Semantic Graph Convolutional Networks for 3D Human Pose Regression

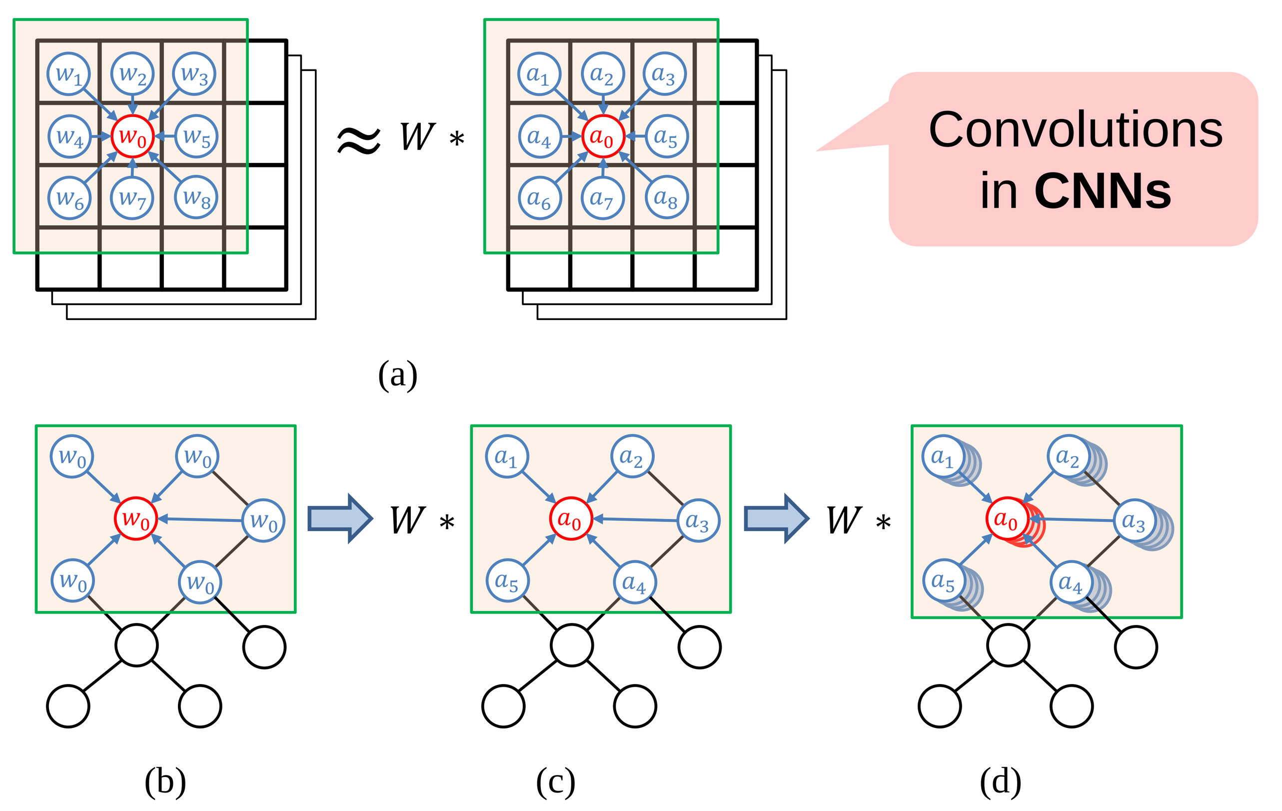
Long Zhao[1] Xi Peng[2] Yu Tian[1] Mubbasir Kapadia[1] Dimitris Metaxas[1]
[1] Rutgers University [2] Binghamton University

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Highlights

- We study the problem of learning Graph Convolutional Networks (GCNs) for **regression**.
- We propose **Semantic Graph Convolutional Networks (SemGCNs)** to address **two limitations** of GCNs.
- We apply SemGCNs to **3D human pose regression**: Both 2D and 3D poses are able to be naturally represented by a canonical skeleton, i.e., a graph.
- The proposed SemGCN outperforms state of the art while using **90% fewer parameters**.

