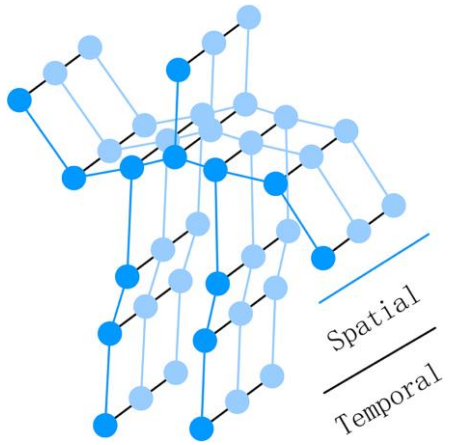
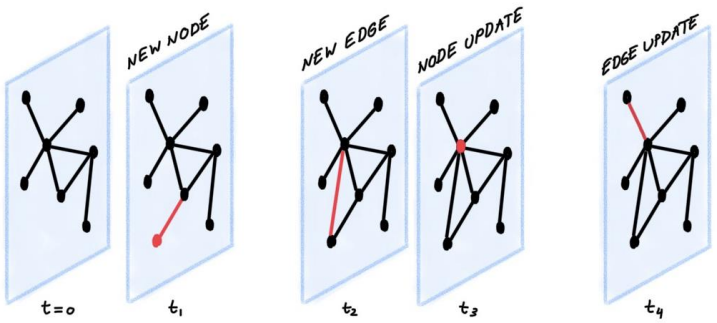
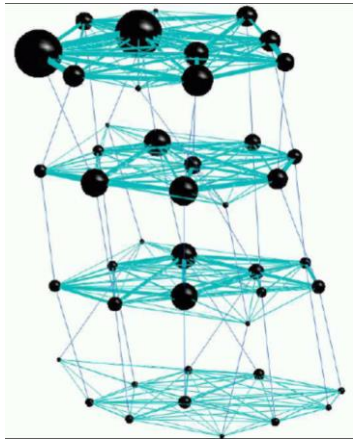
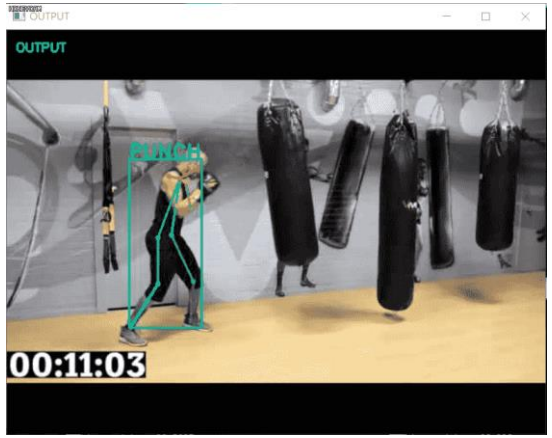
Les GNN pour les graphes spatio-temporels dans la reconnaissance d’actions

