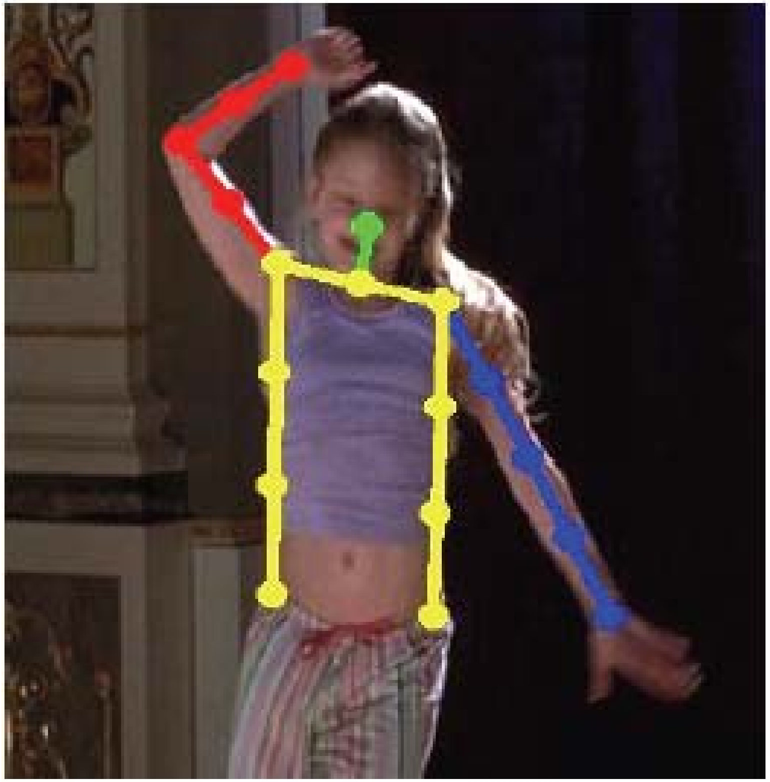
香港中文大學
The Chinese University of Hong Kong

Structured Feature Learning for Human Pose Estimation

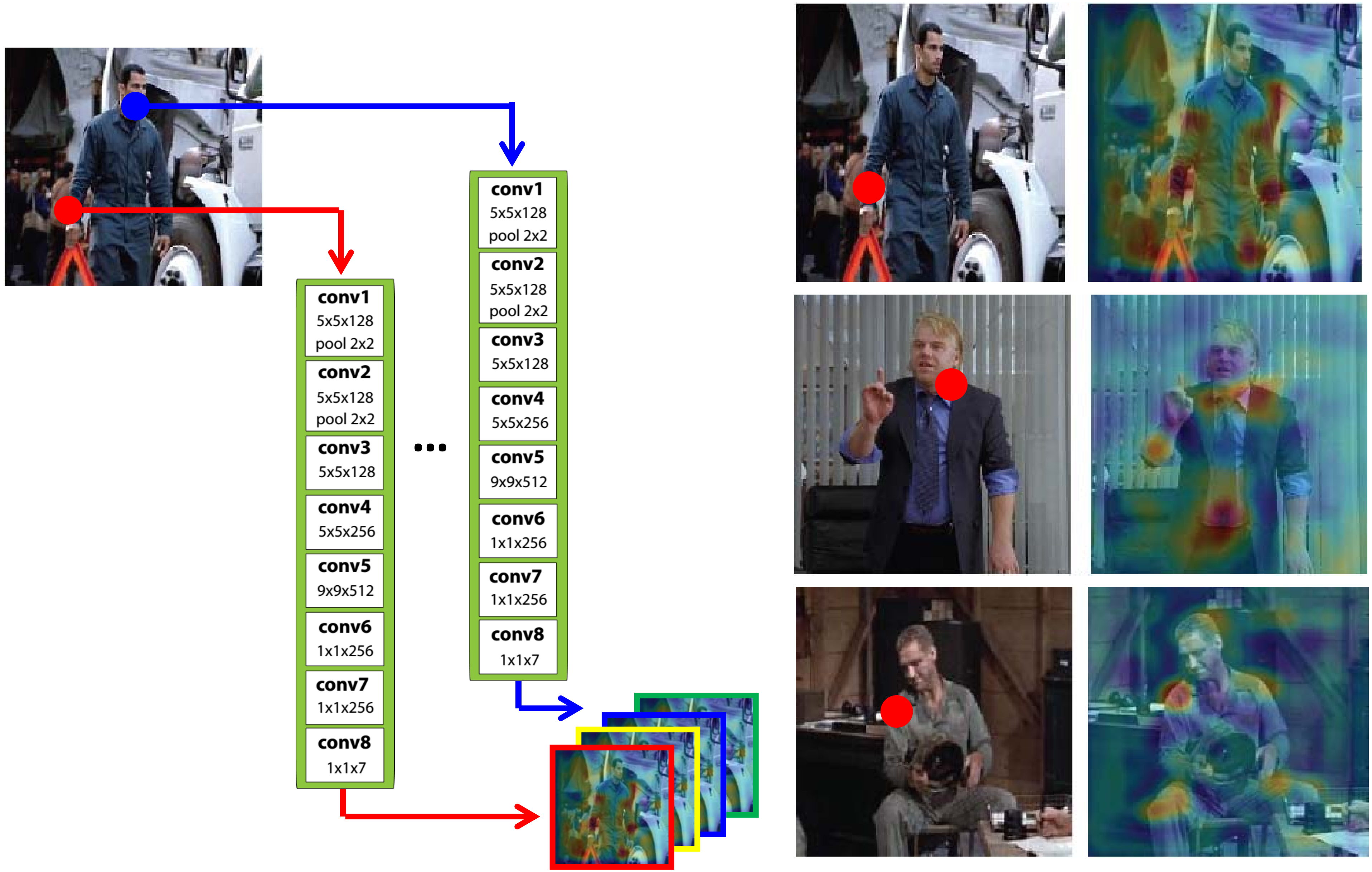
Xiaogang Wang
Department of Electronic Engineering,
The Chinese University of Hong Kong

