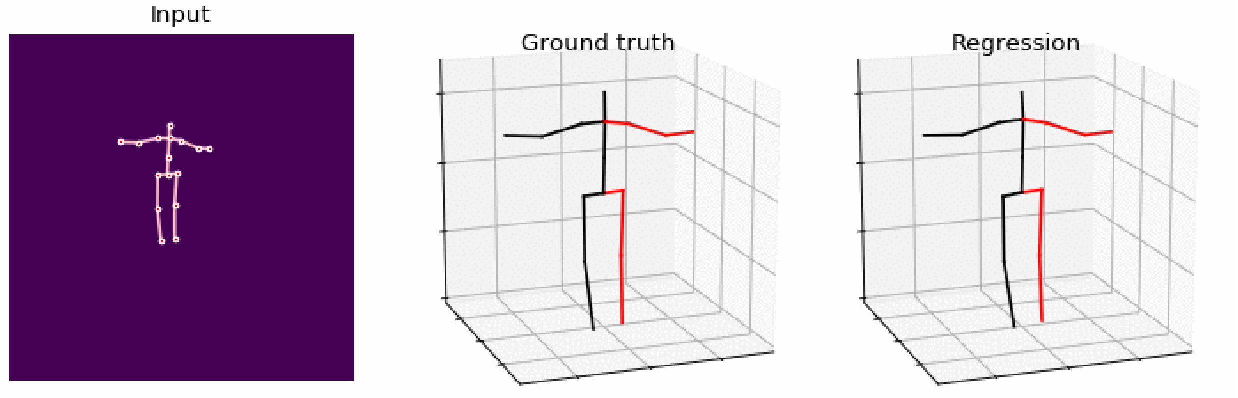


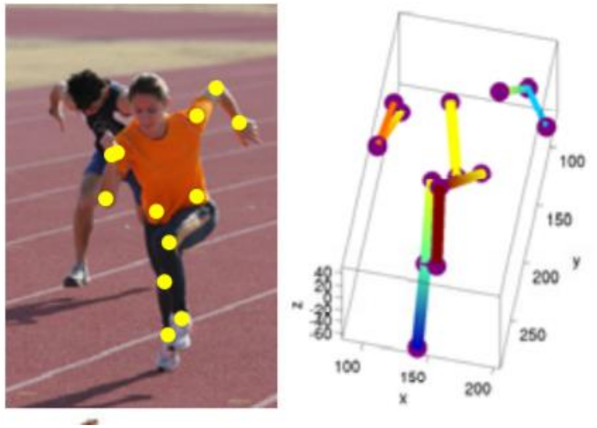
Semantic Graph Convolutional Networks for 3D Human Pose Regression

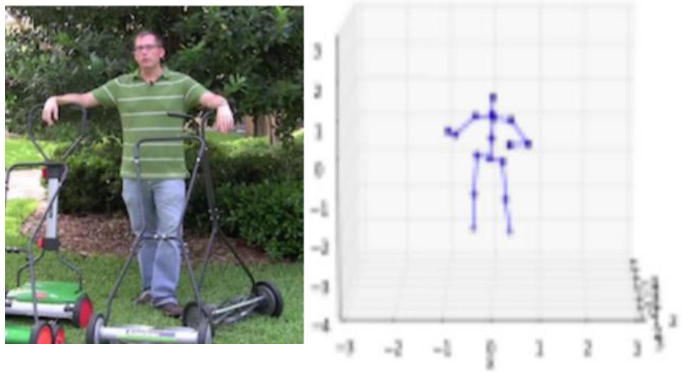
Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, Dimitris N. Metaxas
Rutgers University, Binghamton University (2019)

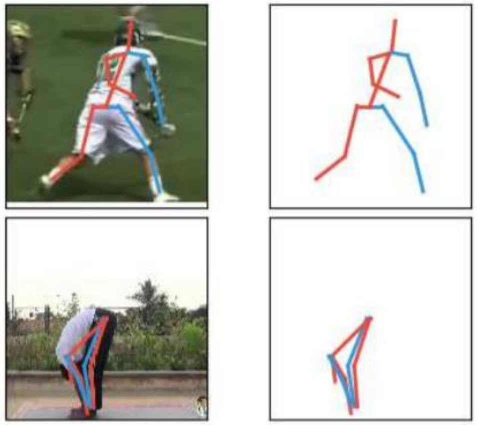
Presented by Cédric Siboyabasore

Regression problem:
Given a 2D human pose and the optional image, the goal is to predict the locations of its corresponding 3D joints in a certain coordinate space

Pose estimation: existing methods







- Nearest neighbor: H. Jiang used a large database of poses to resolve ambiguities based on nearest neighbor queries.