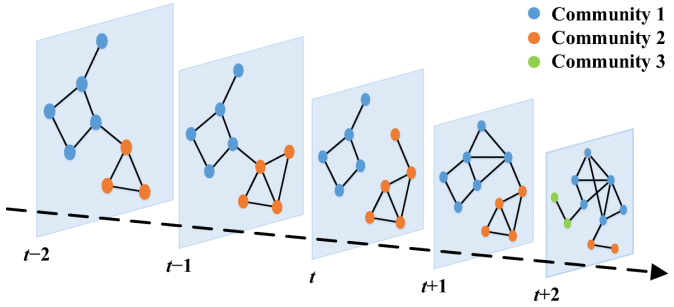






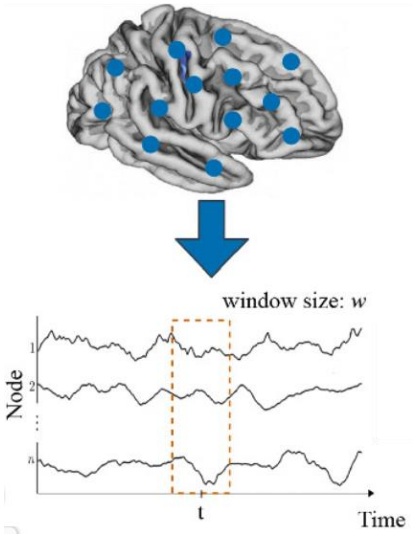


Les sequences de graphes: domaines d'applications



Réseaux sociaux

Vision par ordinateur

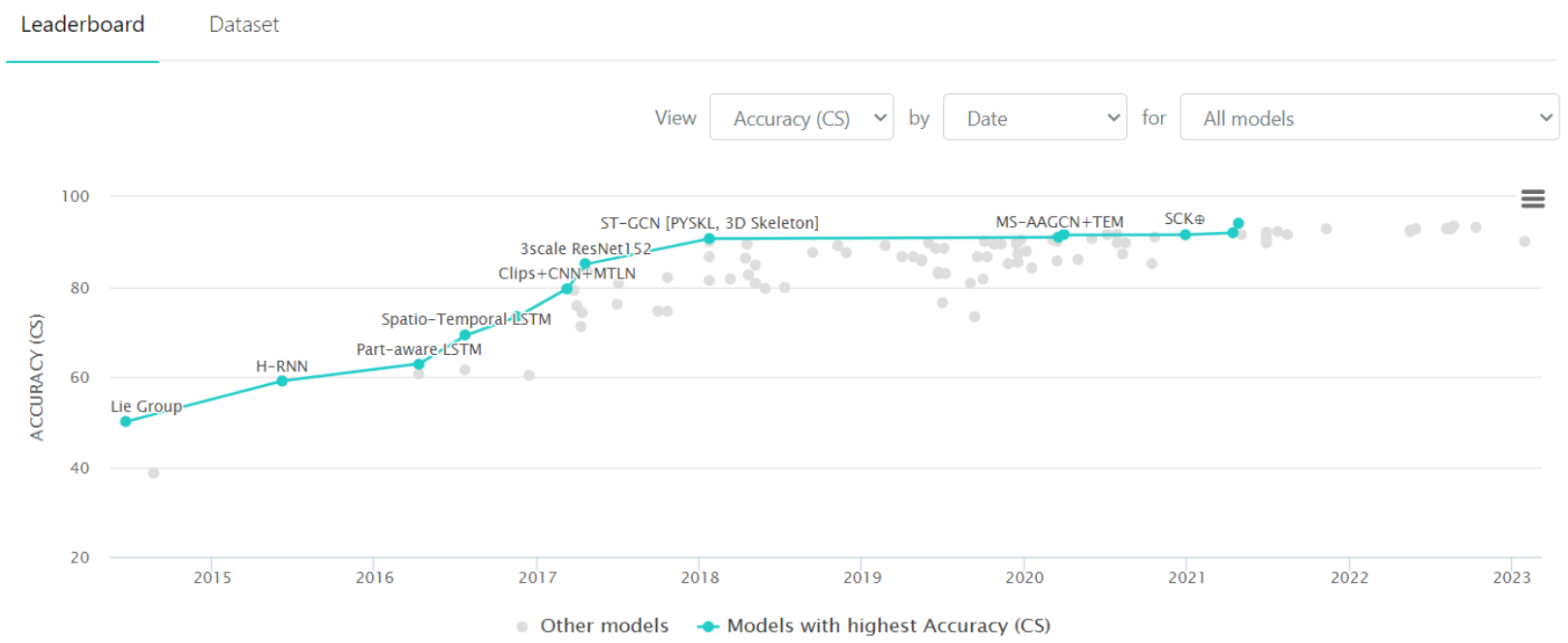


Brain Connectome

STG dans la reconnaissance d’actions

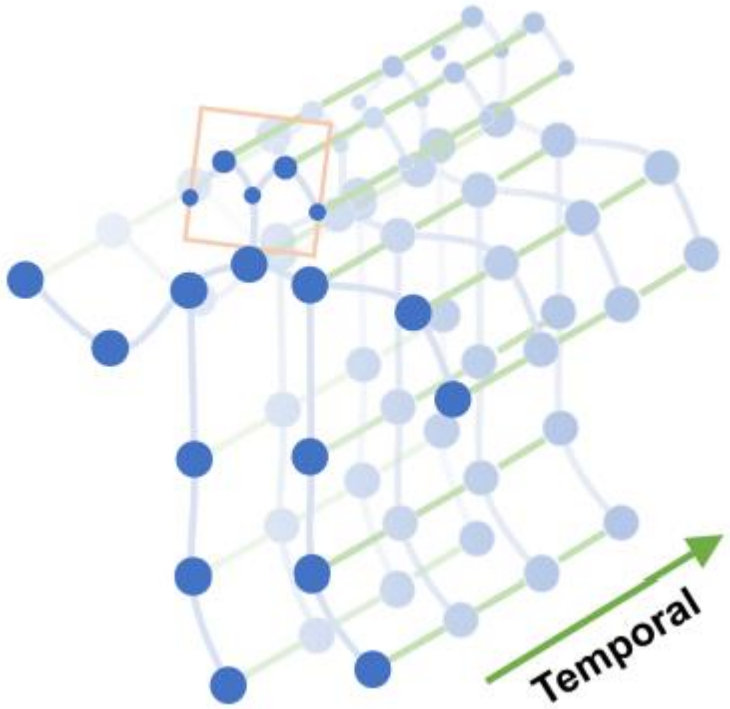