Human Pose Estimation

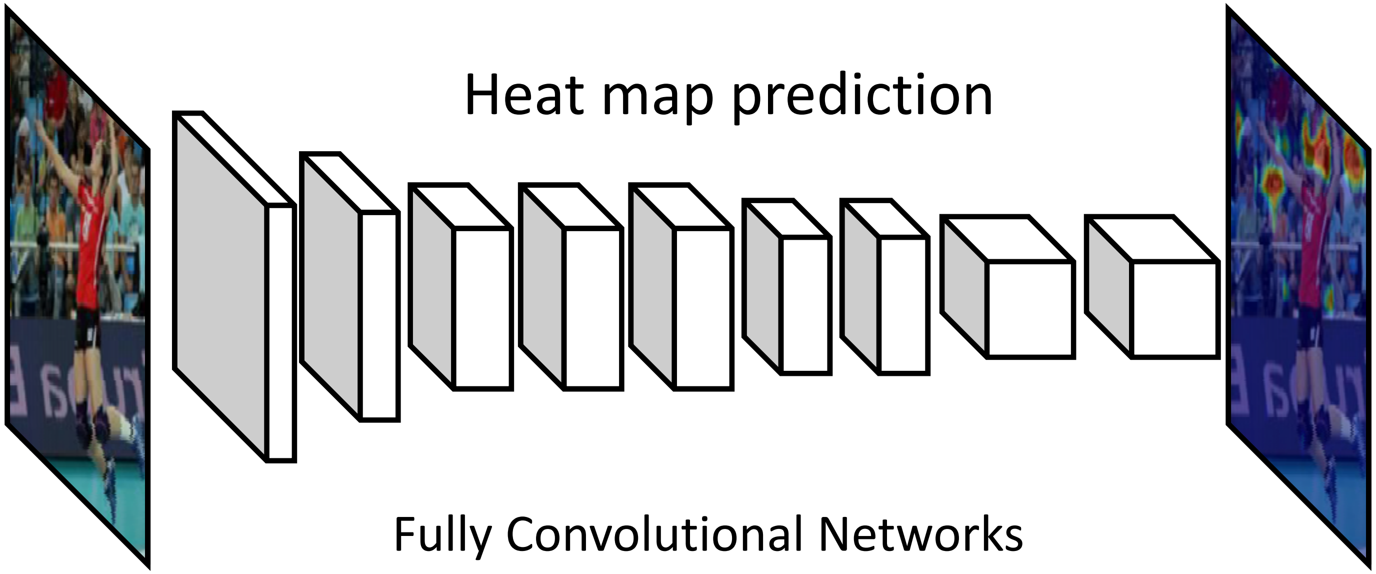
 

Pose estimation result generated by our deep learning algorithm

Using CNN to localize individual joints separately is not reliable

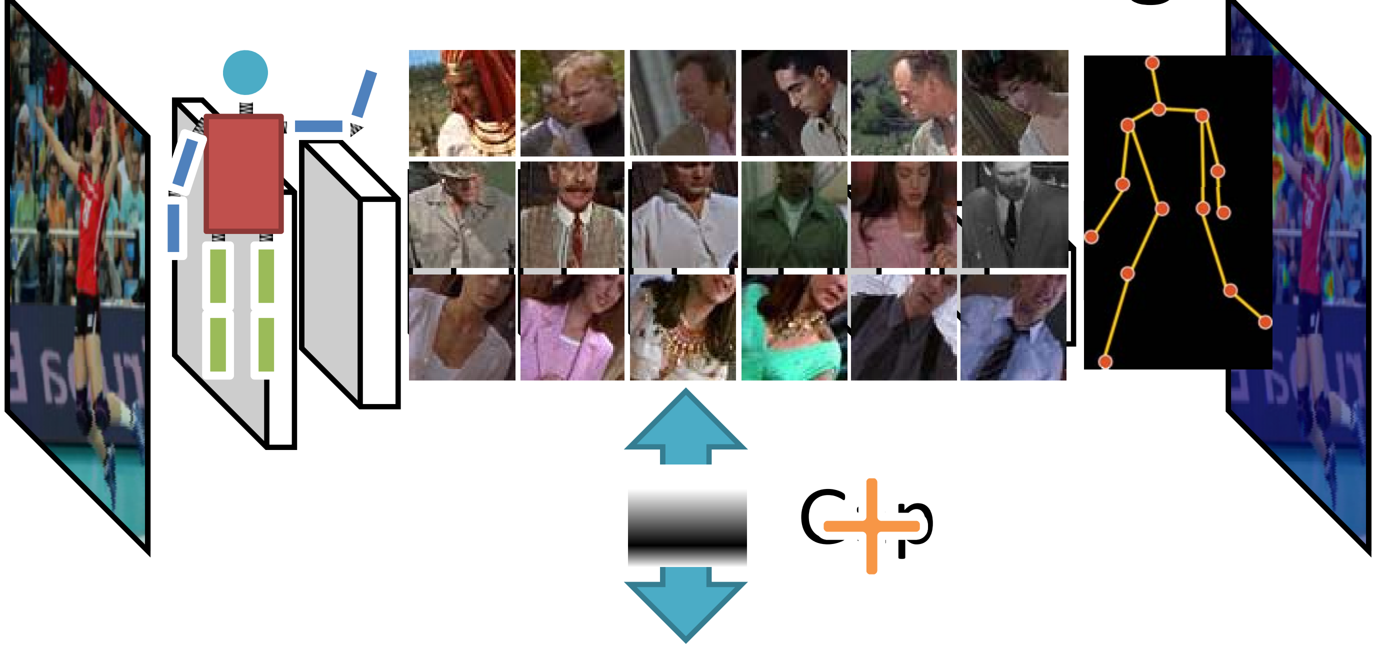


Deep Learning Based Methods

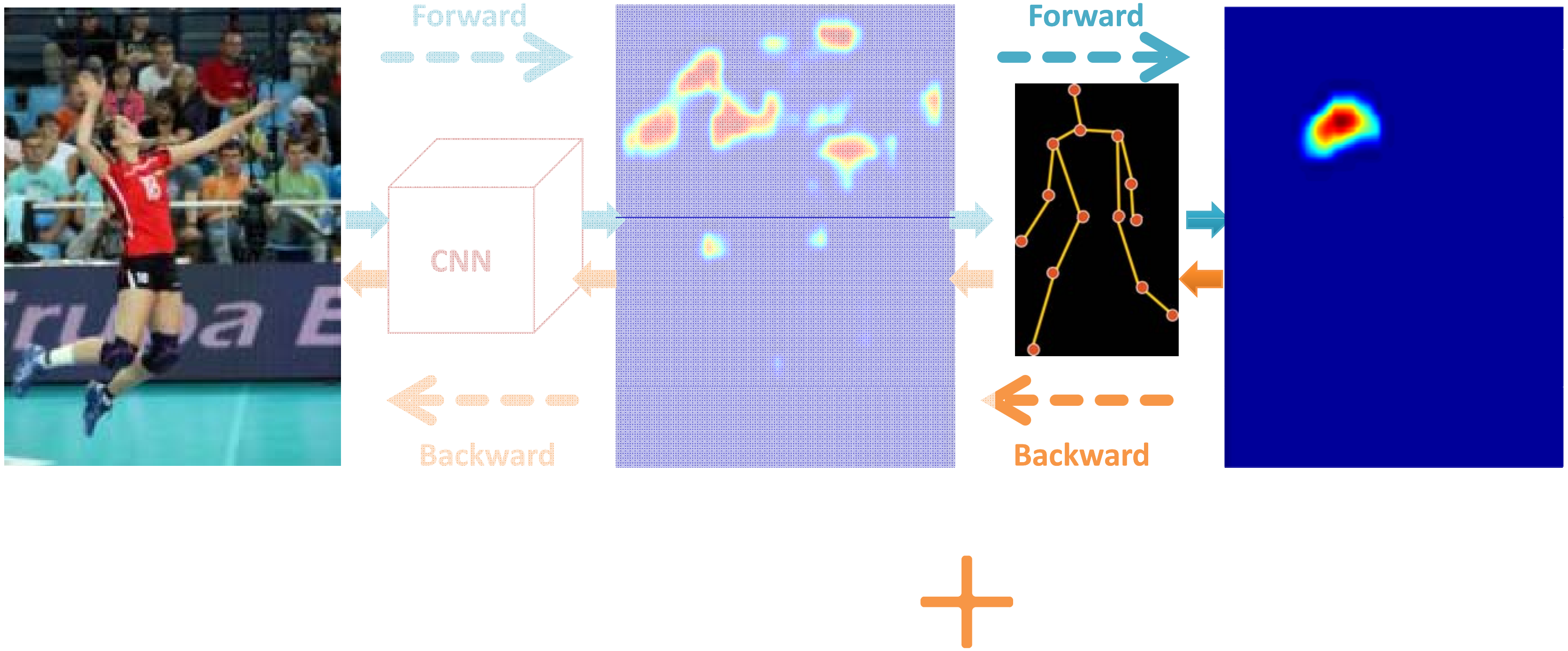


- ✓ Benefits from better neural network architectures
  - VGG
  - GoogLeNet,
  - ResNet
- ✗ Structures are used for post processing