3d human pose reconstruction using millions of Exemplars (2015)

- Deep learning: Xingyi Zhou et al directely predicted 3D pose from the image

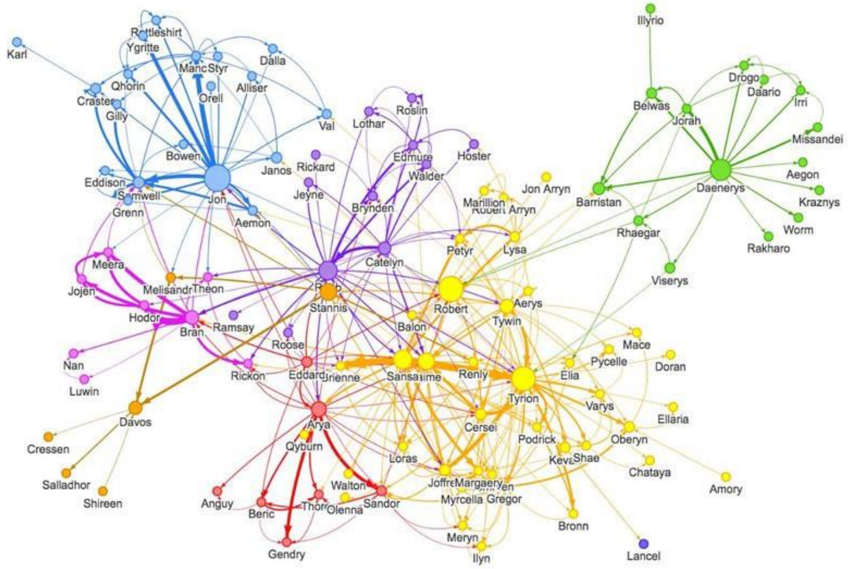
Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach (2017)

- **Deep learning**: Martinez et al, proved that 2D estimation is crucial for 3D pose estimation. Predicted 3D key points purely based on 2D detections

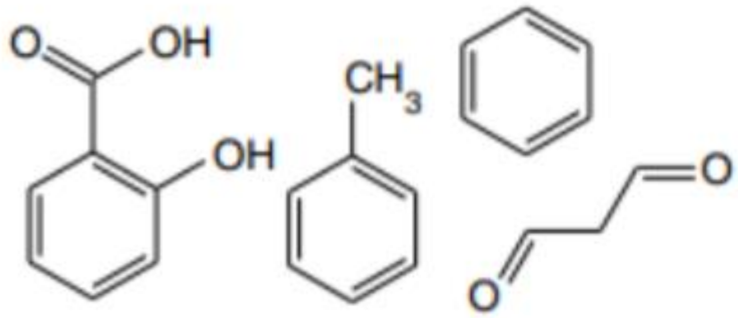
A simple yet effective baseline for 3d human pose Estimation (2017)

In this paper, the authors leverage graph representations

In several applications, information is naturally represented by graphs



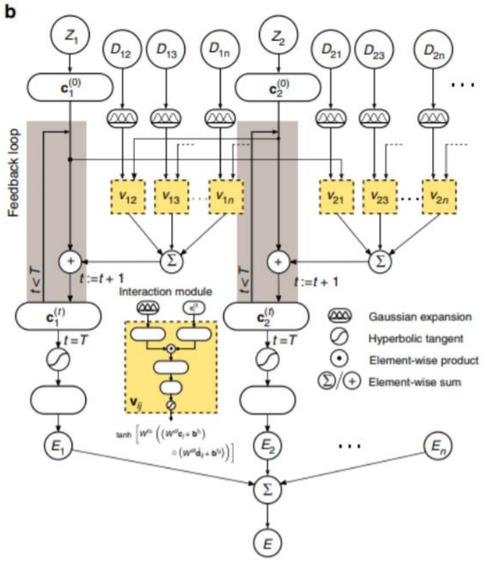
- Social networks: nodes represent users while edges represent the relationship between users



- Molecular structures: atoms and molecules, bounding between atoms and the energy tied to some particular geometry of the molecules can be represented by a graph