- Un cadre bien défini avec des bases de données publiques pour les benchmarks

Skeleton Based Action Recognition on NTU RGB+D



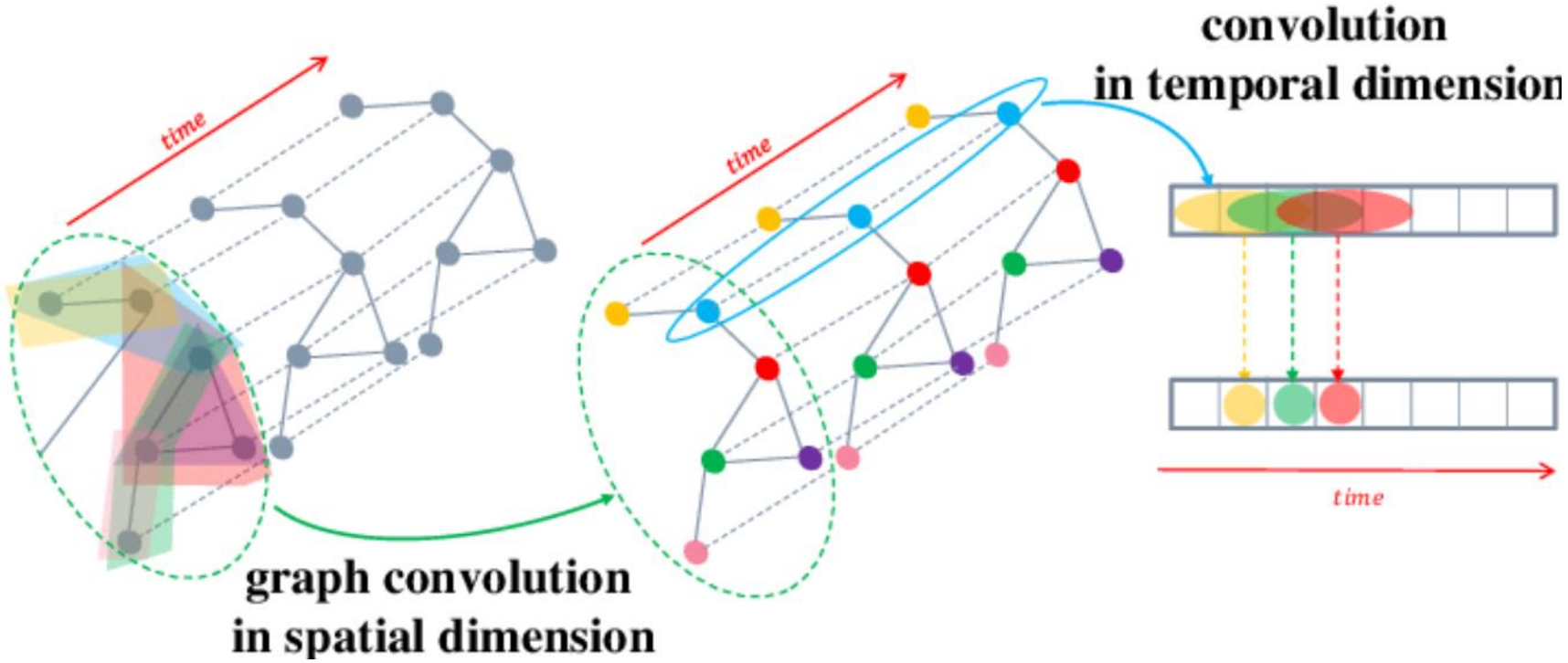
La représentation spatio-temporelle

[1]



[1]Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Thirty-second AAAI conference on artificial intelligence.