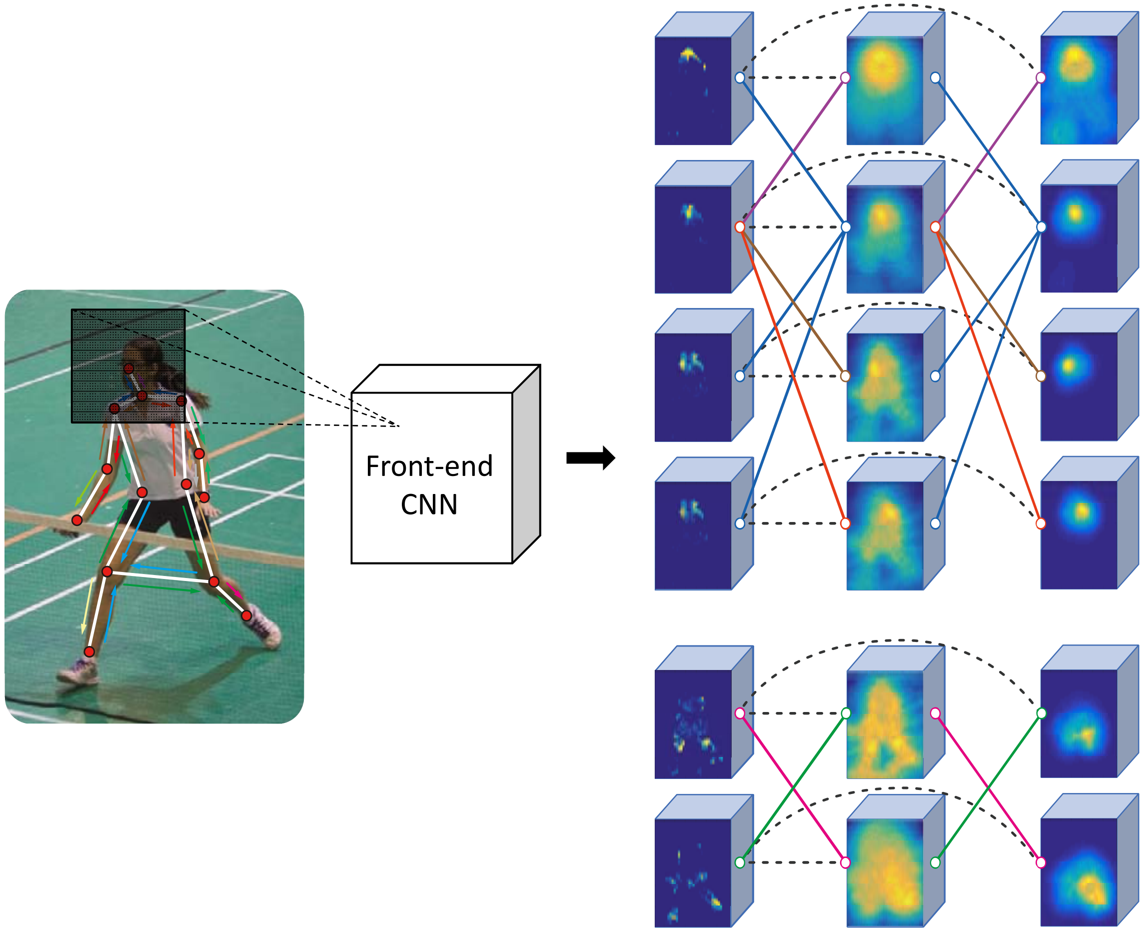
Gap Between Deep Models and Structure Modeling



Motivation: Geometric Constraints Among Body Parts Helps in Learning Better Representation