Graph Neural Networks (GNNs)

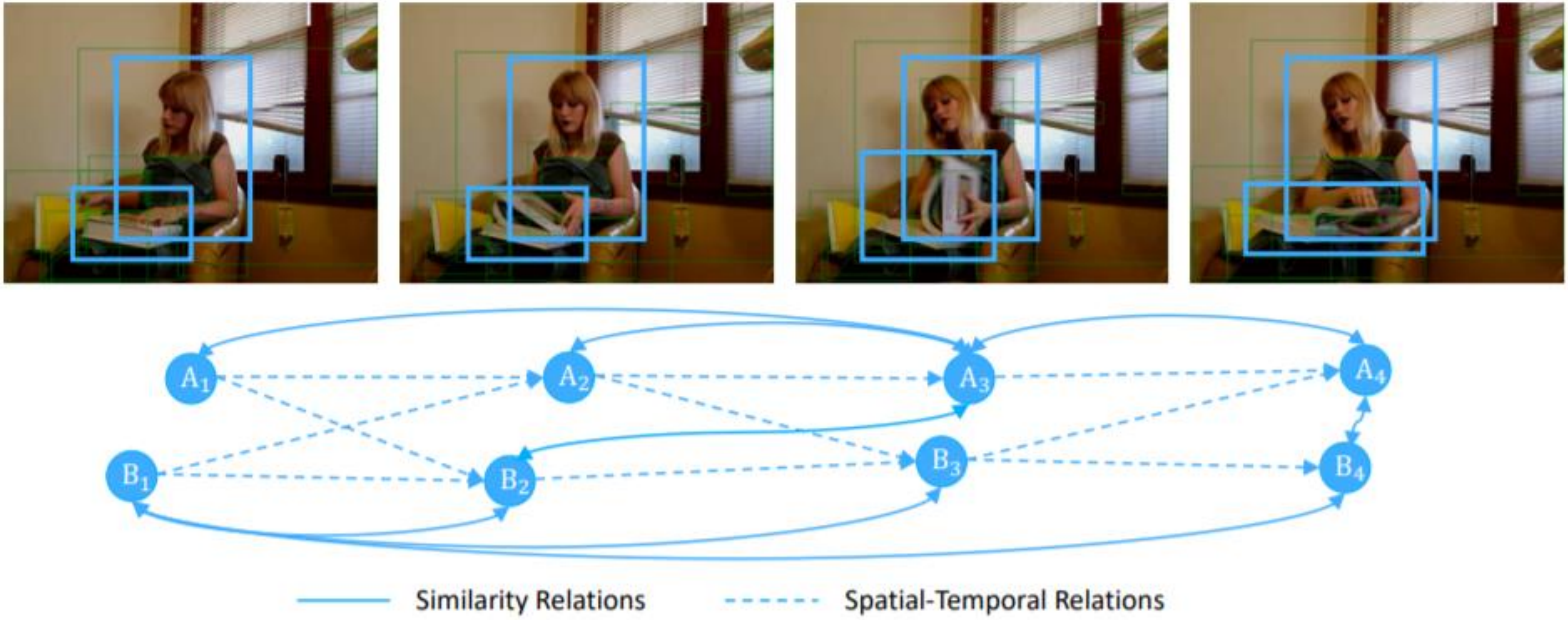
- First presented in 2005 A New Model for Learning in Graph Domains
- Deal with arbitrarily graph-structured data



Quantum-chemical insights from deep tensor neural networks

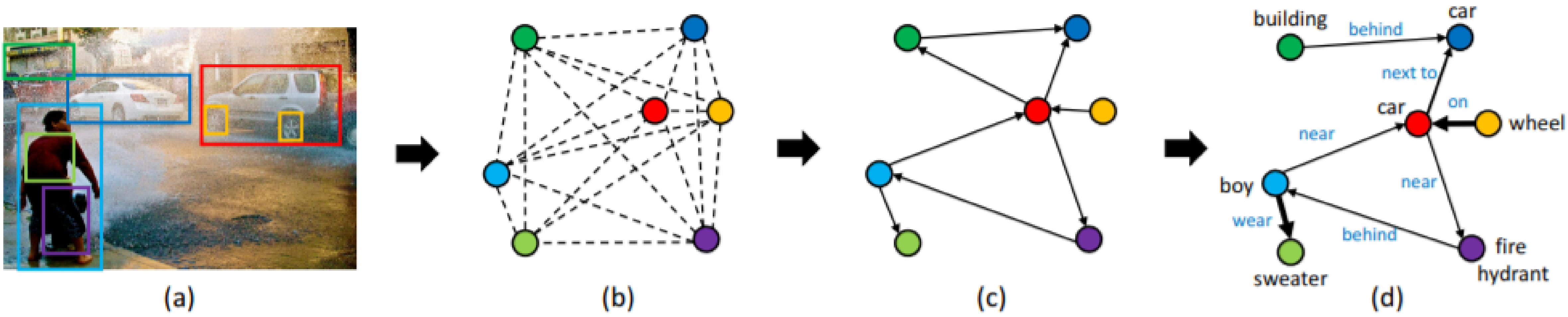
- Graph Convolutional Networks (GCNs)
  - Achieved state-of-the-art results on computer vision tasks