ST-GNN: principe de base



[1]

[1]Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019, July). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 922-929).

Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition

de Zhan Chen,1 Sicheng Li,2 Bing Yang,1 Qinghan Li,3 Hong Liu1

Reprends les considerations de l'article ST-GCN

Partition des voisins en fonction de la proximité du centre de gravité/ l' éloignement du centre de gravité par rapport au nœud sur lequel s' effectue la convolution.
Effectue la convolution frame par frame sur les caractéristiques du squelette
Puis convolution temporelles
Mais …

Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition

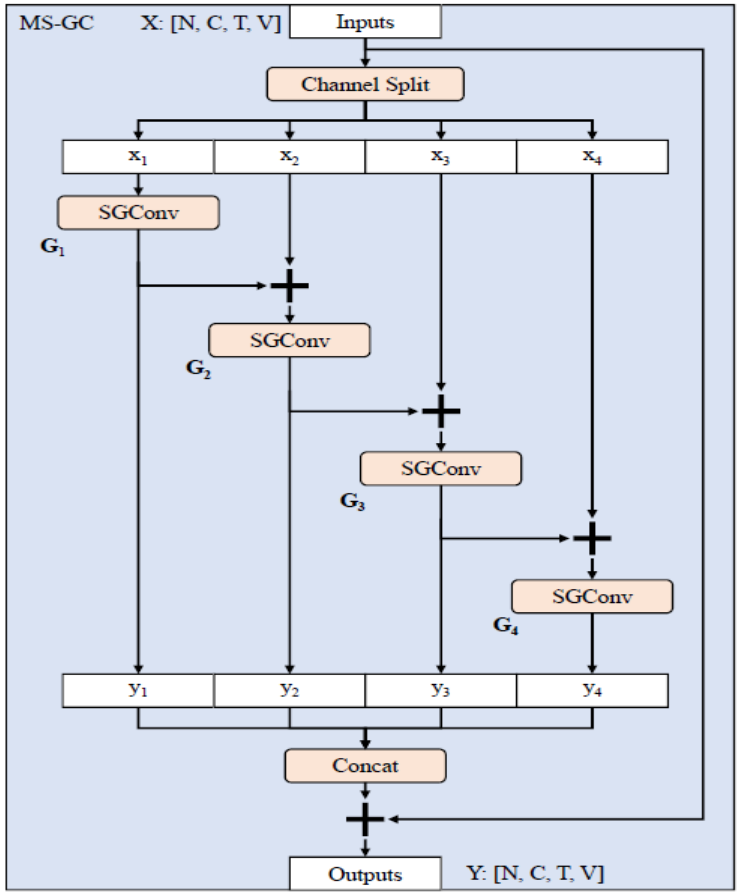


Figure 2: Illustration of multi-scale spatial graph convolution (MS-GC) module. SGConv denotes the spatial graph convolution, N is the batch size.

$$\mathbf{y}_i = \begin{cases} \mathbf{G}_i(\mathbf{x}_i) & i = 1 \\ \mathbf{G}_i(\mathbf{x}_i + \mathbf{y}_{i-1}) & i > 1 \end{cases}$$

Au lieu de faire la convolution sur les nœuds directement :
La convolution s'effectue caractéristique par caractéristique et est agrégé au fur et à mesure : ce qui permets en une étape de convolutions de condenser l'information de voisins éloignés de plus d'un saut.

Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition
Zhan Chen,1 Sicheng Li,2 Bing Yang,1 Qinghan Li,3 Hong Liu1*

Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition

Ajout à la convolution du bloc temporelle

$$\mathbf{y}_i = \begin{cases} \mathbf{T}_i(\mathbf{G}_i(\mathbf{x}_i)) & i = 1 \\ \mathbf{T}_i(\mathbf{G}_i(\mathbf{x}_i + \mathbf{y}_{i-1})) & i > 1 \end{cases}$$

Permets a des frames non contigue de partager de l'information en plus de l 'agrégation d'information entre articulation éloigné.

