25th International Conference on Pattern Recognition

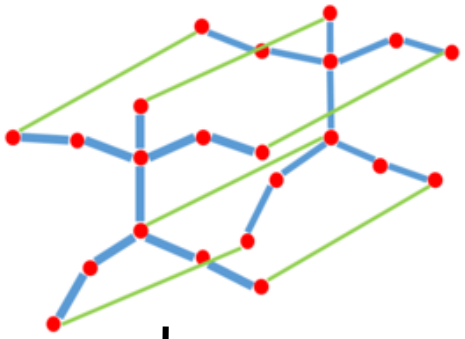
# Vertex Feature Encoding and Hierarchical Temporal Modeling in a Spatio-Temporal Graph Convolutional Network for Action Recognition

**Konstantinos Papadopoulos,** Enjie Ghorbel,
Djamila Aouada, Björn Ottersten

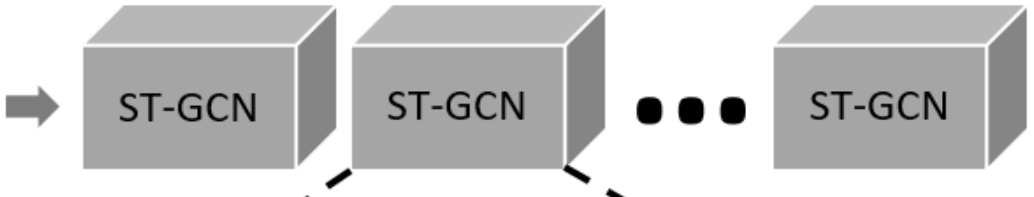
Interdisciplinary Center for Security, Reliability and Trust
University of Luxembourg

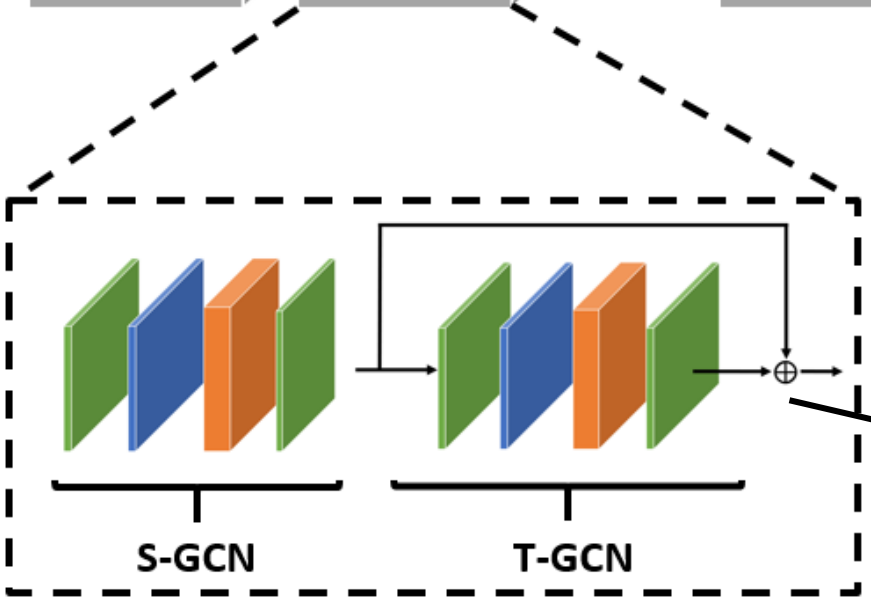
January 13, 2021

# Spatial-Temporal Graph Convolutional Networks



Each graph node $v_i$ has skeleton features $\mathrm{Q}_i$ such that:
$\mathbf{f}_{in}(v_i) = \mathrm{Q}_i$