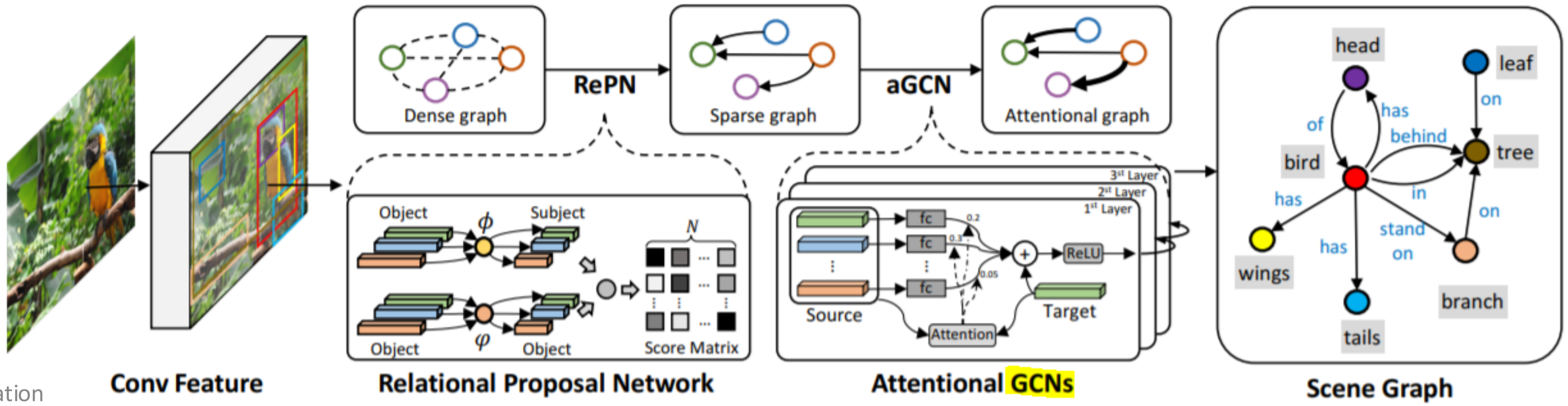
GCNS: state-of-the-art in modeling of relations among temporal sequences