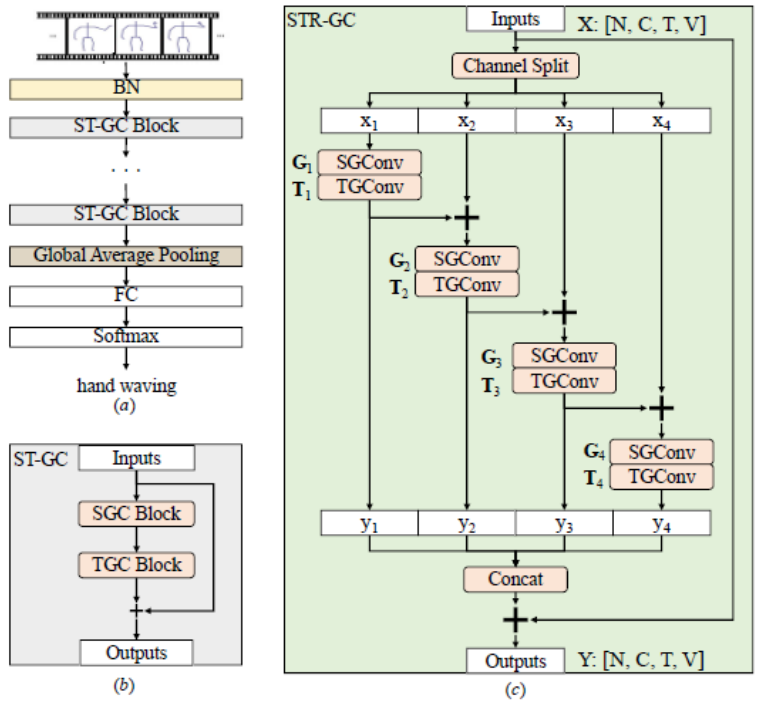


Figure 3: Architecture overview. (a) The full architecture of the ST-GCN backbone, BN is a batch normalization layer, FC is a full connected layer. (b) Illustration of ST-GC block which is used to build the whole network. (c) Illustration of proposed spatial temporal residual graph convolution (STR-GC) module, TGConv is the temporal graph convolution operation.

Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition
Zhan Chen,1 Sicheng Li,2 Bing Yang,1 Qinghan Li,3 Hong Liu1*

Résultat.

