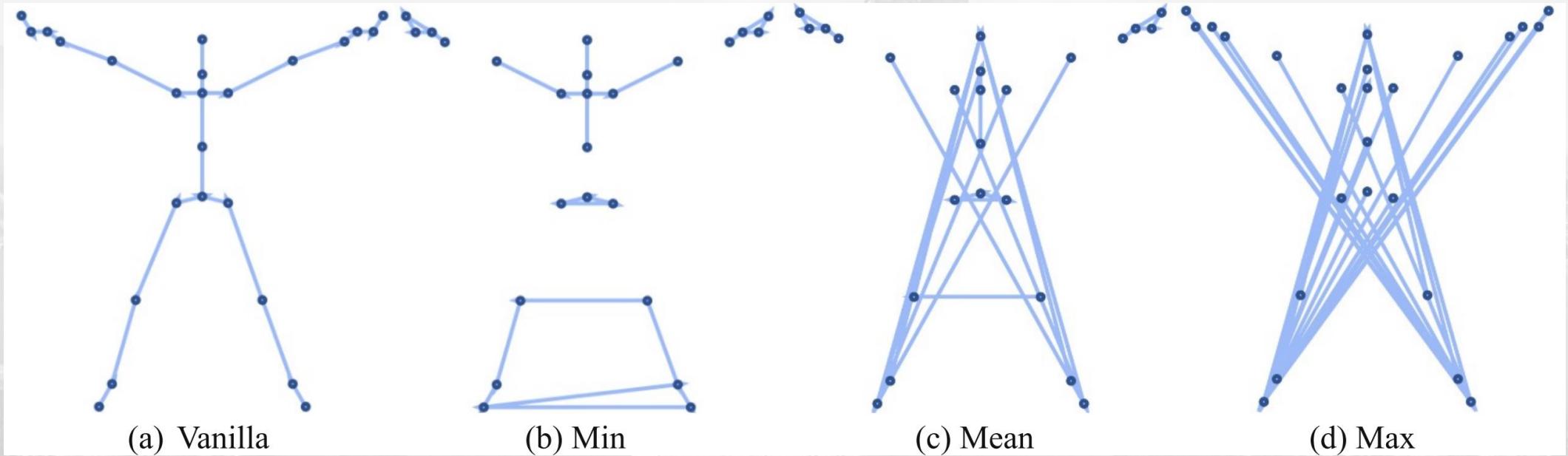
对于任意两个关节点 i, j ，在时刻 t ，设关节特征分别为 $J_i = (x_i, y_i, z_i)$, $J_j = (x_j, y_j, z_j)$ ，设骨骼特征为 $M_t = (x_t, y_t, z_t)$ 其中满足

$$M = J_i - J_j = (x_i - x_j, y_i - y_j, z_i - z_j)$$

定义任意关节 i, j 之间的骨骼标准差：

骨骼的各个特征在时间维度上的标准差的均值，即 $b_{std} = \text{mean}(\sigma(X_t), \sigma(Y_t), \sigma(Z_t))$

经过计算，物理骨骼矩阵中的骨骼链接具有较小的标准偏差总和。基于此，选择标准差之和中最小的骨组合作为新的“骨”表示。**使用GPR图的信息**，即节点间距离对骨骼矢量进行加权，并提取**std 求和最小的骨骼**作为新的骨骼表示。