Model structures on score maps and jointly learn feature representations

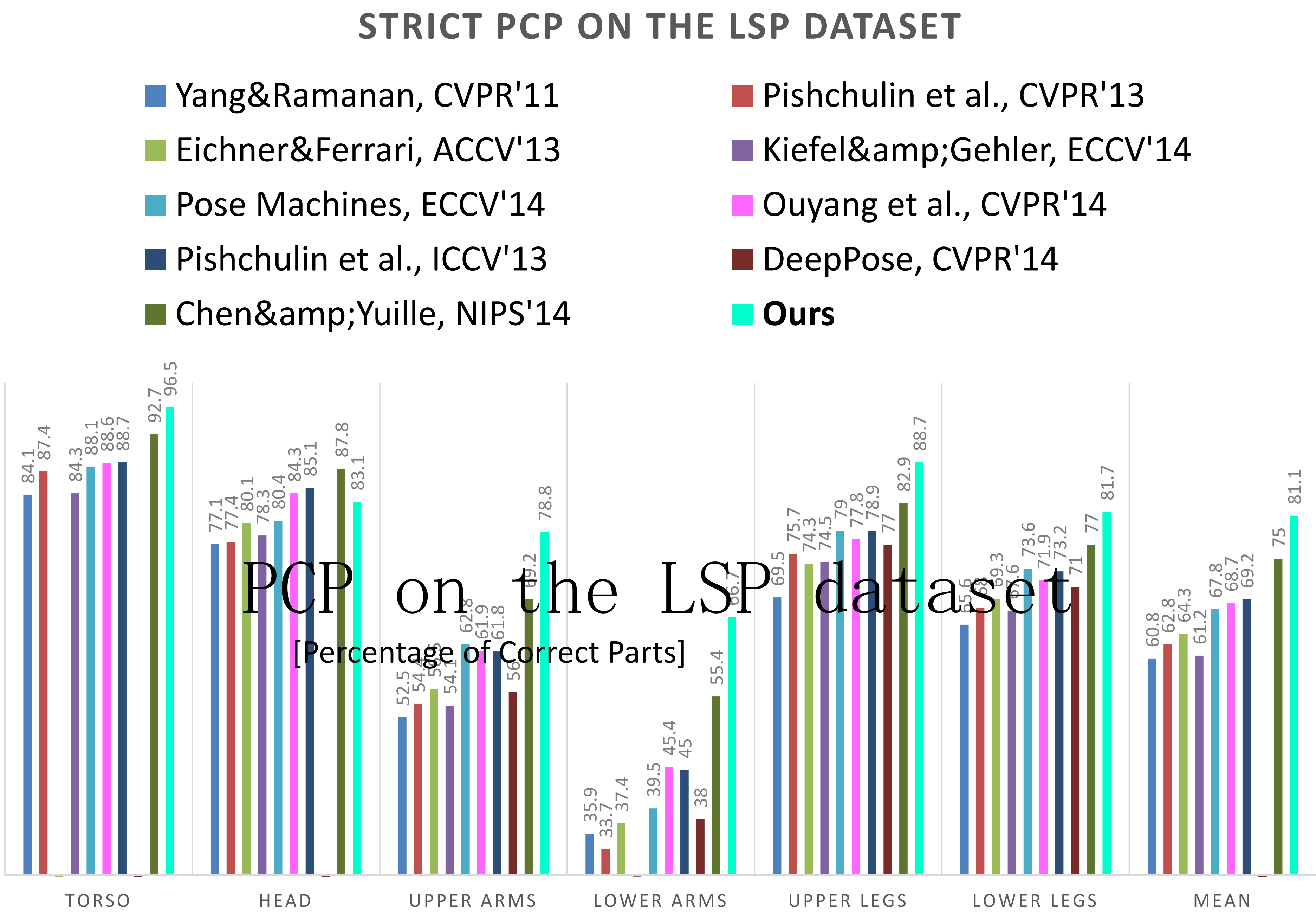


Message passing on score maps

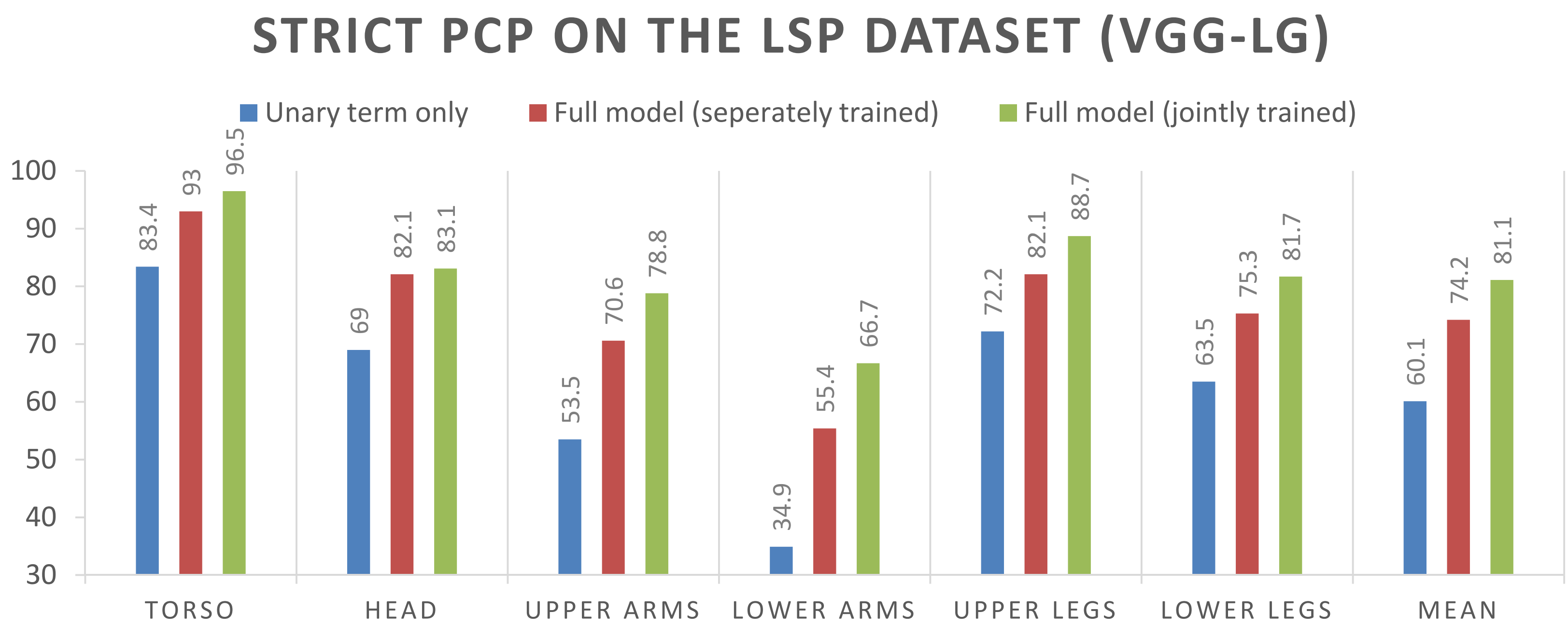
W. Yang, W. Ouyang, H. Li, and X. Wang, “End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation,” CVPR 2016

PCP on the LSP dataset

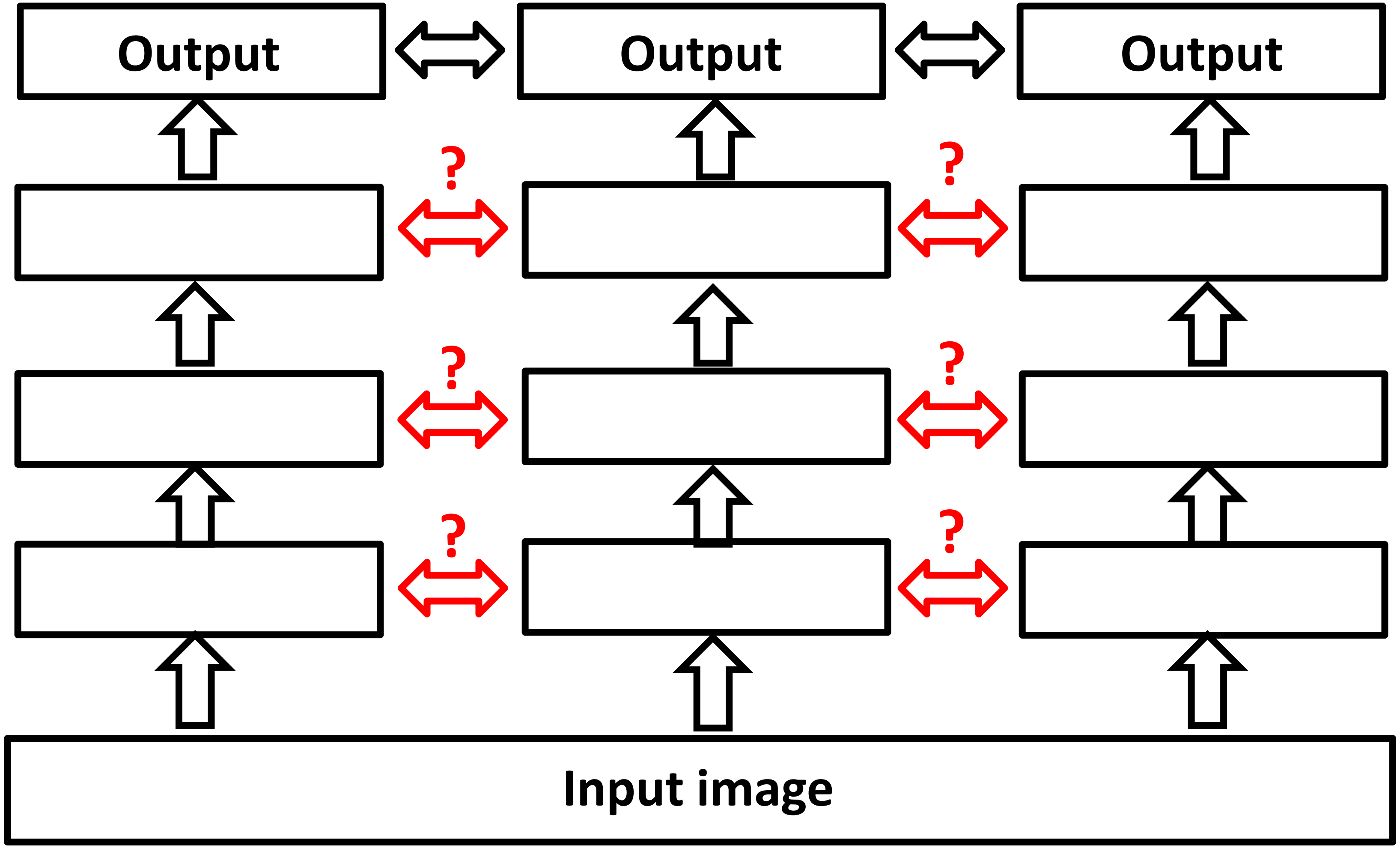
[Percentage of Correct Parts]