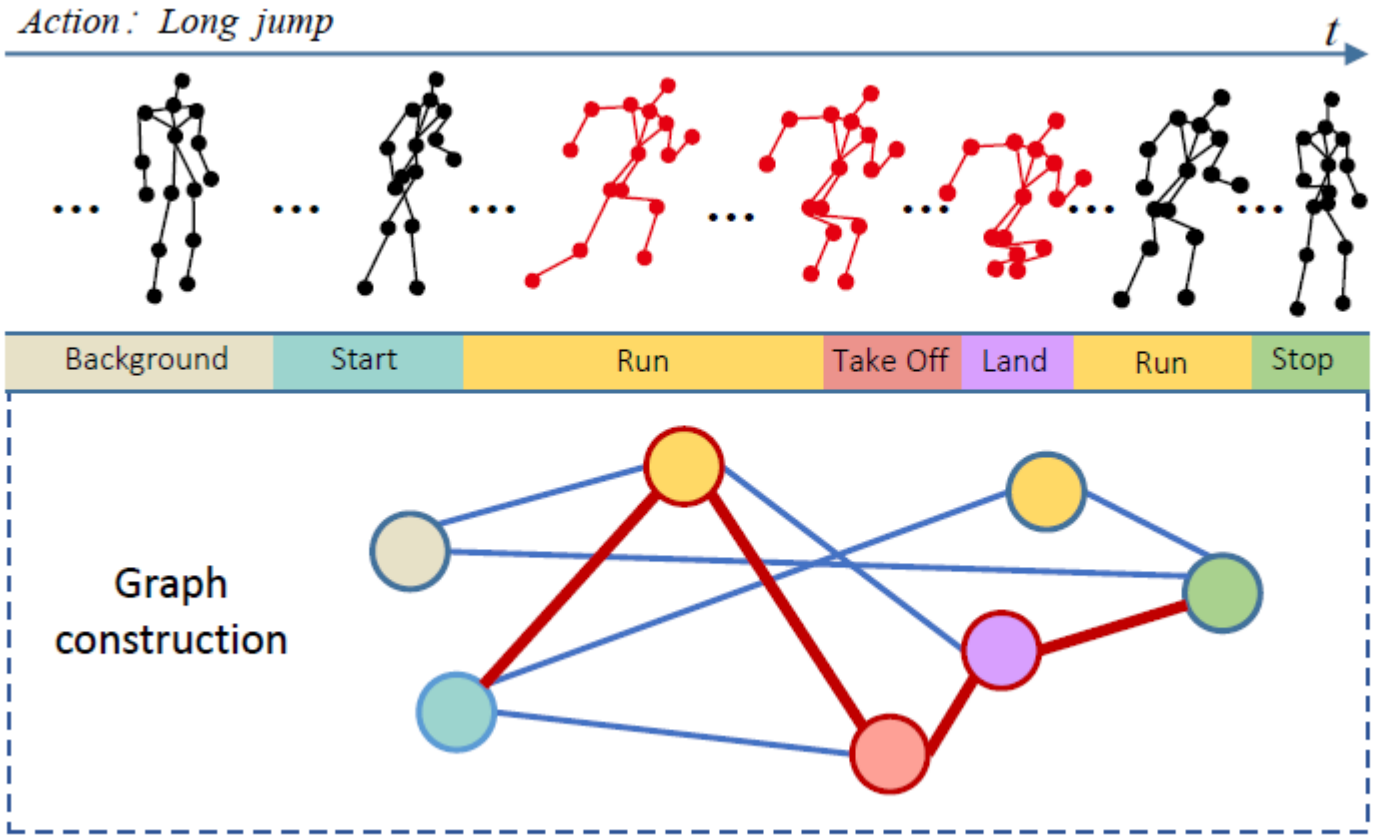
Methods	X-view (%)	X-sub (%)
HBRNN (Du et al. 2015)	64.0	59.1
P-LSTM (Shahroudy et al. 2016)	67.3	60.7
TCN (Soo Kim and Reiter 2017)	83.1	74.3
VA-LSTM (Zhang et al. 2017)	87.7	79.2
ST-GCN (Yan, Xiong, and Lin 2018)	88.3	81.5
AS-GCN (Li et al. 2019a)	94.2	94.2
2s AGC-LSTM (Si et al. 2019)	95.0	89.2
2s AGCN (Shi et al. 2019c)	95.1	88.5
2s NAS-GCN (Peng et al. 2020)	95.7	89.4
4s DGNN (Shi et al. 2019a)	96.1	89.9
4s MS-AAGCN (Shi et al. 2019b)	96.2	90.0
2s MS-G3D (Liu et al. 2020)	96.2	**91.5**
4s Shift-GCN (Cheng et al. 2020)	96.5	90.7
Js MST-GCN (ours)	95.1	89.0
Bs MST-GCN (ours)	95.2	89.5
2s MST-GCN (ours)	96.4	91.1
4s MST-GCN (ours)	**96.6**	**91.5**

Table 4: Comparisons of the Top-1 accuracy with the state-of-the-art methods on the NTU RGB+D dataset.

Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition
Zhan Chen,1 Sicheng Li,2 Bing Yang,1 Qinghan Li,3 Hong Liu1*

Temporal Graph Modeling for Skeleton-based Action Recognition

Idée général :
Prendre en compte l'aspect temporel en construisant un graphe des relation temporelle



Temporal Graph Modeling for Skeleton-based Action Recognition
Jianan Li, Xuemei Xie, Zhifu Zhao, Yuhan Cao, Qingzhe Pan and Guangming Shi Fellow, IEEE

Temporal Graph Modeling for Skeleton-based Action Recognition

Construction de ce graphe des relations temporelles

Matrice d’adjacence avec
corrélations entre frame
2 technique :

Temporal Graph Modeling for Skeleton-based Action Recognition

Techniques de calcul de corrélations :

calculatoire vs learned.