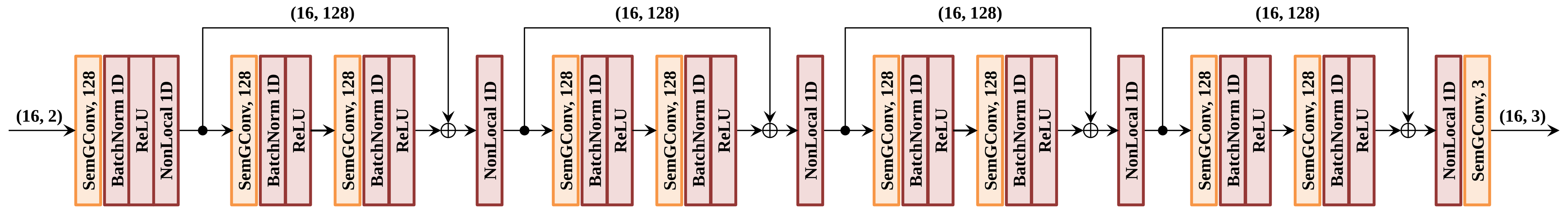
Semantic Graph Convolutions



- A convention kernel of CNNs in (a) can be approximated by learning a weighting vector for each position and a shared transformation matrix.
- Conventional **Graph Convolutions** in (b) only learn a shared transformation matrix for all nodes.
- The approximated formulation in (a) is extended to (b) to build **Semantic Graph Convolutions** as defined in (c).
- We further extend (c) to learn a channel-wise weighting vector for each node as shown in (d).

Semantic Graph Convolutional Networks

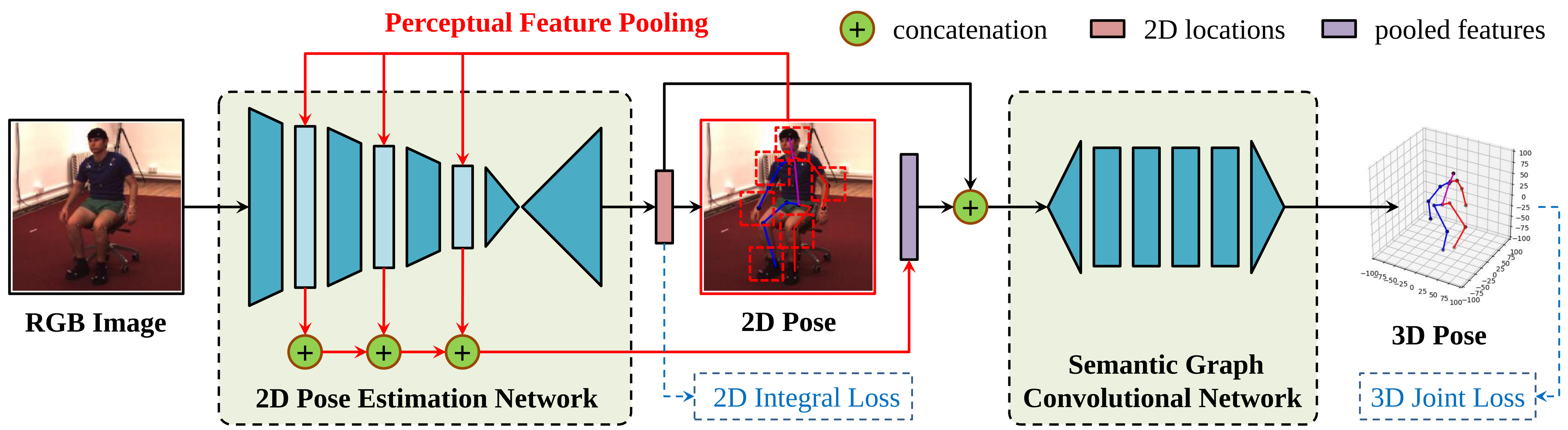
- **Semantic Graph Convolutional Layers** learn local relationships between nodes.
  → to address Limitation #1 of GCNs: shared transformation matrix for each node.
- **Non-local Blocks** capture global and long-range relationships among nodes.
  → to address Limitation #2 of GCNs: small receptive fields of convolution filters.
- Our proposed **Semantic Graph Convolutional Networks** interleave Semantic Graph Convolutional Layers and Non-local Blocks to capture local and global semantic relations of nodes in a graph.