Unary Term vs. Full Model

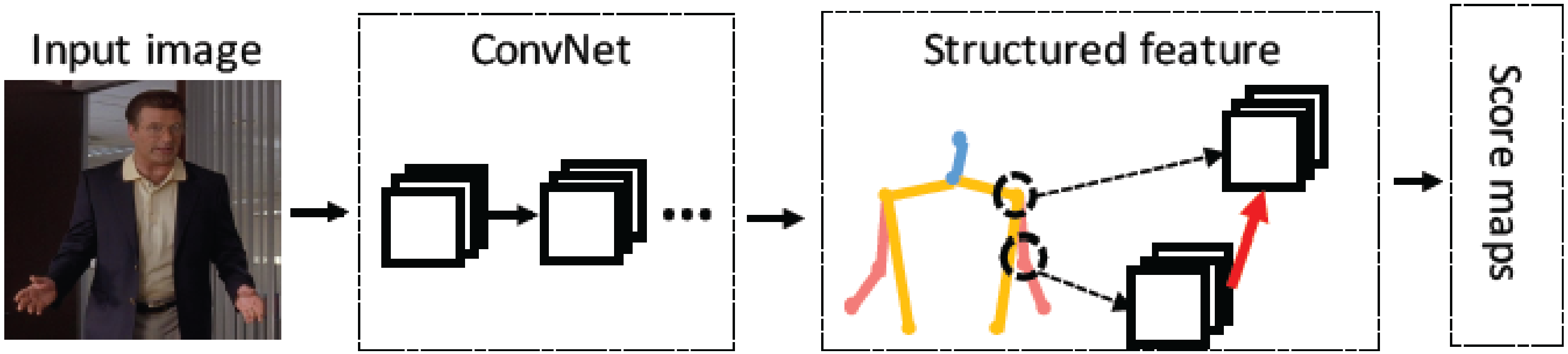


Model interaction between neurons in the same layer?



Structured Feature Learning

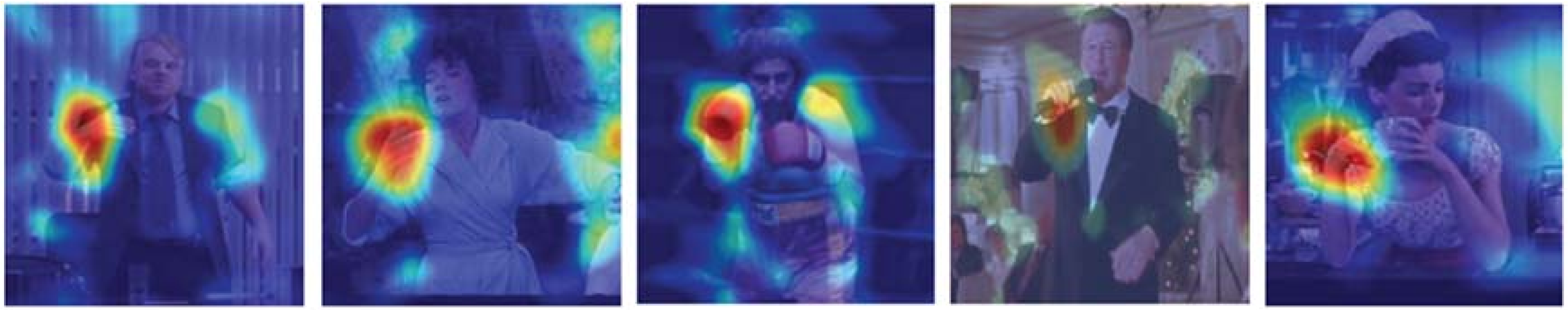
- Rich information is preserved at feature map level
- Reason the correlations among body joints at the feature level



X. Chu, W. Ouyang, W. Yang, and X. Wang, "Structured Feature Learning for Pose Estimation," CVPR 2016.

- **Understand the semantic meanings of feature maps**

High responding images for channel 1 for neck

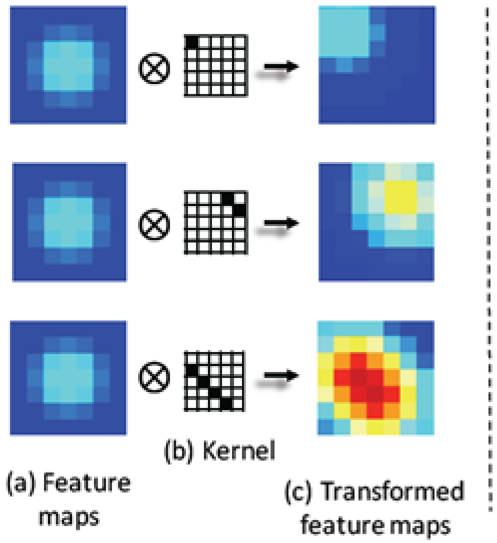


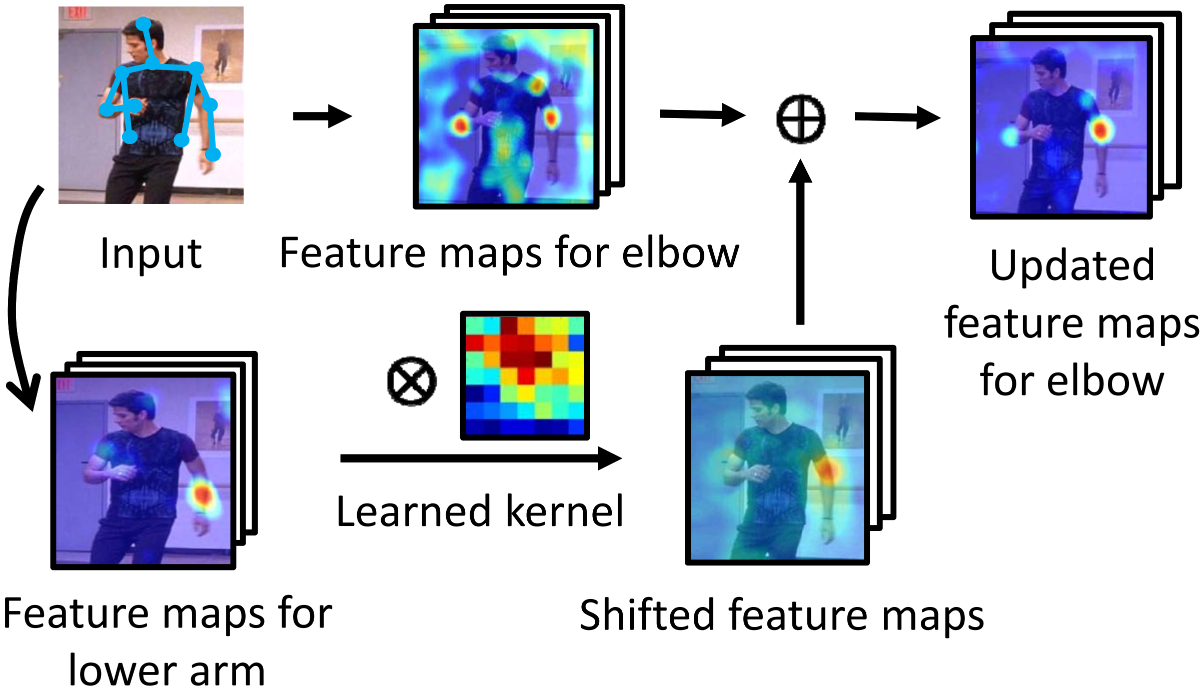
High responding images for channel 2 for left shoulder