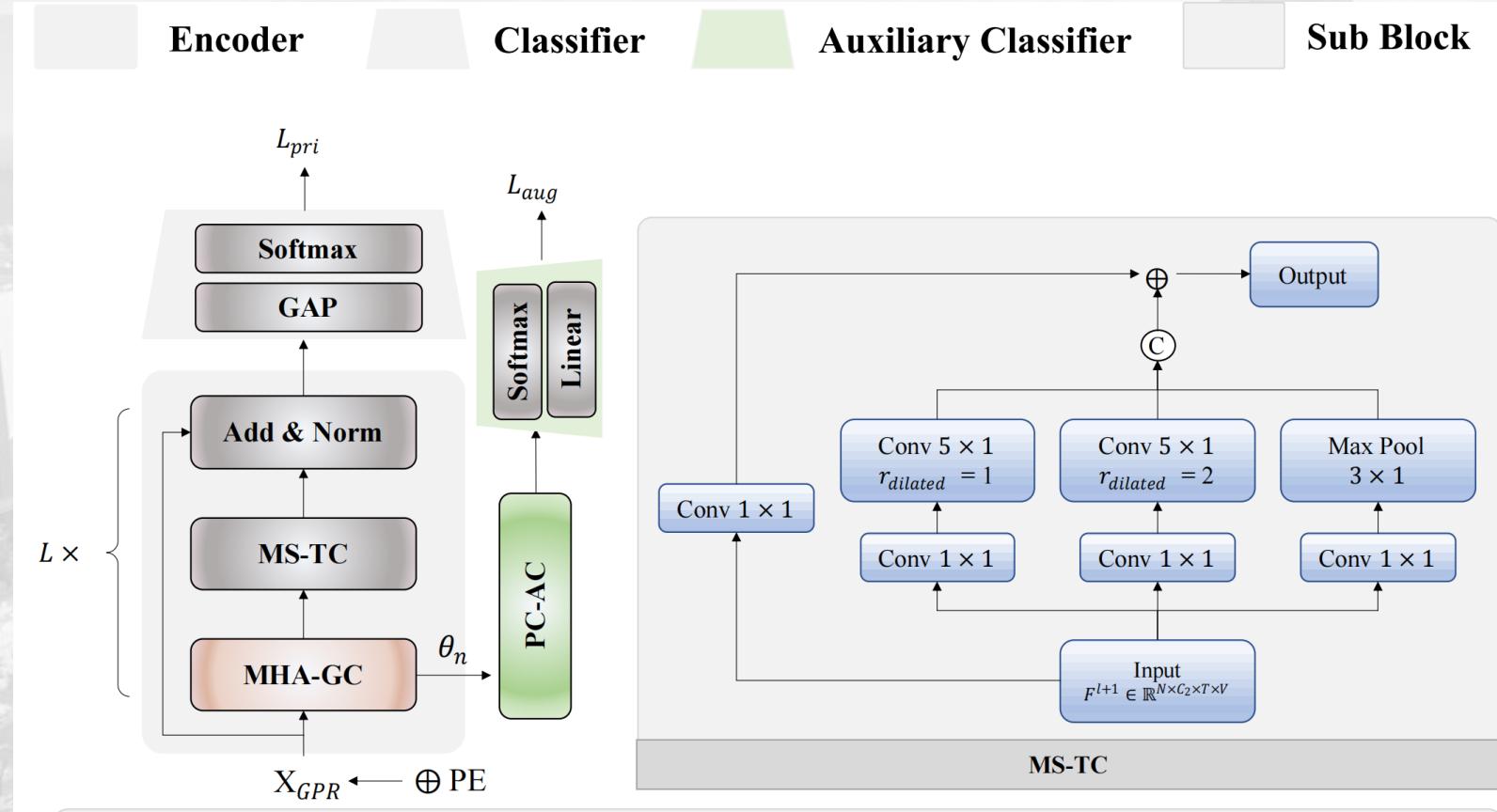
先验模态表示法的两个主要优点：

- (1) 它可以捕捉没有直接连接的远程节点之间的交互。这种交互可能有助于识别某些动作。
- (2) 它包含类属性的上下文依赖性

3 *Part Three*

模型结构

介绍论文中的模型结构，并补充两个创新点，以及引用论文与其进行比较分析。



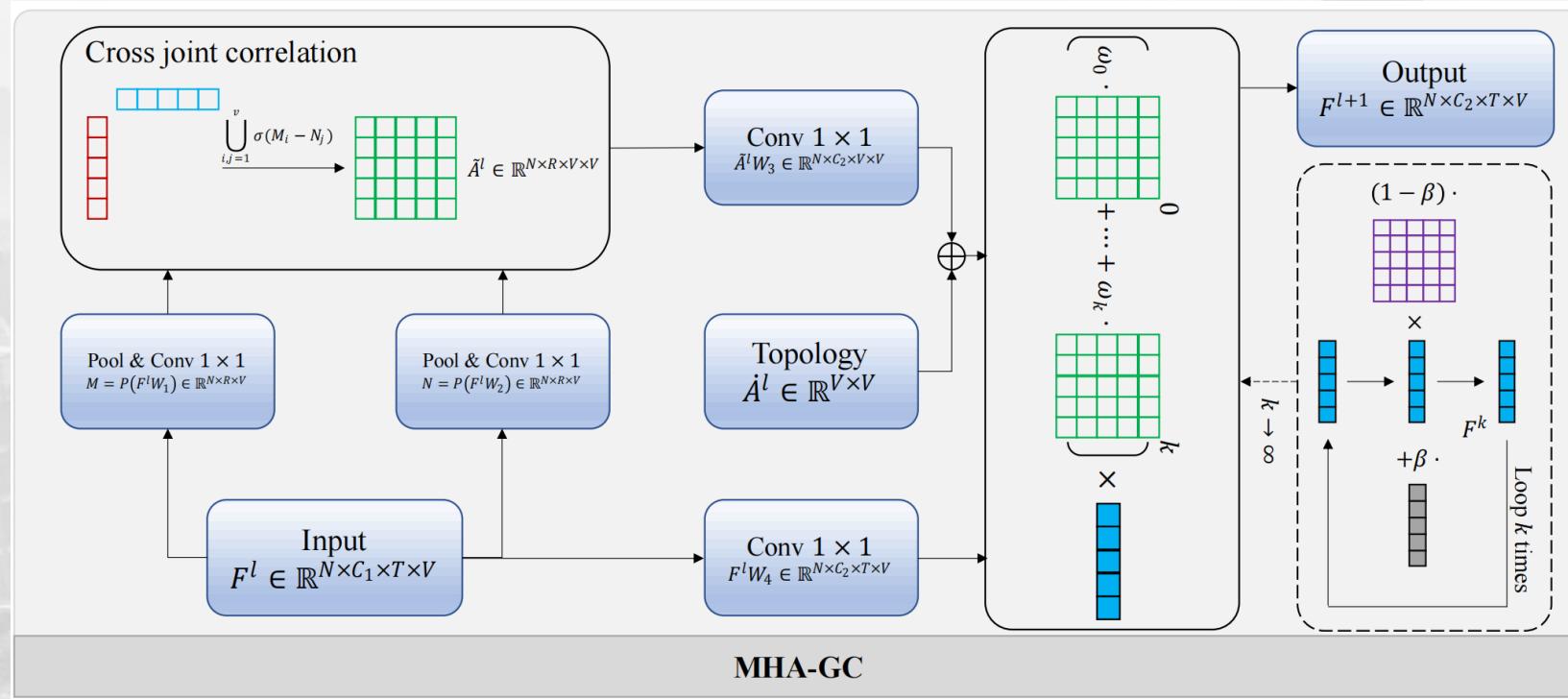
Temporal Conv:

- ✓ 引入多尺度的时间卷积模块，分别用来提取不同时间帧长度的动作特征
- ✓ 提取出的不同长度的特征进行concatenate操作
- ✓ 引入residual残差连接，避免模型深度的增加造成梯度消失，同时有助于原始特征的传播

模型结构

MHA-GC

XDU



Spacial Conv:

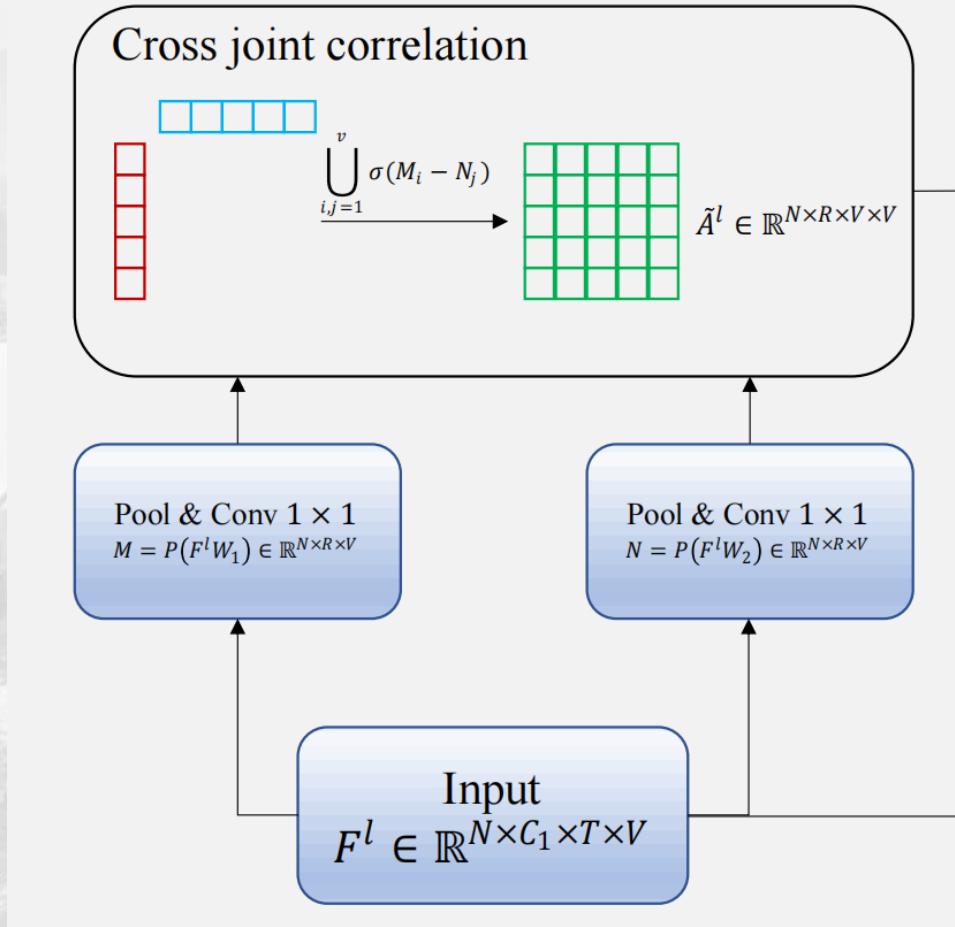
$$\text{Input } \left\{ \begin{array}{l} F^0 = X_{GPR}W_0 + PE \\ F^{(0)} \in \mathbb{R}^{N \times C \times T \times V} \\ PE \in \mathbb{R}^{C \times V} \end{array} \right.$$

经过Conv和Pool:

Channel-Specific Correlations ← Cross joint correlation ←

$$\left[\begin{array}{l} M = P(F^l W_1) \in \mathbb{R}^{N \times R \times V} \\ N = P(F^l W_2) \in \mathbb{R}^{N \times R \times V} \end{array} \right]$$