Videos as Space-Time Region Graphs

GCNS: state-of-the-art in modeling of relations among visual objects



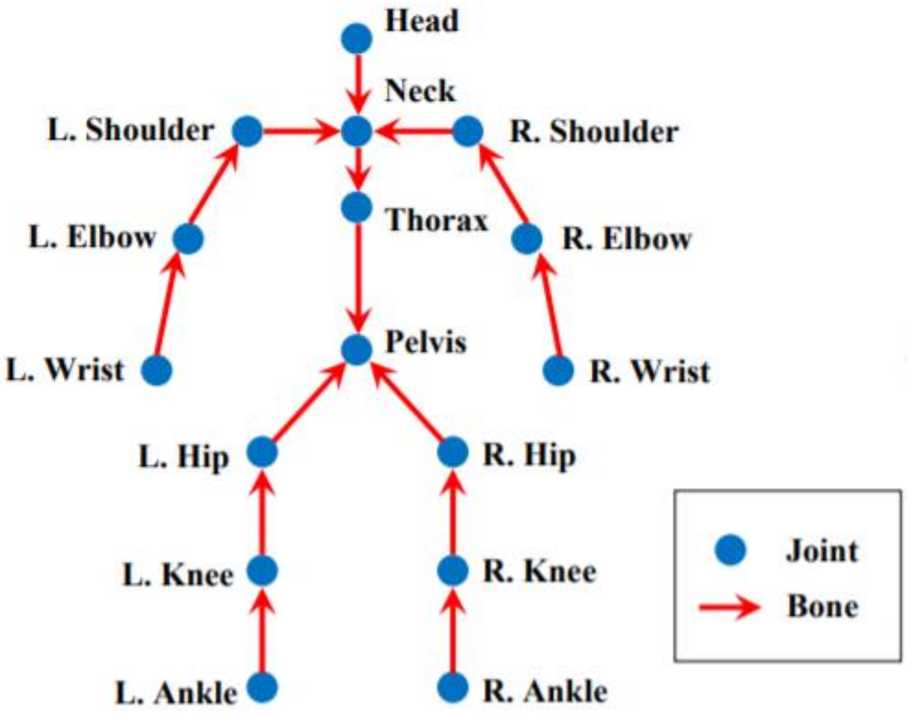


Graph R-CNN for Scene Graph Generation

The human body is a graph!

- Nodes = joints
- Edges = bones

=> The authors of our paper aim at leveraging GCNs



ResGCN : Vanilla GCNs

- The K nodes of the graph are represented in a matrix $\mathbf{X}^{(l)} \in \mathbb{R}^{D_l \times K}$
- Nodes representations are transformed by parameter matrix $\mathbf{W} \in \mathbb{R}^{D_{l+1} \times D_l}$
- The graph convolution operation is written as

$$\mathbf{X}^{(l+1)} = \sigma\left(\mathbf{W}\mathbf{X}^{(l)}\tilde{\mathbf{A}}\right)$$

where $\tilde{\mathbf{A}}$ is the adjacency matrix, $\tilde{a}_{ij} = 1$ if i and j are neighbors, 0 otherwise

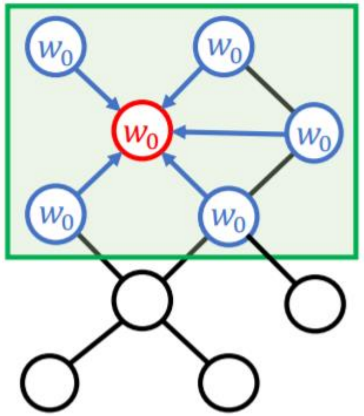
and σ is a non-linear activation function

ResGCN : Vanilla GCNs

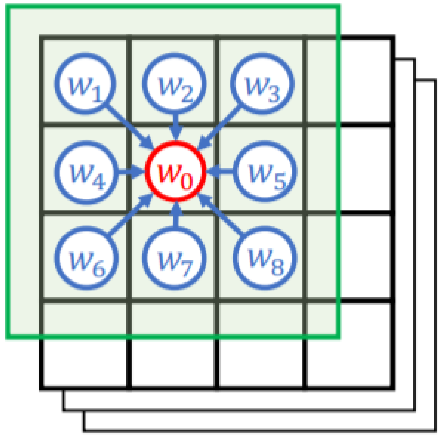
$$\mathbf{X}^{(l+1)} = \sigma\left(\mathbf{W}\mathbf{X}^{(l)}\tilde{\mathbf{A}}\right)$$

Limitations

- Matrix W is shared for all edges => relationships between neighboring nodes is not well exploited



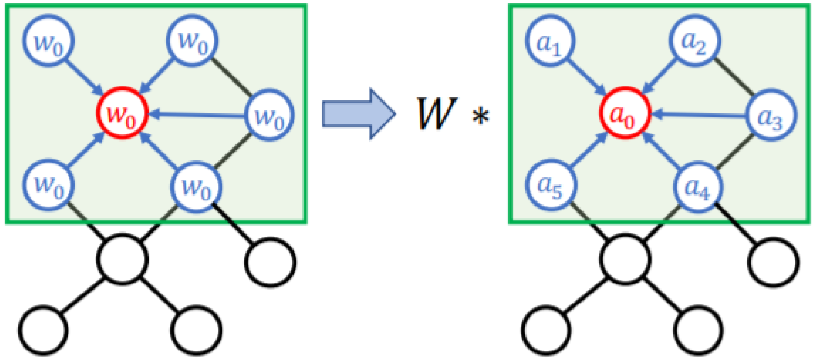
(a) Graph convolutional layer



(b) Convolutional layer

- In previous papers, receptive field is fixed to 1 => only first-order neighbors of each node are taken into account