Simple 2D, single layer convolution

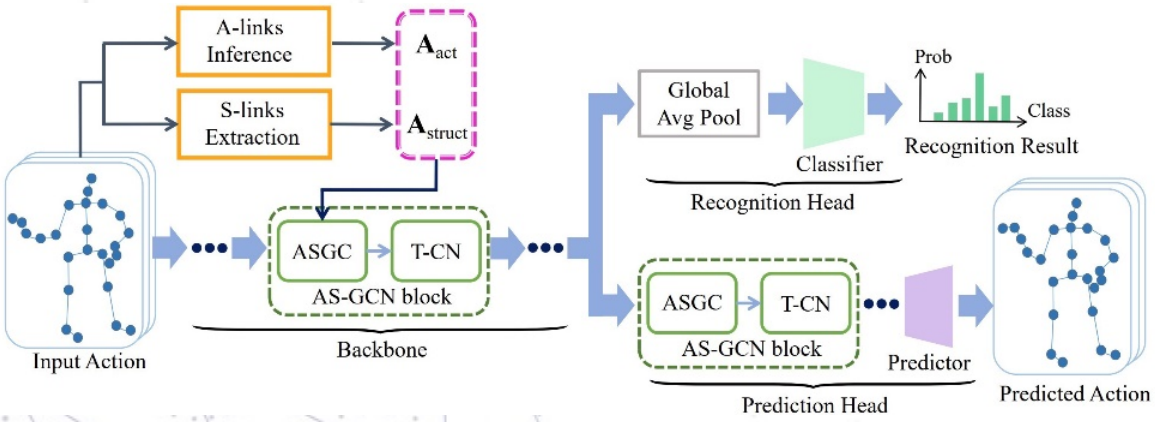
Yan et al. “Spatial temporal graph convolutional networks for skeleton-based action recognition”, AAAI 2018

# Background – ST-Graph Convolutional Networks

ST-GCN variants

learn more appropriate skeleton joint links
(**AS-GCN**[1] , **2S-AGCN**[3])

**AS-GCN**[1]: Encoder-decoder which captures action-specific latent dependencies and structural links.




[1]: Li, et al., "Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition", CVPR 2019
[2]: Si, et al. "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition", CVPR 2019
[3]: Shi, et al. "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition", CVPR 2019

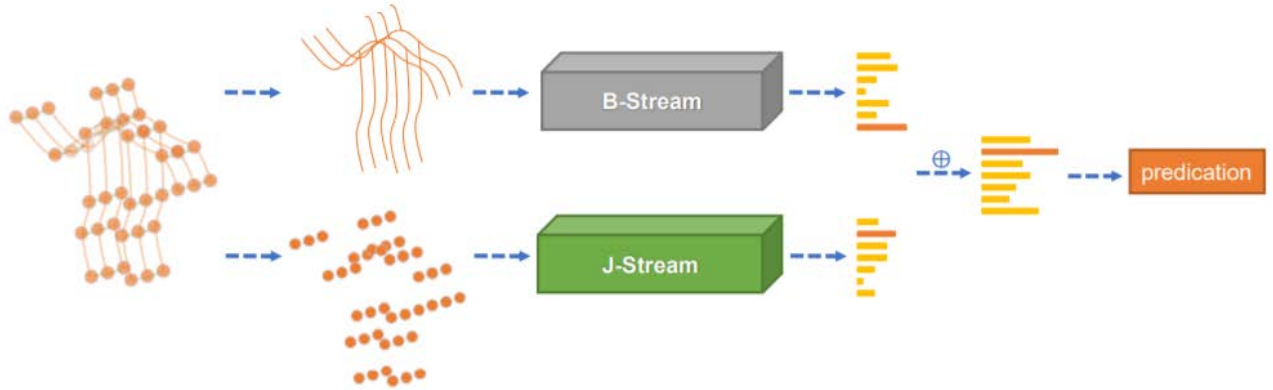
# Background – ST-Graph Convolutional Networks

ST-GCN variants

learn more appropriate skeleton joint links
(**AS-GCN**[1] , **2S-AGCN**[3])

Use more features
(**2S-AGCN**[3])

**2S-AGCN**[3]: Skeleton joint coordinates and the lengths and directions of bones are used as input features.




What about:

- Suitability of raw features?
- Model compactness?