Cross joint correlation



Feature vectors:

$$M, N \in \mathbb{R}^{N \times R \times V}$$

Any pair of vertices (v_i, v_j) in M and N is computed separately to obtain the first-order neighborhood information

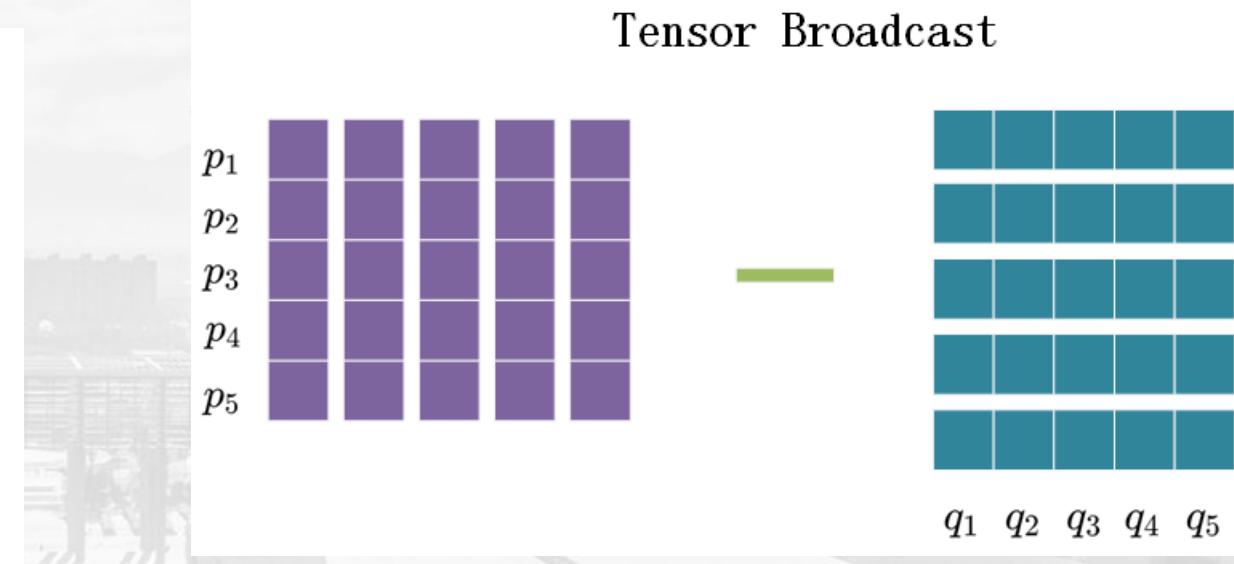
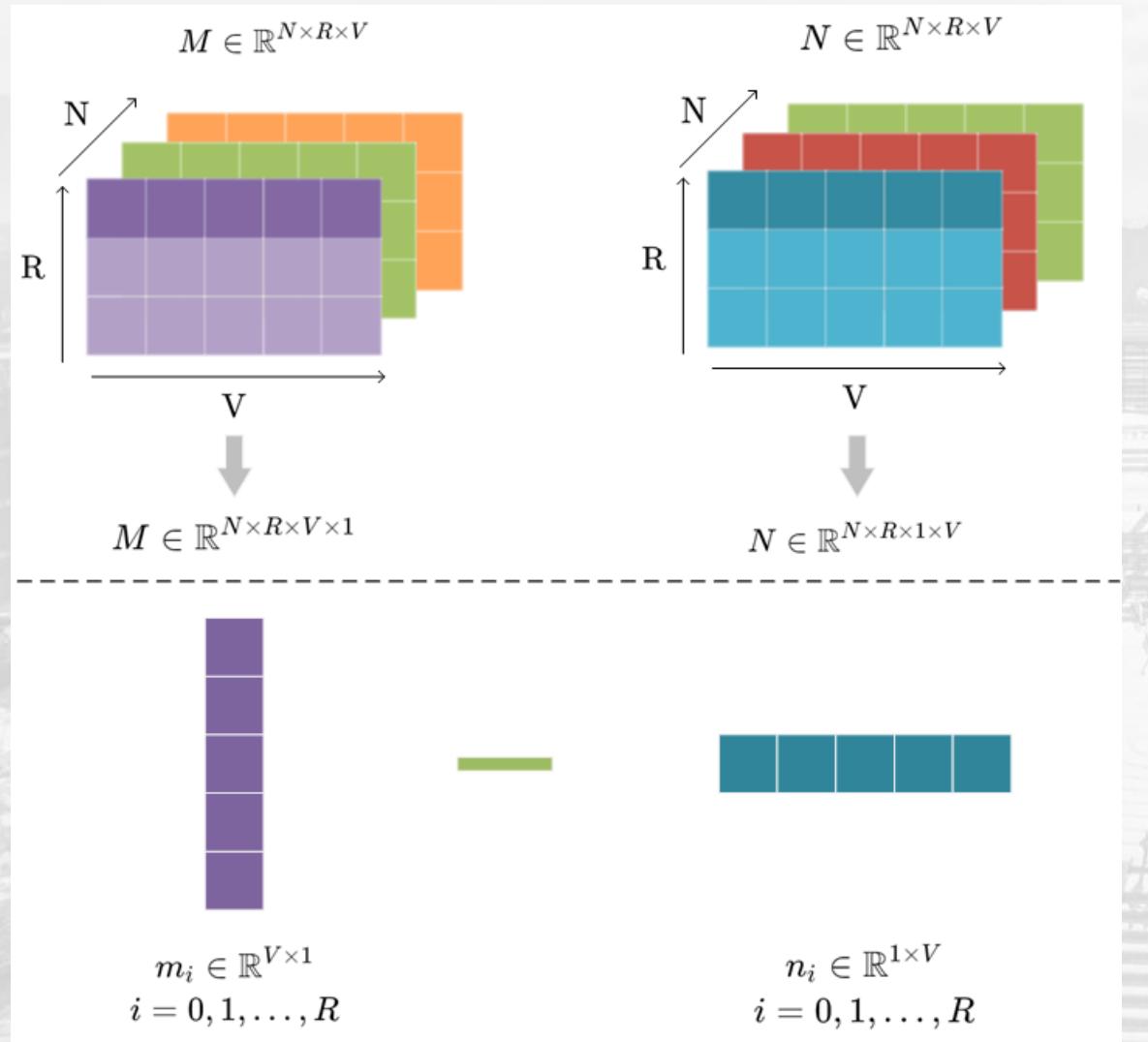
$$\tilde{A}^l = \sum_{i,j=1}^v \sigma(M_i - N_j)$$

where σ is the activation function, v indicates all nodes, and \tilde{A}_{ij}^l denotes the messages aggregation from node j to node i .