Semantic Graph Convolution (SemGConv)



Idea to exploit neighboring nodes relationships

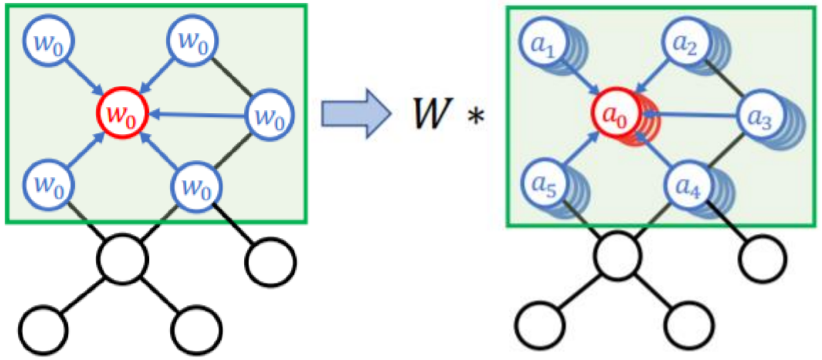
- Graph convolution can be decomposed as learning a weight vector a_i for each node and combining them with a shared matrix W
- The semantic graph convolution operation is written as

$$\mathbf{X}^{(l+1)} = \sigma\Big(\mathbf{W}\mathbf{X}^{(l)}\rho_i(\mathbf{M} \odot \mathbf{A})\Big)$$

where $\mathbf{M} \in \mathbb{R}^{K \times K}$ is a learnable weighting matrix; ρ_i is a Softmax non-linearity

and A enforces that for each node, only the weights of its neighboring nodes are computed

Semantic Graph Convolution (SemGConv)



Idea to exploit neighboring nodes relationships

- Graph convolution can be decomposed as learning a weight vector a_i for each node and combining them with a shared matrix W
- The semantic graph convolution operation is written as

$$\mathbf{X}^{(l+1)} = \Big\Vert_{d=1}^{D_{l+1}} \sigma\left(\vec{\boldsymbol{w}}_d \mathbf{X}^{(l)} \rho_i (\mathbf{M}_d \odot \mathbf{A})\right)$$

where $\mathbf{M}_d \in \mathbb{R}^{K \times K}$ is a learnable weighting matrix for channel d; w_d is the d-th row of transformation matrix W; ρ_i and A same as before

Semantic Graph Convolution Network(SemGCN)

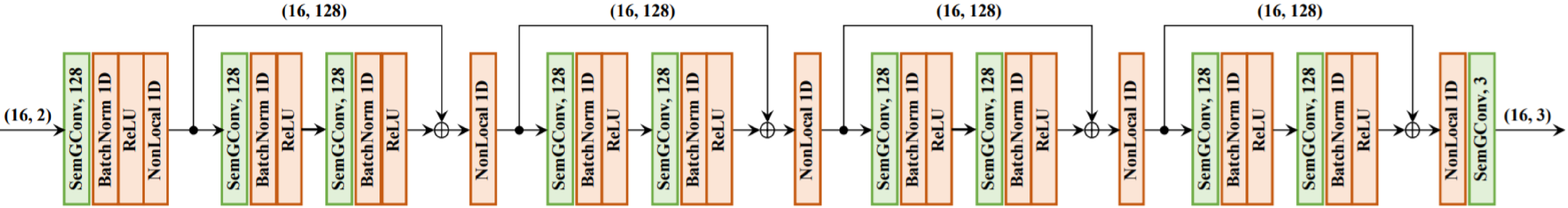
Idea to address problem of limited receptive field

Graph convolution layers are followed by a weighted sum of the learned representations to capture global and long-range relationships among nodes in the graph (Non-local layer)

$$\vec{\boldsymbol{x}}_i^{(l+1)} = \vec{\boldsymbol{x}}_i^{(l)} + \frac{W_x}{K} \sum_{j=1}^{K} f(\vec{\boldsymbol{x}}_i^{(l)}, \vec{\boldsymbol{x}}_j^{(l)}) \cdot g(\vec{\boldsymbol{x}}_j^{(l)})$$

Semantic Graph Convolution Network (SemGCN)