High responding images for channel 3 for lower arm

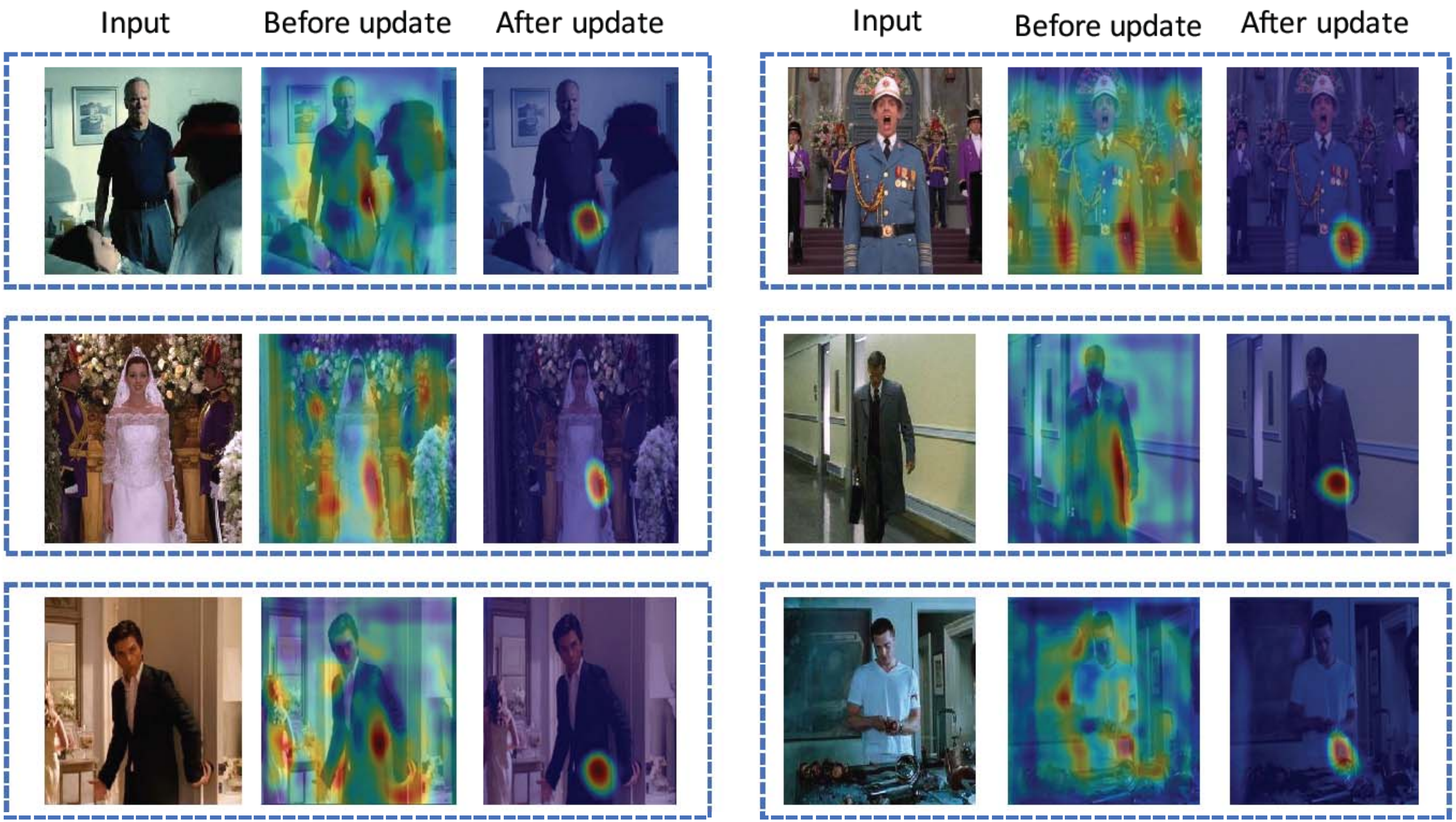
Geometrical Transform Kernels

- Pass information through convolution between feature maps and geometrical transform kernels