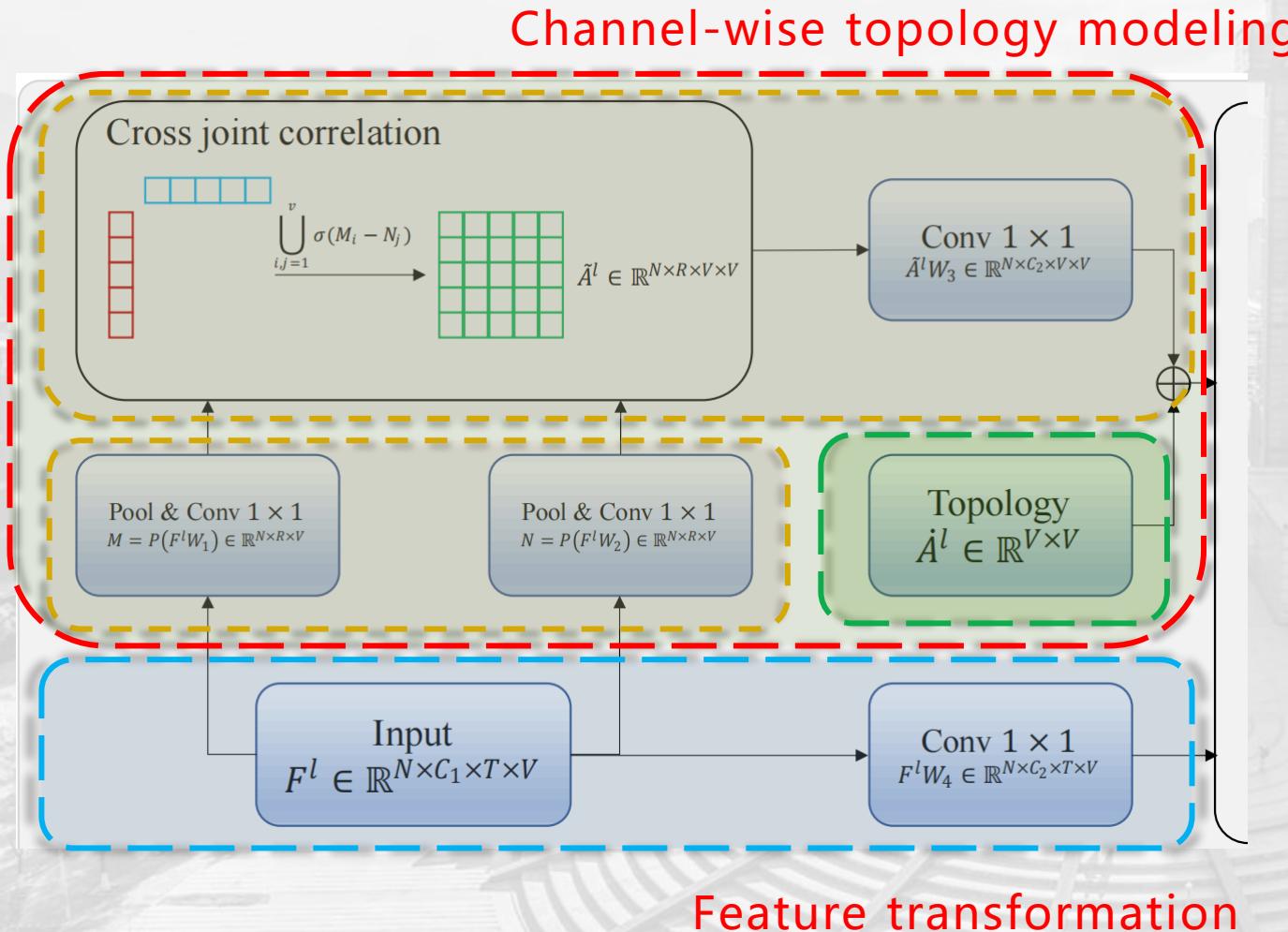
模型结构

Cross joint correlation

XDU



$$Z = \begin{pmatrix} p_1 - q_1 & p_1 - q_2 & p_1 - q_3 & p_1 - q_4 & p_1 - q_5 \\ p_2 - q_1 & p_2 - q_2 & p_2 - q_3 & p_2 - q_4 & p_2 - q_5 \\ p_3 - q_1 & p_3 - q_2 & p_3 - q_3 & p_3 - q_4 & p_3 - q_5 \\ p_4 - q_1 & p_4 - q_2 & p_4 - q_3 & p_4 - q_4 & p_4 - q_5 \\ p_5 - q_1 & p_5 - q_2 & p_5 - q_3 & p_5 - q_4 & p_5 - q_5 \end{pmatrix}$$

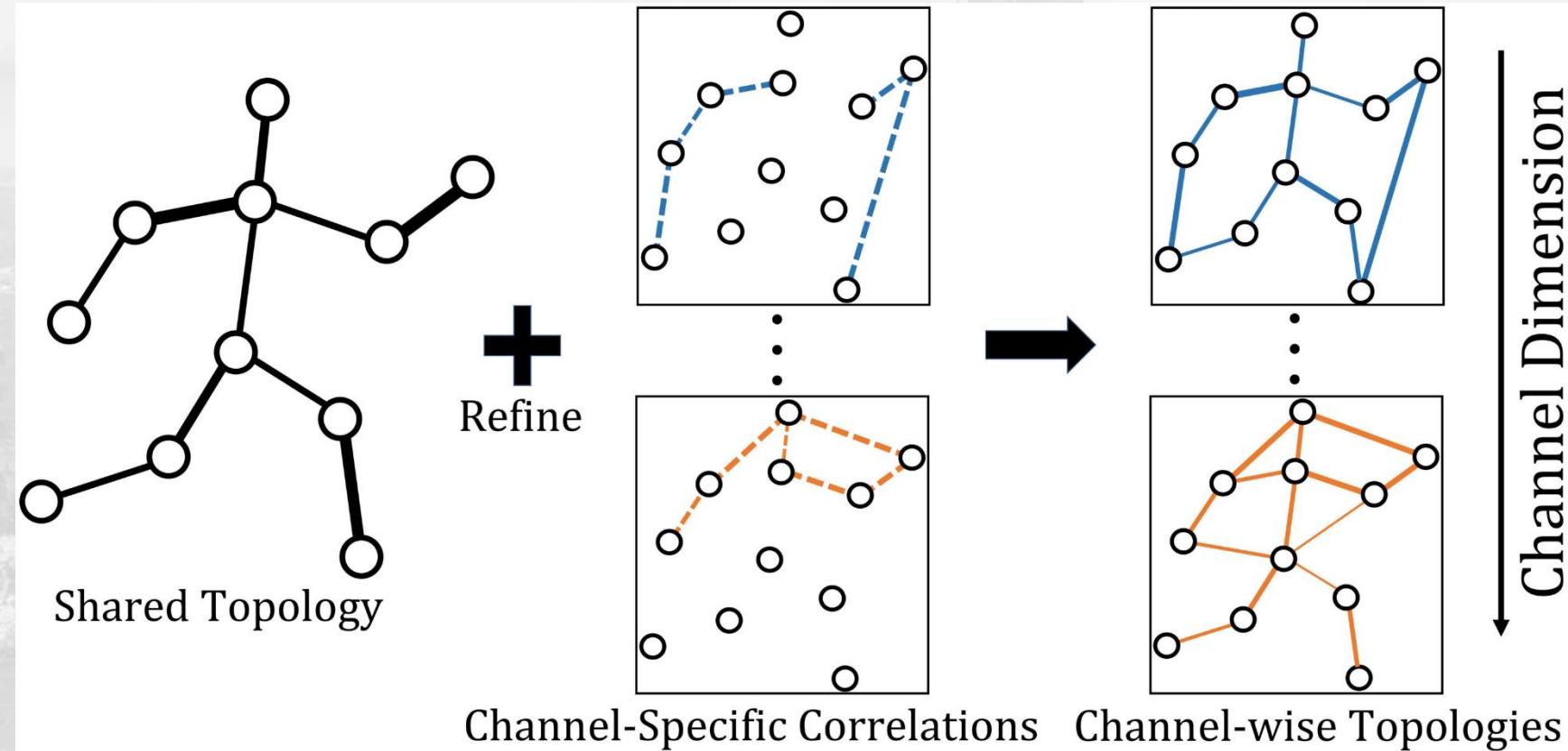


Spacial Conv:

$$\bar{A}^l = \dot{A}^l + \gamma \tilde{A}^l W_3$$

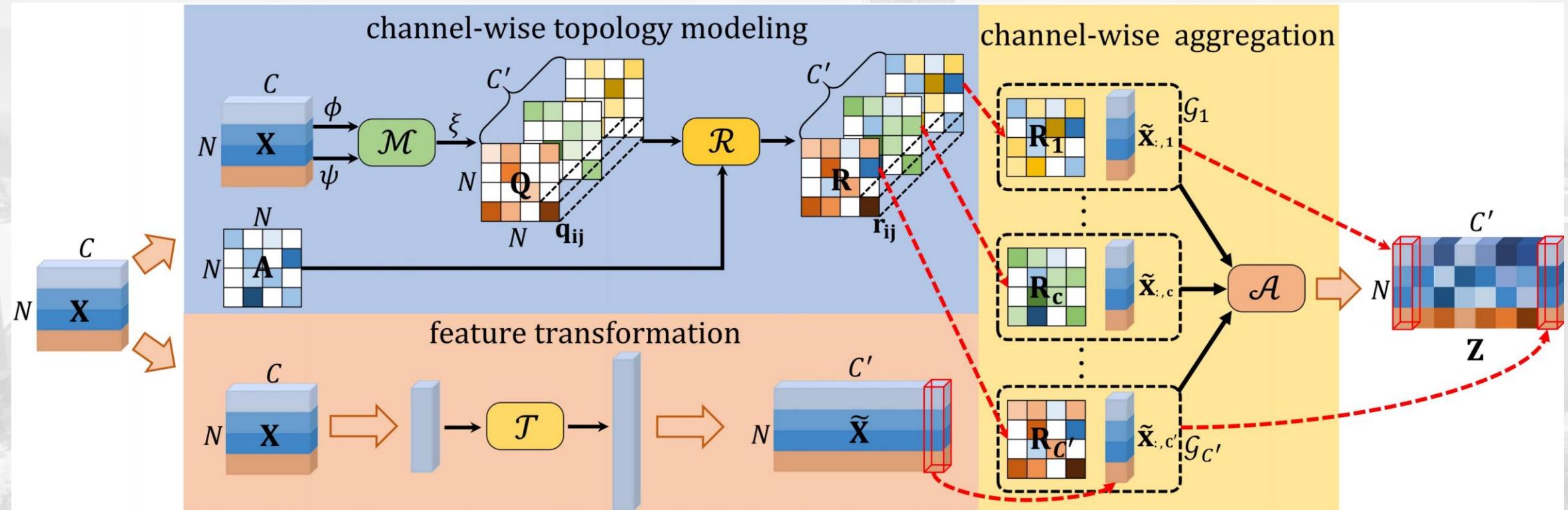
Shared Topology

Channel-Specific Correlations



动态非共享拓扑：共享拓扑+特定通道的关系矩阵 == 通道级拓扑

Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition



Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition[3]

Static Topology-shared GCs

Static Topology-non-shared GCs

Dynamic Topology-shared GCs

Dynamic Topology-non-shared GCs

GCNs已成功应用于基于骨架的动作识别中，其中大部分遵循论文[4]的特征更新规则，**特征更新规则**包括两个步骤：

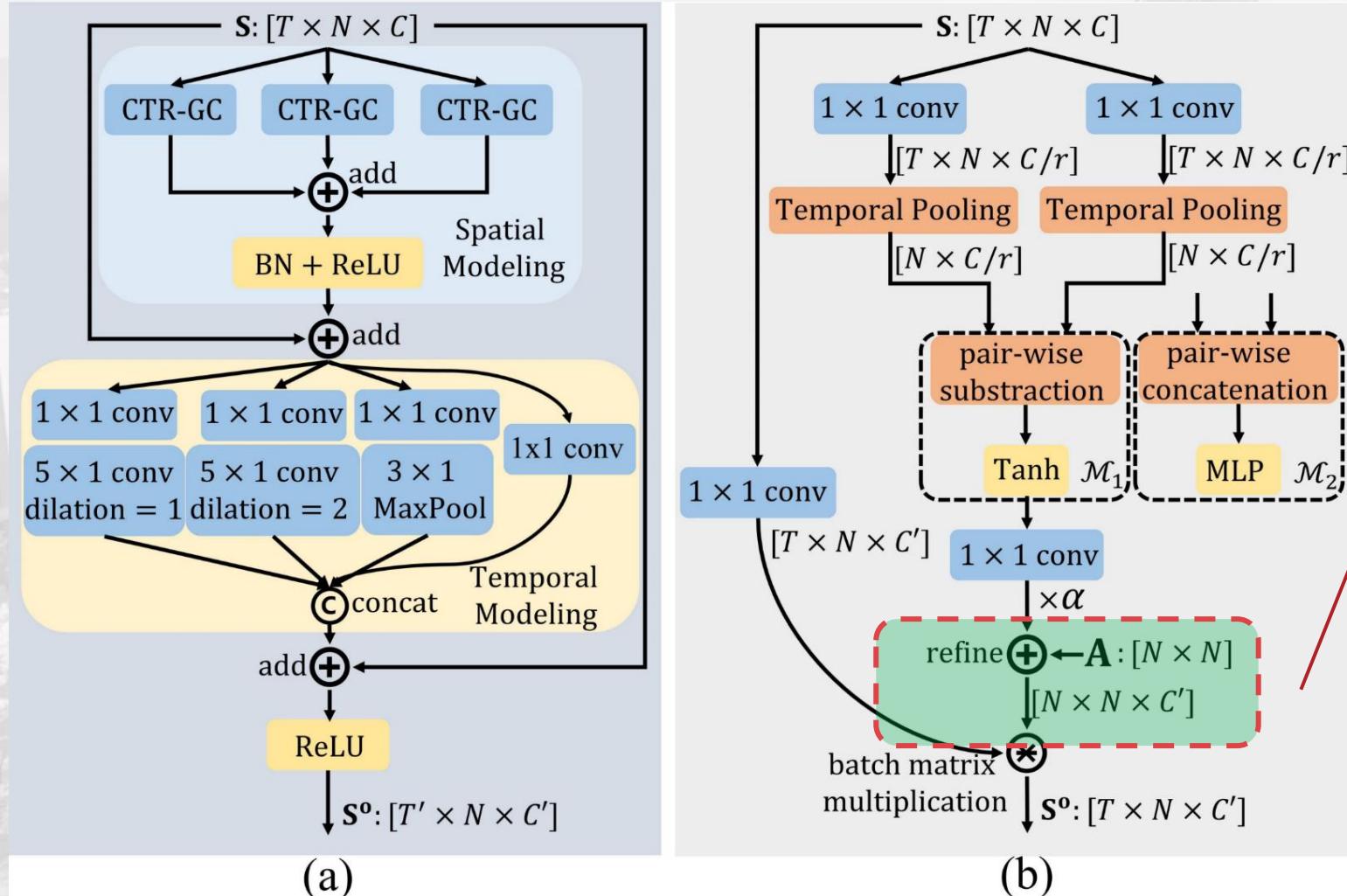
- (1) Transform features into high-level representations;
- (2) Aggregate features according to graph topology.