[1]: Li, et al., “Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition”, CVPR 2019
[2]: Si, et al. “An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition”, CVPR 2019
[3]: Shi, et al. “Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition”, CVPR 2019

# Background – ST-Graph Convolutional Networks

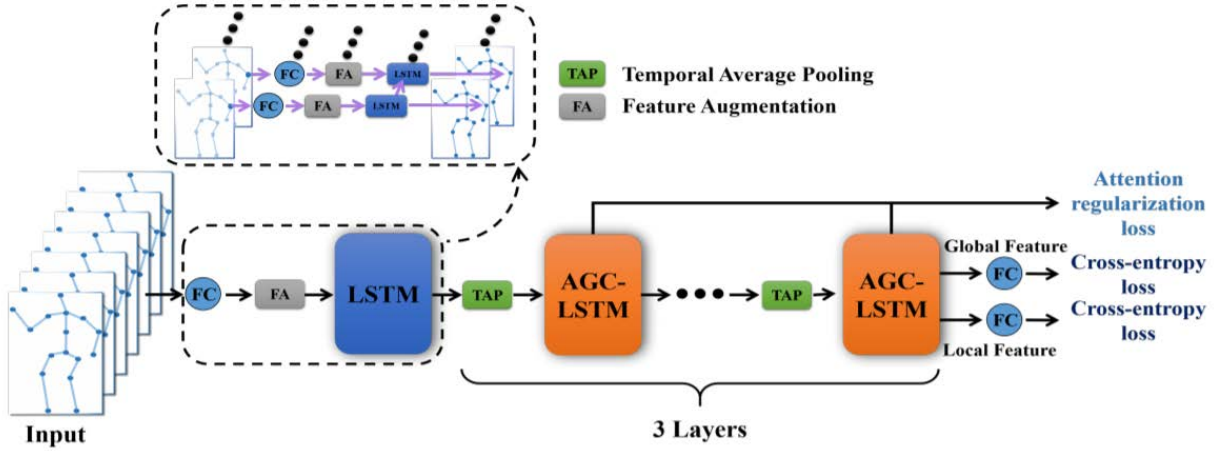
ST-GCN variants

- learn more appropriate skeleton joint links (**AS-GCN**[1] , **2S-AGCN**[3])
- Use more features (**2S-AGCN**[3])
- Model short-term and long-term dependencies (**AGC-LSTM**[3])

What about:

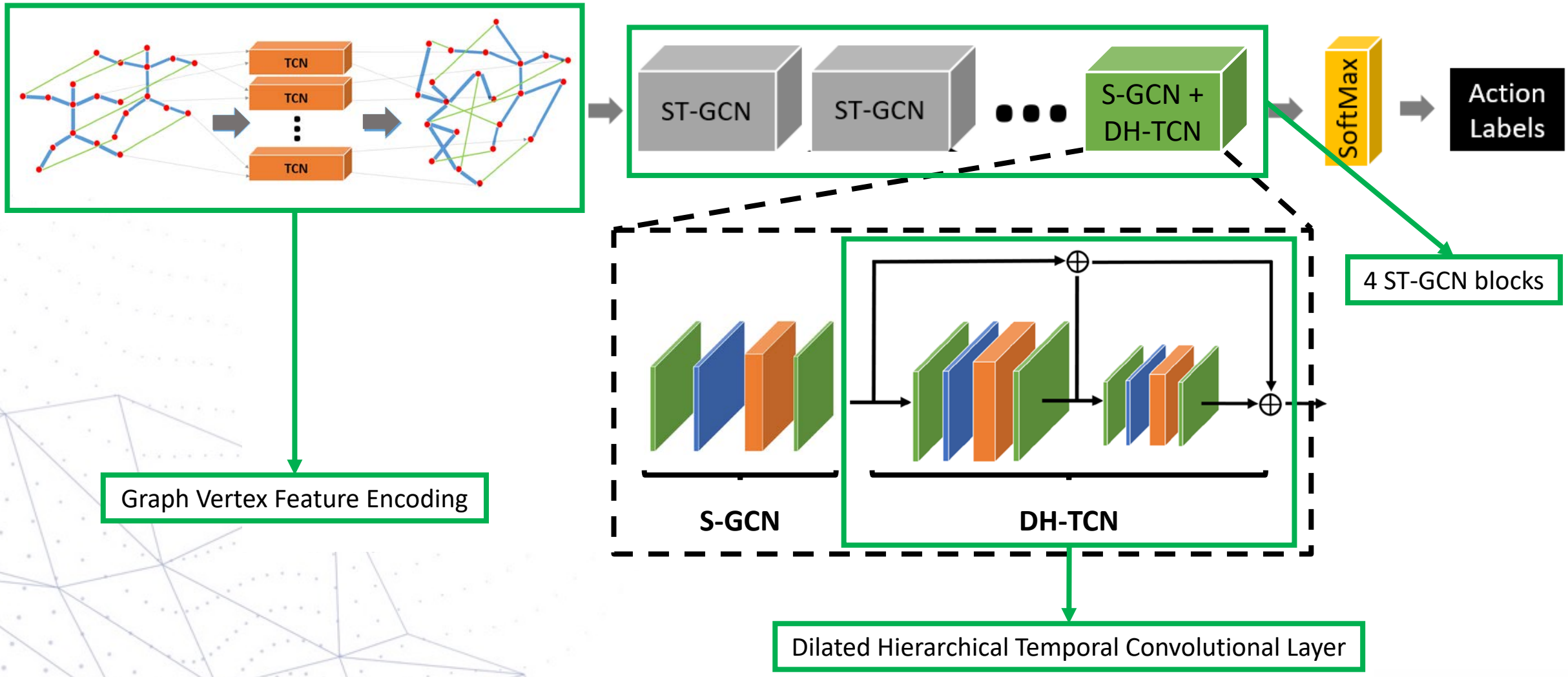
- Suitability of raw features?
- Model compactness?
- Modeling temporal dependencies in an efficient and effective way?