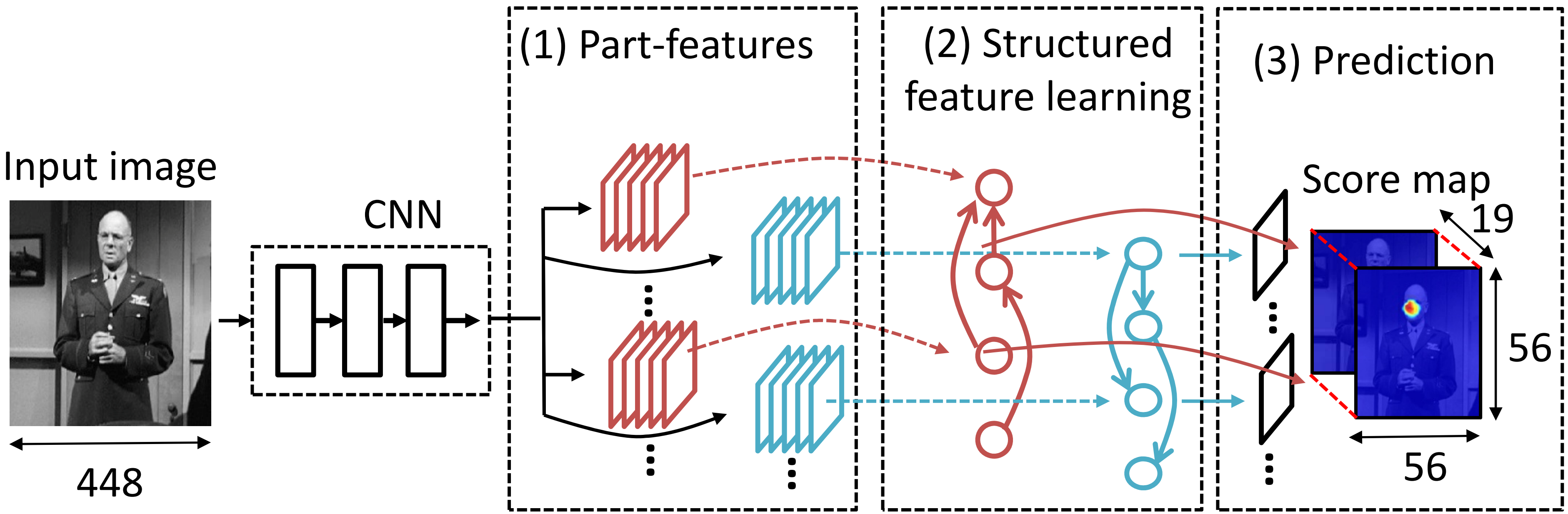
Feature map update --- Torso

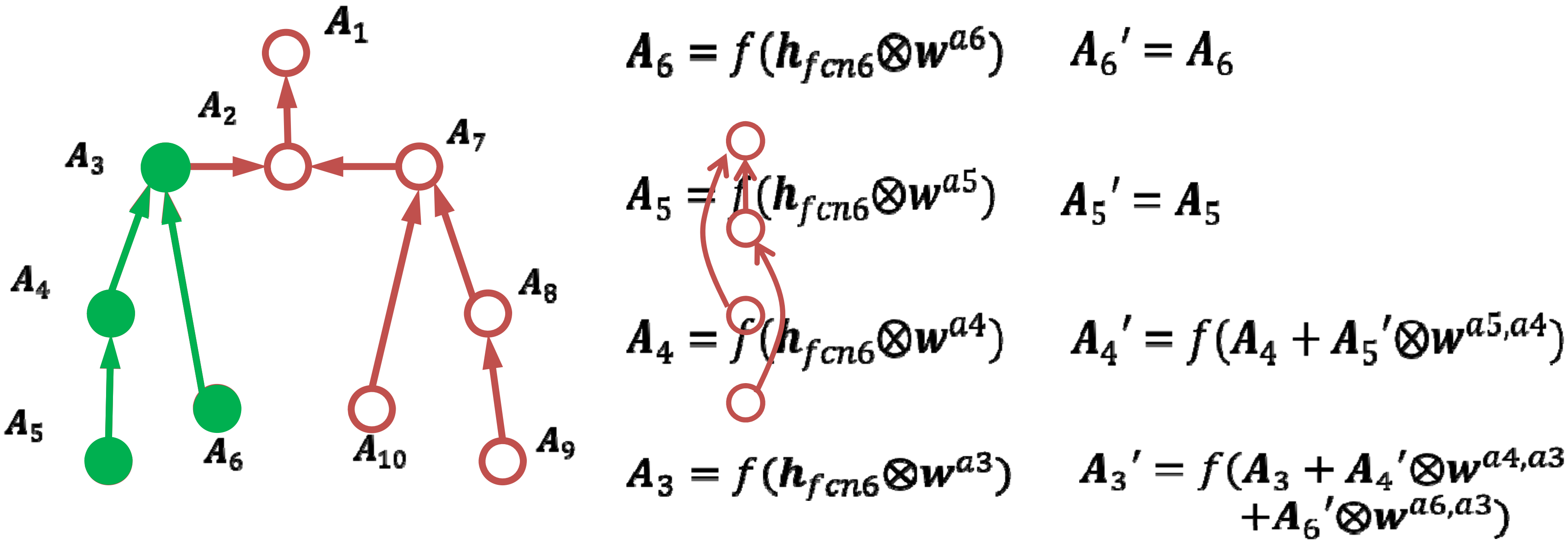


Feature map update --- Shoulder

Bidirectional Tree

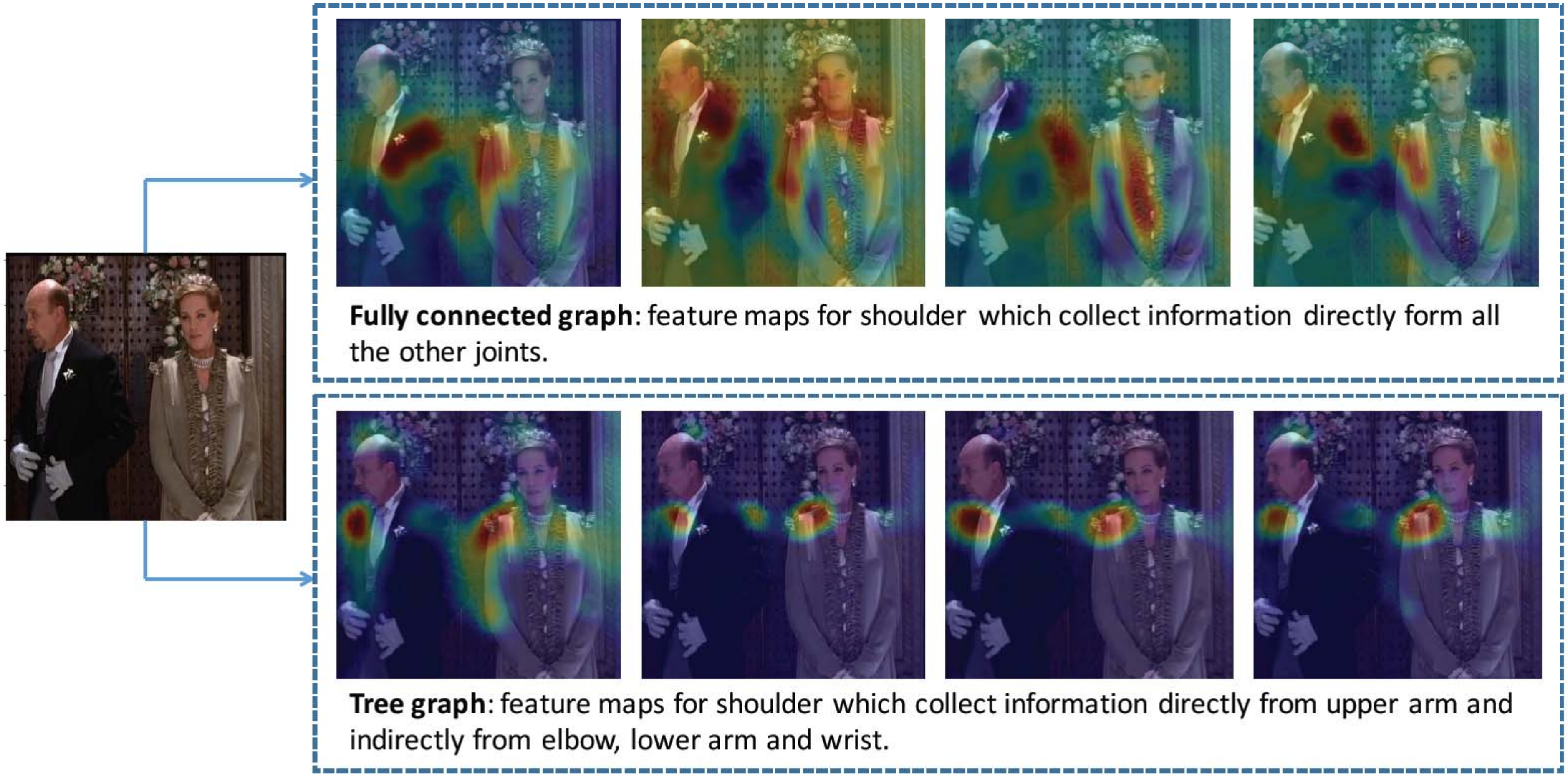
- Fully connected graph is not a good solution
  - Large transform kernels are required to model joints in distance
  - Relationship between some joints are unstable
- Propagate information through intermediate joints on a designed graph
- On a bi-directional tree, feature channels at a joint well receives information from other joints


Input image
448
CNN
(1) Part-features
(2) Structured feature learning
(3) Prediction
Score map
19
56
56



A_1
A_2
A_3
A_7
A_4
A_8
A_5
A_6
A_{10}
A_9
$A_6 = f(h_{fcn6} \otimes w^{a6})$ $A_6{}' = A_6$
$A_5 = f(h_{fcn6} \otimes w^{a5})$ $A_5{}' = A_5$
$A_4 = f(h_{fcn6} \otimes w^{a4})$ $A_4{}' = f(A_4 + A_5{}' \otimes w^{a5,a4})$
$A_3 = f(h_{fcn6} \otimes w^{a3})$ $A_3{}' = f(A_3 + A_4{}' \otimes w^{a4,a3} + A_6{}' \otimes w^{a6,a3})$