根据拓扑结构的差异，基于GCN的方法可以分为如下：

- (1) 根据推理过程中拓扑结构是否被动态调整，基于GCN的方法可以分为静态方法和动态方法；
- (2) 根据拓扑是否在不同的通道上共享，基于GCN的方法可以分为拓扑共享方法和拓扑非共享方法。

LA-GCN属于动态非共享拓扑

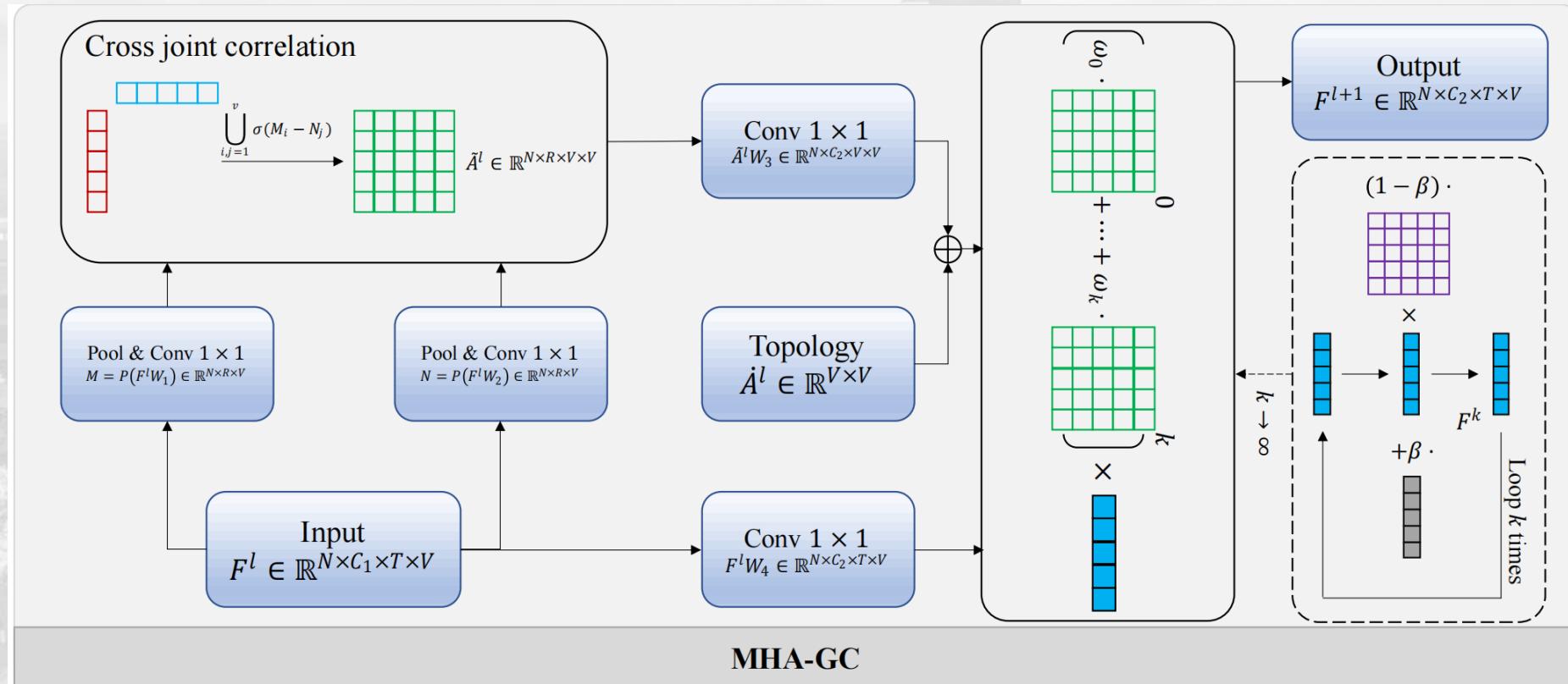
Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition



$$\bar{A}^l = \dot{A}^l + \gamma \tilde{A}^l W_3$$

Before batch matrix
Multiplication and
feature aggregation:

K-Loop



$$F^l \in \mathbb{R}^{N \times C_2 \times T \times V} \quad \bar{\mathcal{A}}^l \in \mathbb{R}^{N \times C_2 \times V \times V} \quad F^{l+1} \in \mathbb{R}^{N \times C_2 \times T \times V}$$

$$F^l \quad \times \quad \bar{\mathcal{A}}^l \quad = \quad F^{l+1}$$

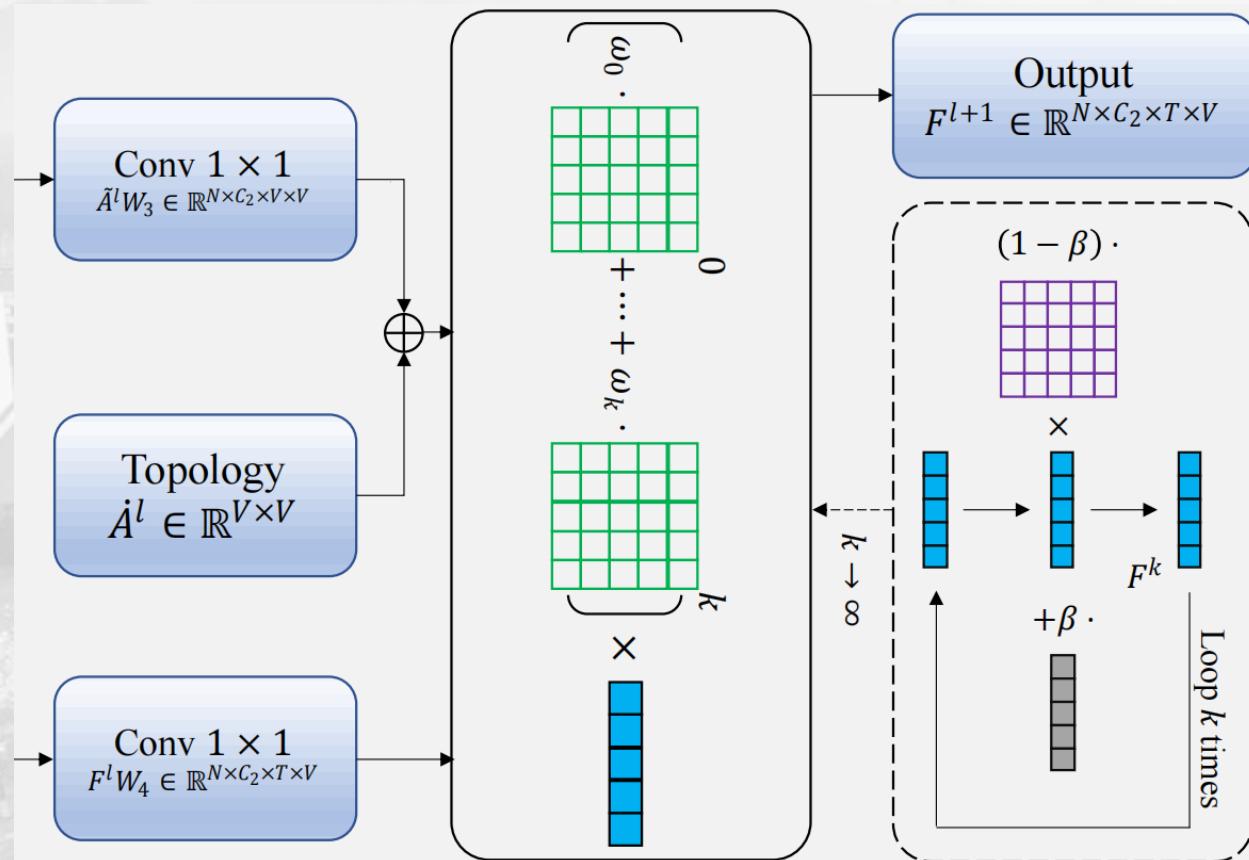
Γ T
V V

$$\times \left[w_0 \cdot \begin{matrix} 0 \\ \vdots \\ V \times V \end{matrix} + \cdots + w_k \cdot \begin{matrix} k \\ \vdots \\ V \times V \end{matrix} \right] =$$