3D Human Pose Regression

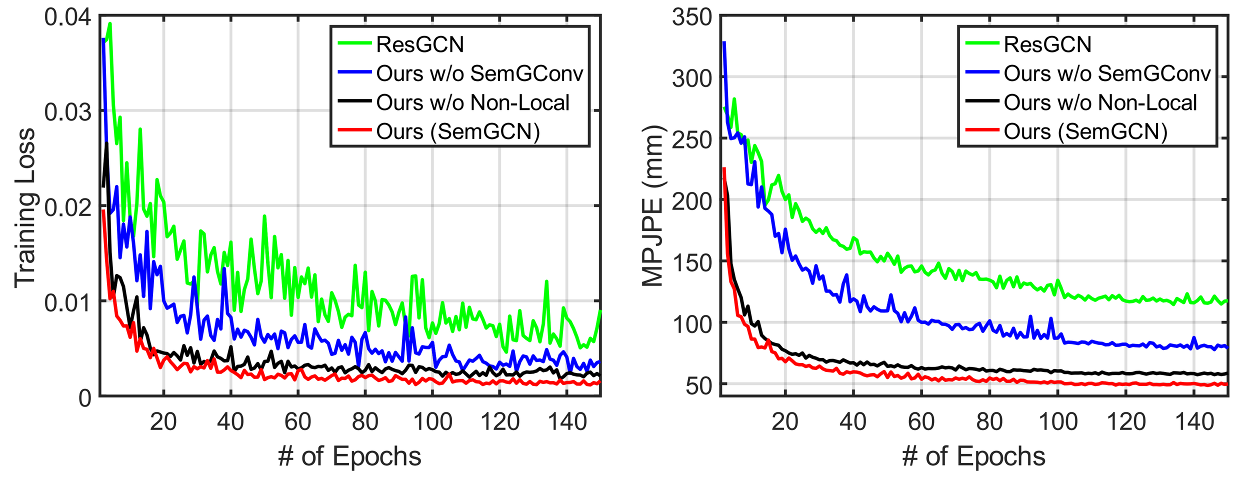
Motivation: We argue that image content is able to offer important cues for solving ambiguous cases of 3D human poses. Therefore, we treat image content as an additional constraint for the human skeleton.

- The pre-trained **2D pose estimation network** encodes the perceptual features of the input image.
- **Perceptual Feature Pooling** is proposed to extract intermediate features from the 2D pose detector.
- Perceptual features are concatenated with the 2D coordinates and fed into the proposed SemGCN.



Results

- **Ablation Study: Training Curves and Testing Errors**



Settings	MPJPE
ResGCN	94.4
Ours w/o SemGConv	65.9
Ours w/o Non-Local	52.5
Ours (SemGCN)	**43.8**

- **Evaluation on 3D Human Pose Regression**

Methods	# of params	MPJPE (GT)
aGCN / GAT	0.16M	82.9
ST-GCN	0.27M	57.4
Martinez et al.	4.29M	45.5
Ours	0.43M	**43.8**

Methods	MPJPE
Martinez et al. (ICCV' 17)	62.9
Yang et al. (CVPR' 18)	58.6
Ours (HG)	60.8
Ours (RN w/ FP)	**57.6**

- **Qualitative Results**