Fully connected graph is not a suitable structure for our method

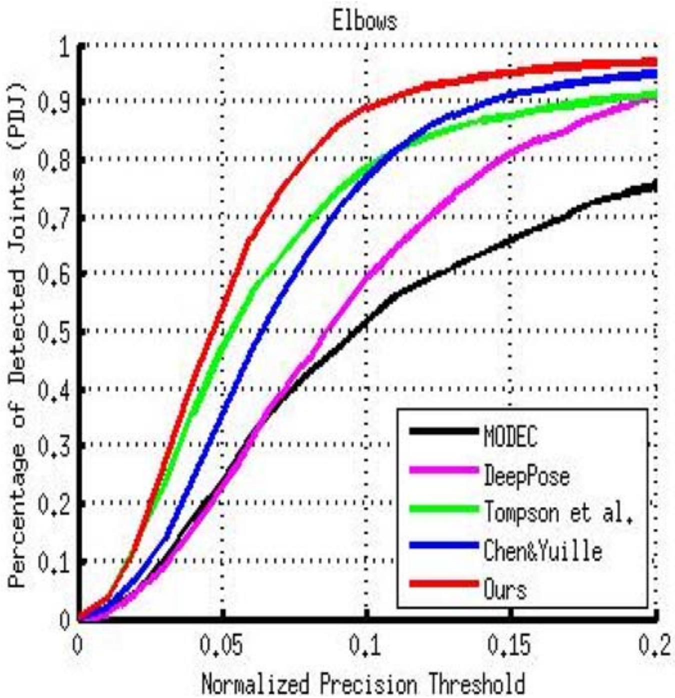


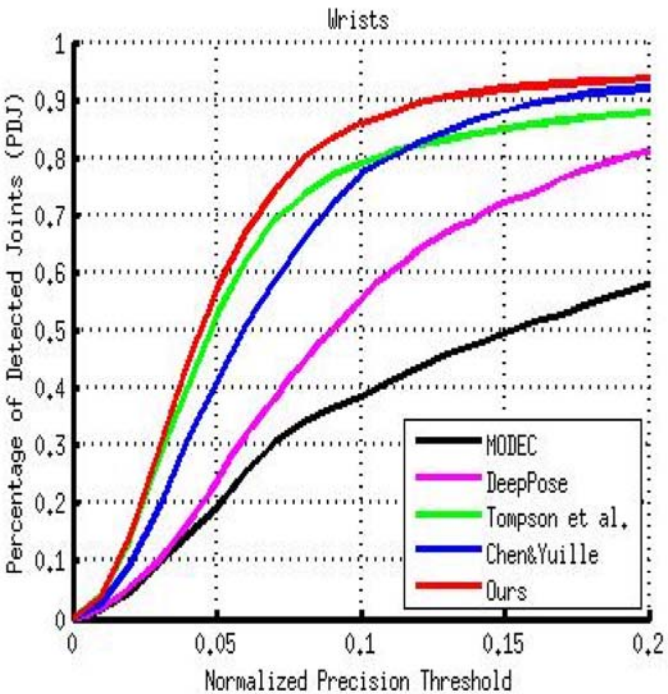
Fully connected graph: feature maps for shoulder which collect information directly form all the other joints.

Tree graph: feature maps for shoulder which collect information directly from upper arm and indirectly from elbow, lower arm and wrist.

Results(FLIC dataset)

Experiment	Head	Torso	U.arms	L.arms	Mean
MODEC [25]	-	-	84.4	52.1	68.3
Tompson *et al.* [31]	-	-	93.7	80.9	87.3
Chen& Yuille [7]	-	-	97.0	86.8	91.9
Ours	98.6	93.9	97.9	92.4	95.2





Results(LSP dataset)

Experiment	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
Andriluka *et al*. [2]	80.9	74.9	46.5	26.4	67.1	60.7	55.7
Yang&Ramanan [37]	82.9	79.3	56.0	39.8	70.3	67.0	62.8
Pishchulin *et al*. [22]	87.5	78.1	54.2	33.9	75.7	68.0	62.9
Eichner&Ferrari *et al*. [10]	86.2	80.1	56.5	37.4	74.3	69.3	64.3
Ouyang *et al*. [18]	85.8	83.1	63.3	46.6	76.5	72.2	68.6
Pishchulin *et al*. [23]	88.7	85.1	61.8	45.0	78.9	73.2	69.2
Chen&Yuille [7]	92.7	87.8	69.2	55.4	82.9	77.0	75.0
Ours	**95.4**	**89.4**	**76.0**	**64.3**	**87.6**	**83.5**	**80.8**

CRF-CNN

(a) Multi-layer neural network (b) Structured output space (c) Structured hidden layer (d) Our implementation

$$p(\mathbf{z}|\mathbf{I},\Theta) = \sum_{\mathbf{h}} p(\mathbf{z},\mathbf{h}|\mathbf{I},\Theta), \text{ where } p(\mathbf{z},\mathbf{h}|\mathbf{I},\Theta) = \frac{e^{-En(\mathbf{z},\mathbf{h},\mathbf{I},\Theta)}}{\sum_{\mathbf{z}\in\mathcal{Z},\mathbf{h}\in\mathcal{H}} e^{-En(\mathbf{z},\mathbf{h},\mathbf{I},\Theta)}}$$

$$En(\mathbf{z},\mathbf{h},\mathbf{I},\Theta) = \sum_{\substack{(k,l)\in\mathcal{E}_h \\ k<l}} \psi_h(h_k,h_l) + \sum_{\substack{(i,j)\in\mathcal{E}_z \\ i<j}} \psi_z(\mathbf{z}_i,\mathbf{z}_j) + \sum_{(i,k)\in\mathcal{E}_{zh}} \psi_{zh}(\mathbf{z}_i,h_k) + \sum_k \phi_h(h_k,\mathbf{I})$$

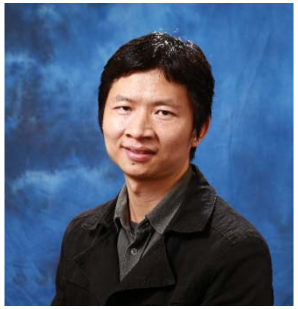
X. Chu, W. Ouyang, H. Li, and X. Wang, “ CRF-CNN: Modeling Structured Information in Human Pose Estimation,” NIPS 2016.

Conclusions

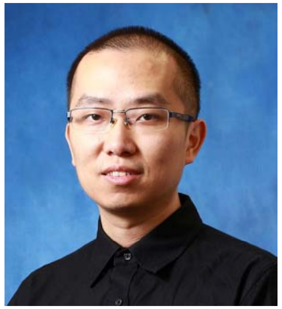
- Structural information is important in feature learning
- End-to-end joint training bridges the gap between structure modeling and feature learning
- A new message passing layer, which is flexible to tree-structured/loopy relational graph
- Feature level information passing delivers more detailed descriptions about body joints than score maps. It is implemented with geometrical transform kernels.
- Propose a CRF-CNN framework to simultaneously model structural information in both output and hidden feature layers in a probabilistic way.

Wei Yang

Xiao Chu



Wanli Ouyang



Hongsheng Li

Xiaogang Wang