$$\omega_i = \beta(1 - \beta)^i, \beta \in (0, 1]$$

(在特征聚合前, F^l 已经完成输出层的线性映射, 即 $F^l = F^l W_4^l$)

$$\left\{ \begin{array}{l} \overline{\mathcal{A}}^l = \sum_{i=0}^k \omega_i \overline{A}^i \\ \omega_i = \beta(1-\beta)^i, \beta \in (0,1] \\ \omega_i \quad \overline{A} \text{ 的延迟因子} \\ F^{l+1} = \sigma(\overline{\mathcal{A}}^l F^l W_4^l) \end{array} \right.$$

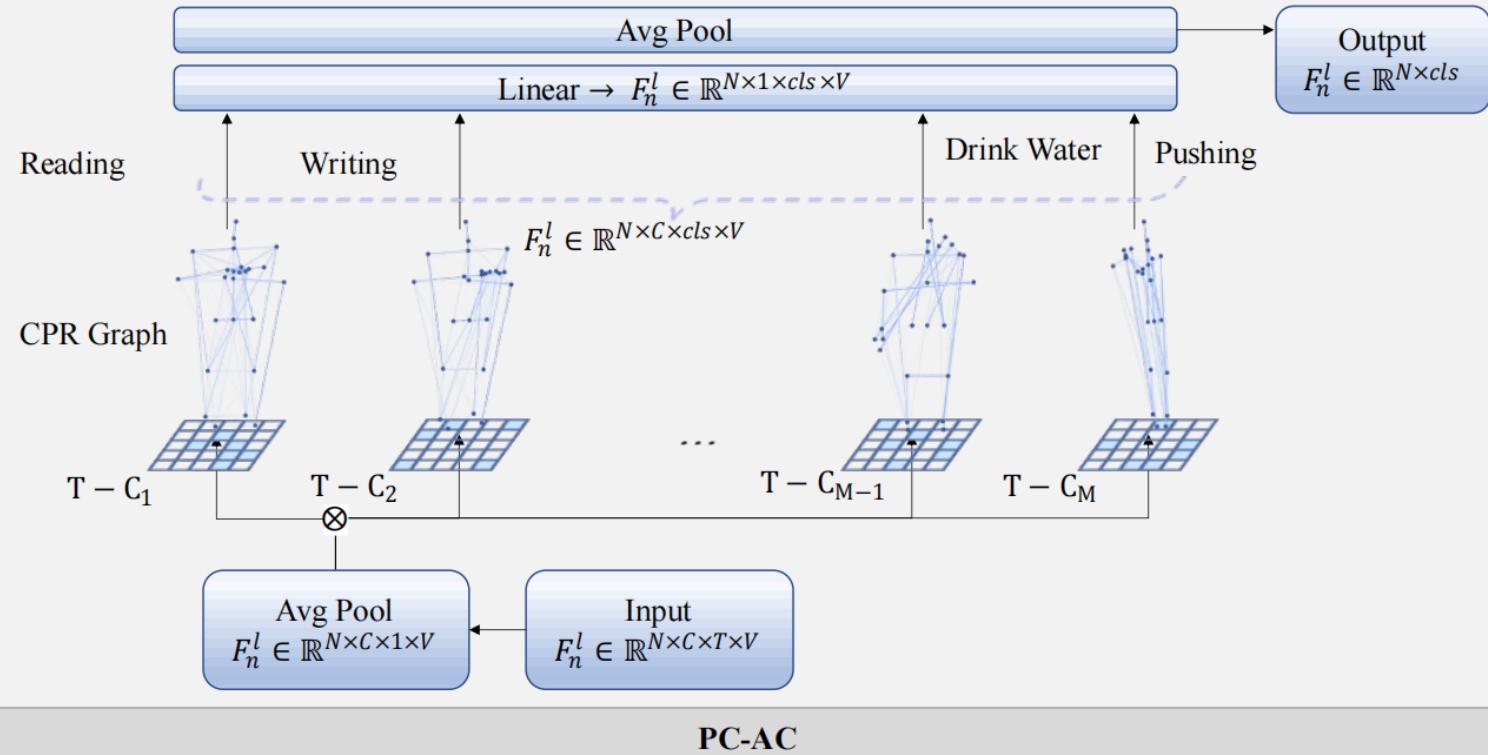


$$F^{k+1} = (1 - \beta) \bar{A} F^k + \beta F^l$$

$$\lim_{K \rightarrow \infty} F^K = \bar{A} F^l$$

骨架序列包含固定数量的边 ε ，通过上述近似，我们得出了**多跳注意的复杂性与单跳注意没有太大的区别**。复杂度都可以表示为 $O(|\varepsilon|)$

PC-AC+A Priori Consistency-Assisted Classification Module



先验一致性辅助分类 (PC-AC)

利用**类关系拓扑图CPR**，包含与预测操作相关的先验节点关系，以帮助GCN执行**节点特征聚合**。

PC-AC模块在训练时向GCN添加了额外的**分支协同监督**，迫使每个特征通过一个包含特定类别拓扑的分支，从而为该类进行预测。

PC-AC+A Priori Consistency-Assisted Classification Module

损失函数： the cross-entropy loss

$$L_{aux} = - \sum_k y_k \cdot \log(\hat{y}_k)$$

$$L_{pri} = - \sum_k y_k \cdot \log(\hat{y}_k)$$

$$\hat{y}_k = \frac{e^{z_i}}{\sum_k e^{z_k}}$$

目标： $\arg \min_{\theta} (L(f_{\theta}^{pri}(x), \hat{y}) + \lambda L(f_{\theta_n}^{aux}(x), \hat{y}))$

4 Part Four

实验设计及结果

简要介绍实验的设计和实验结果，突出本篇论文的创新点

实验设计及结果

多流融合策略 (关节、骨骼、关节运动、骨骼运动)

XDU

TABLE 1: The comparison of Top-1 accuracy (%) on the NTU RGB+D [31] bench

—— UCLA [33] benchmark.