**AGC-LSTM**[2]: LSTM extension to graph convolutions.

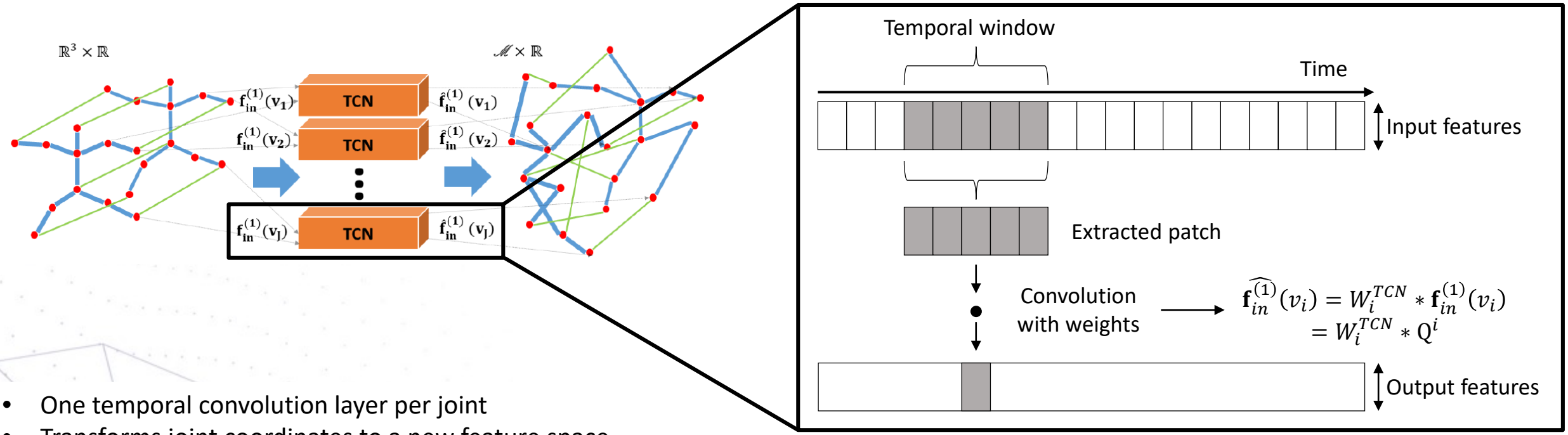



[1]: Li, et al., "Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition", CVPR 2019
[2]: Si, et al. "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition", CVPR 2019
[3]: Shi, et al. "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition", CVPR 2019