Accompagné d’un mécanisme multi-tête

Temporal Graph Modeling for Skeleton-based Action Recognition

Origine

$$f_{out}^{s} = \sum_{k}^{K_s} W_k f_{in} (\Lambda_k^{-\frac{1}{2}} \tilde{\mathrm{A}}_{\mathrm{k}} \Lambda_k^{-\frac{1}{2}} \odot M_k),$$

Arrivé (origine + matrice des relations temporelles + Multi-head-mechanisme)

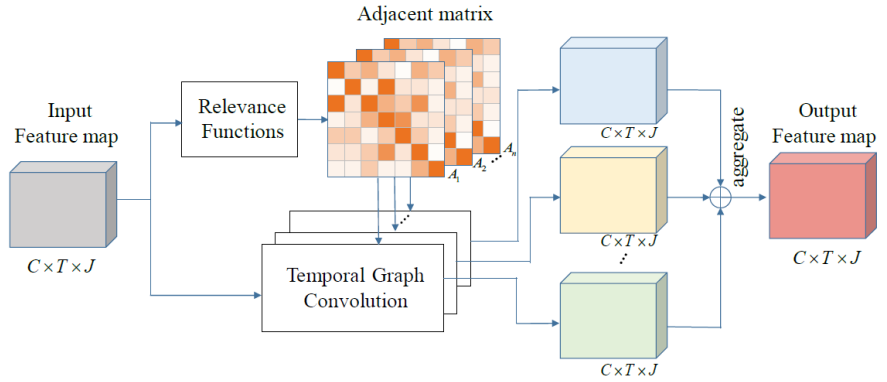


Fig. 3: The illustration of the proposed multi-head temporal graph convolution. After obtaining skeleton feature by graph convolution, we first utilize designed relevance function to obtain multi-head adjacent matrices. Subsequently, the temporal graph convolution is performed on individual temporal relation graphs using corresponding temporal adjacent matrix to extract multi-kinds of temporal relations. Finally, the obtained multi-kinds of features are aggregated by element-wise sum to get temporal enhanced features.

$$f_{out}^{st} = \sum_{n}^{N} \mathrm{A}_t^n \left[\sum_{k}^{K_s} \mathrm{W}_k f_{in} (\Lambda_k^{-\frac{1}{2}} \tilde{\mathrm{A}}_{\mathrm{k}} \Lambda_k^{-\frac{1}{2}} \odot M_k) \right] \mathrm{W}^n$$

Temporal Graph Modeling for Skeleton-based Action Recognition
Jianan Li, Xuemei Xie, Zhifu Zhao, Yuhan Cao, Qingzhe Pan and Guangming Shi Fellow, IEEE

Resultat

TABLE V: Comparisons of the recognition accuracy (%) with the state-of-the-art methods on NTU-60 RGB+D skeleton dataset.

Methods	CS	CV	Conference
VA-LSTM [1]	79.2	87.7	CVPR2017
RotClip+MTCNN [30]	81.1	87.4	TIP2018
ST-GCN [3]	81.5	88.3	AAAI2018
DPRL [31]	83.5	89.8	CVPR2018
SR-TSL [32]	84.8	92.4	ECCV2018
STGR-GCN [10]	86.9	92.3	AAAI2019
AS-GCN [15]	86.8	94.2	CVPR2019
NAS-GCN [33]	89.4	95.7	AAAI2020
MS-AAGCN [34]	90.0	96.2	TIP2020
MS TE-GCN(our)	**90.8**	**96.2**	-

Temporal Graph Modeling for Skeleton-based
Action Recognition
Jianan Li, Xuemei Xie, Zhifu Zhao, Yuhan Cao, Qingzhe Pan and Guangming Shi Fellow, IEEE

Conclusion :

2 articles :

2 façons de se prémunir de la particularité de la convolution sur graphes :

Aspect local ne permettant pas de transmettre des informations entre nœuds distants.

- L’une basé sur la convolution non pas sur les nœud mais sur les canaux de caractéristiques.
- L’autre basé sur la création d’un graphes non trivial basé sur les corrélations entre frames permettant à des frames éloignés de communiquer.

Des questions ?

Merci