VA	25.5
ST	57.9
AS	61.8
2s-AGC	73.2
Direc	84.9
S	84.9
Shi	87.1
DC-G	87.6
Dyna	88.4
M	88.6
DI	88.8
MS	88.9
Effic	90.6
CT	90.7
Inf	90.7
LA	88.0

Methods	Top1
Lie Group [57]	74.2
Actionlet ensemble [58]	76.0
HBRNN-L [59]	78.5
Ensemble TS-LSTM [38]	89.2
AGC-LSTM [39]	93.3
4s Shift-GCN [40]	94.6
DC-GCN+ADG [54]	95.3
CTR-GCN [8]	96.5
InfoGCN [6]	97.0
LA-GCN (ours)	97.6

X-Set
25.5
57.9
61.8
73.2
84.9
84.9
87.1
87.6
88.4
88.6
88.8
88.9
90.6
90.7
90.7
88.0
90.9
91.3
91.8

TABLE 4: (i) The comparison of accuracy without and with PC-AC, where the λ of L_{aug} is 0.2. (ii) Contribution of different text prompts.

Methods	Top1
w/o L_{aug}	84.9
L_{total}	86.1 \uparrow 1.2
p1: [J] function in [C].	85.6 \uparrow 0.7
p2: What happens to [J] when a person is [C]?	85.8 \uparrow 0.9
p3: What will [J] act like when [C]?	86.1 \uparrow 1.2
p4: When [C][J] of human body.	85.5 \uparrow 0.6
p5: When [C] what will [J] act like?	85.7 \uparrow 0.8
p6: When a person is [C], [J] is in motion.	85.5 \uparrow 0.6

多流融合策略 (关节、骨骼、关节运动、骨骼运动)

TABLE 7: λ selection of PC-AC with different text prompts T-C.

λ	0.1	0.2	0.3	0.5
p1: [J] function in [C].	85.0	85.6	85.3	84.8
p2: What happens to [J] when a person is [C]?	85.1	85.8	85.5	85.1
p3: What will [J] act like when [C]?	85.3	86.1	85.6	85.3
p4: When [C][J] of human body.	85.0	85.5	85.0	84.8
p5: When [C] what will [J] act like?	85.2	85.7	85.2	84.9
p6: When a person is [C], [J] is in motion.	85.1	85.5	85.2	85.0

TABLE 9: The accuracy (%) of the five-stream (5s) and six-stream (6s) ensemble on the NTU RGB+D 120 X-Sub split between the traditional four-stream (4s) and the new “bone” representatives. The basis is the joint acc: 86.5%. The model accuracy (%) of the new “bone” representations $\{B_{p1}, B_{p2}, B_{p3}, B_{p4}, B_{p5}, B_{p6}\}$ on the NTU RGB+D 120 X-Sub split are 85.9, 86.0, 84.9, 85.9, 85.8, and 85.8, respectively.

Modes	Components	Acc
2s	Joint + Bone	89.7 \uparrow 3.2
4s	J + B + JM + BM	89.9 \uparrow 3.4
5s	4s + B_{p1}	90.1 \uparrow 3.6
	4s + B_{p2}	90.4 \uparrow 3.9
	4s + B_{p3}	90.3 \uparrow 3.8
	4s + B_{p4}	90.1 \uparrow 3.6
	4s + B_{p5}	90.4 \uparrow 3.9
	4s + B_{p6}	90.2 \uparrow 3.7
6s	4s + $B_{p2} + B_{p5}$	90.7 \uparrow 4.2

5 Part Five

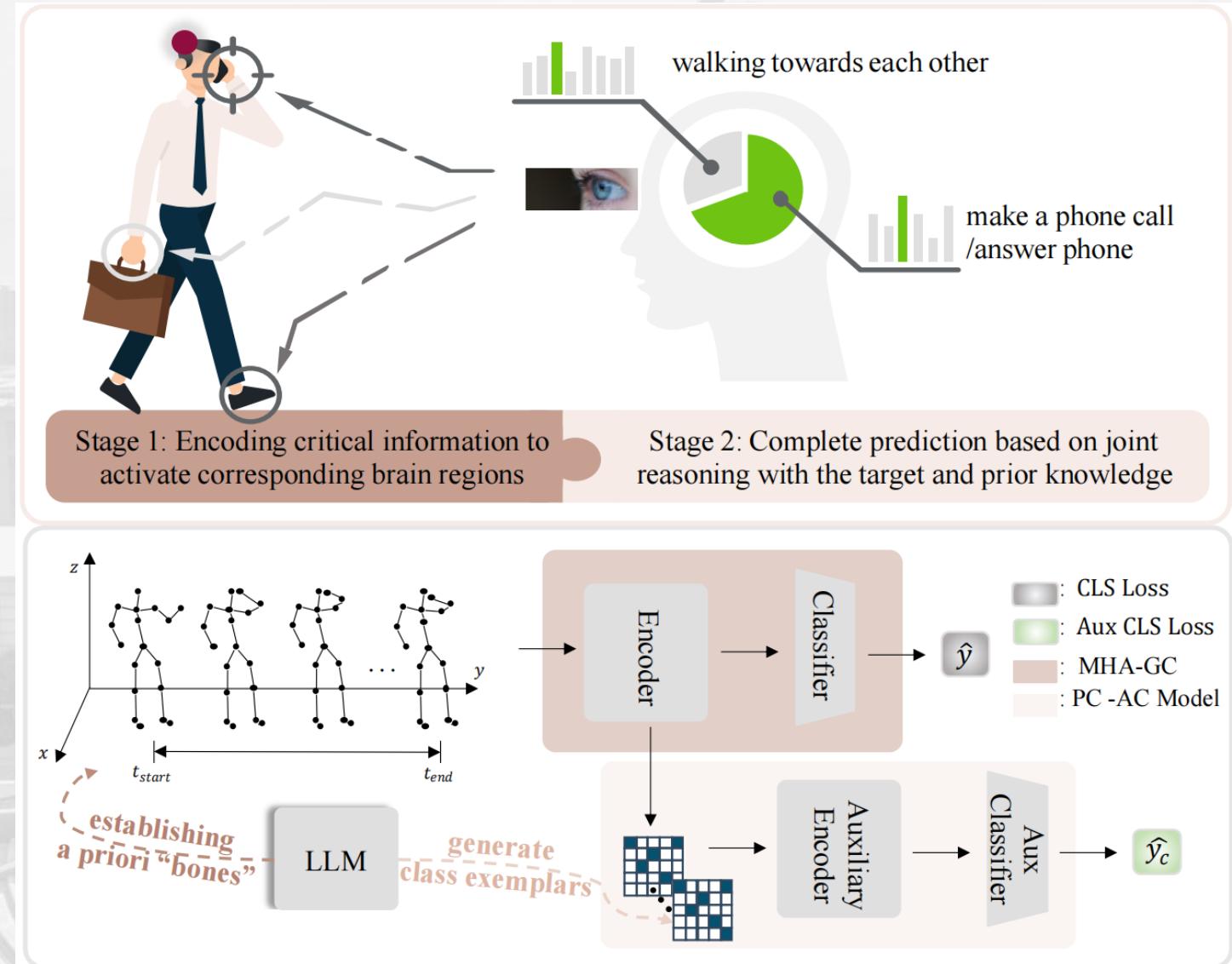
论文总结

对论文的整体思想，突出的创新点和结论进行归纳总结

论文总结

整体结构+突出创新点

XDU



- [1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 7444–7452. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135>
- [2] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2019, pp. 12 026–12 035.
- [3] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel wise topology refinement graph convolution for skeleton-based action recognition," in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021, pp. 13 339–13 348.
- [4] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016. 2

XDU

THANK FOR ATTENTION

厚德 求真 励学 笃行