# Proposed Approach – Overview

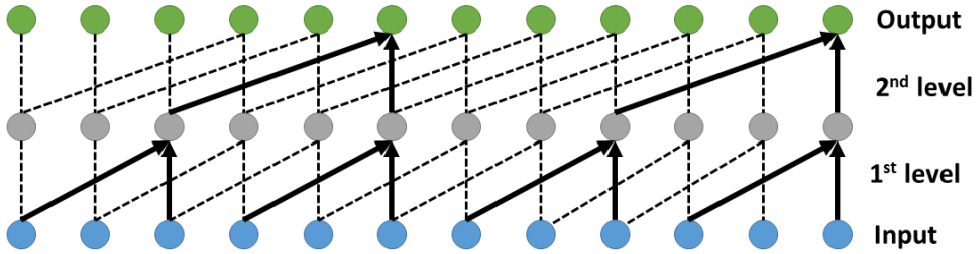


# Graph Vertex Feature Encoding

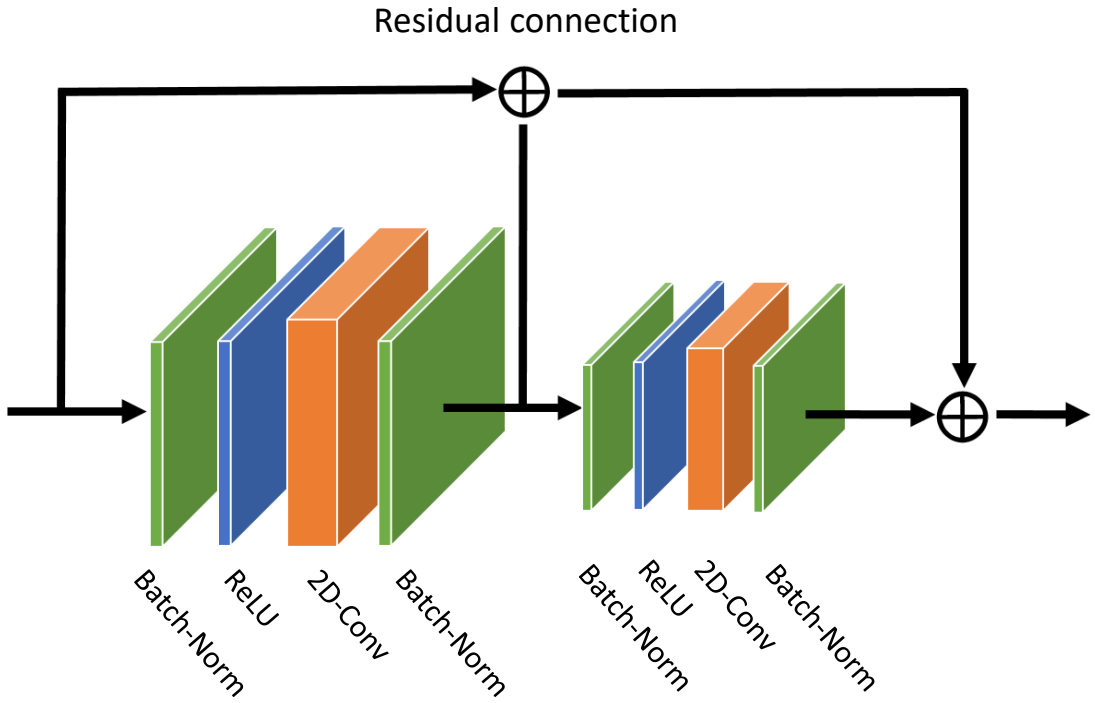



- One temporal convolution layer per joint
- Transforms joint coordinates to a new feature space
- Skeleton joints $\mathrm{Q}^i$ are given as input in vertices $v_i$ at $l = 1$.
- Trained in an end-to-end manner with the network.
- The graph structure is preserved.

# Dilated Hierarchical Temporal Convolutional Layer




Dilated convolutions are efficient in modeling long-term dependencies while at the same time they maintain efficiency.




- DH-TCN models both short-term and long-term dependencies.
- Placed at the last ST-GCN block.
- $\mathbf{f}_{temp}^{(k,n)} = f(\mathbf{W}_i^{DH} *_{\lambda} \mathbf{f}_{temp}^{(k,n-1)})$, with $\mathbf{f}_{temp}^{(k,0)} = \mathbf{f}_{out}^{(k)}$
- $\mathbf{f}_{out}^{(k)}$: output feature map from the $k^{\text{th}}$ (S-GCN) block
- $*_{\lambda}$: convolution with dilation $\lambda = 2^{n-1}$

# Results

Table 1: Accuracy of recognition (%) on NTU-60, NTU-120 and Kinetics datasets.

| Method | NTU-60 | | NTU-120 | | Kinetics | |
|---|---|---|---|---|---|---|
| | X-subject | X-view | X-subject | X-setup | Top-1 | Top-5 |
| SkeleMotion | 76.5 | 84.7 | 67.7 | 66.9 | - | - |
| Body Pose Evolution Map | **91.7** | **95.3** | 64.6 | 66.9 | - | - |
| Multi-Task CNN with RotClips | 81.1 | 87.4 | 62.2 | 61.8 | - | - |
| Two-Stream Attention LSTM | 76.1 | 84.0 | 61.2 | 63.3 | - | - |
| Skeleton Visualization (Single Stream) | 80.0 | 87.2 | 60.3 | 63.2 | - | - |
| Multi-Task Learning Network | 79.6 | 84.8 | 58.4 | 57.9 | - | - |
| ST-GCN (10 blocks) | 81.5 | 88.3 | 72.4* | 71.3* | 30.7 | 52.8 |
| **GVFE + ST-GCN w/ DH-TCN (4 blocks - ours)** | 79.6 | 88.0 | 72.3 | 71.7 | 29.0 | 50.9 |
| AS-GCN (10 blocks) | 86.8 | 94.2 | 77.7* | 78.9* | **34.8** | **56.5** |
| **GVFE + AS-GCN w/ DH-TCN (4 blocks - ours)** | 86.4 | 92.9 | **79.2** | **81.2** | - | - |

**Reduced number of ST-GCNs – Almost similar performance**

Table 2: Ablation study: accuracy of recognition (%) on NTU-120 dataset for cross-setup settings using ST-GCN as a baseline.

| Method | Accuracy (%) |
|---|---|
| ST-GCN (4 blocks) | 51.8* |
| GVFE + ST-GCN (4 blocks) | 70.9 |
| ST-GCN w/ DH-TCN (4 blocks) | 68.3 |
| **GVFE + ST-GCN w/ DH-TCN (4 blocks - ours)** | **71.7** |