Final architecture



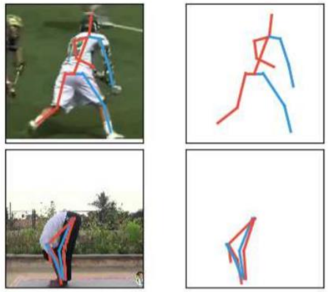
3D Human Pose Regression

- Martinez et al predicted 3D pose estimation using only 2D joints,

From a dataset of N pairs of 2D joints $\mathbf{P} \in \mathbb{R}^{K\times 2}$ and corresponding 3D joints $\mathbf{J} \in \mathbb{R}^{K\times 3}$,

learn

$$\mathcal{F}^* = \underset{\mathcal{F}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathcal{F}(\mathbf{P}_i), \mathbf{J}_i)$$



What our paper proposes:

Learn

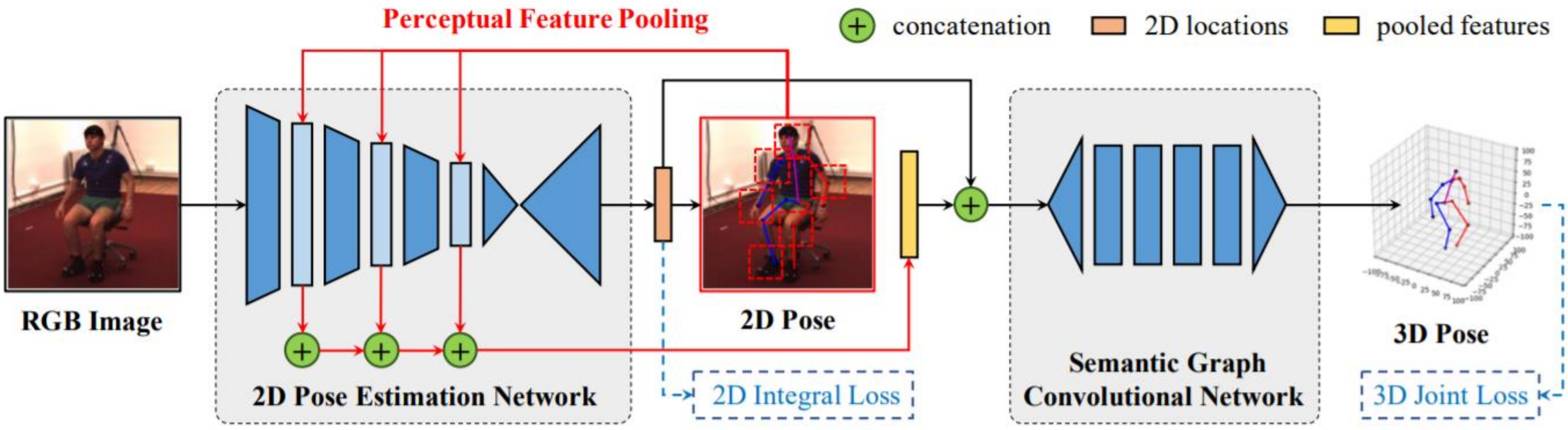
$$\mathcal{F}^* = \underset{\mathcal{F}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathcal{F}(\mathbf{P}_i | I_i), \mathbf{J}_i)$$

=>2D joints can be detected in the image I_i instead of being given as input

3D Human Pose Regression

Final architecture: 2 neural networks

- 1) ResNet for 2D pose estimation. Input : RGB Image. Output: 2D joint locations and image features
- 2) Semantic Graph Convolution Network for 3D pose estimation. Input: 2D joint locations and image features. Output: 3D pose estimation



3D Human Pose Regression

Loss function = sum of losses of two previous papers

$$\mathcal{L}(\mathcal{B}, \mathcal{J}) = \underbrace{\sum_{i=1}^{M} ||\tilde{\mathbf{B}}_i - \mathbf{B}_i||^2}_{\text{bone vectors}} + \underbrace{\sum_{i=1}^{K} ||\tilde{\mathbf{J}}_i - \mathbf{J}_i||^2}_{\text{joint positions}},$$

where the bones are computed from the predicted 3D joints positions

=>This loss = Novelty of this paper

Compositional human pose regression

A simple yet effective baseline for 3d human pose estimation

Novelties of this paper

- Semantic Graph Convolution operation, which exploits neighboring nodes relationship
- Semantic Graph Convolutional network
- Estimate 3D joint locations directly from an image, by estimating 2D joints and taking into account image content
- New loss taking into account bones and joints positions

Thank you!