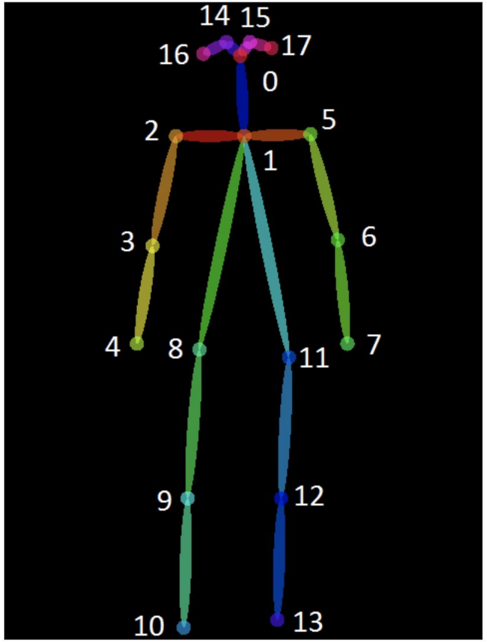
Pose estimation

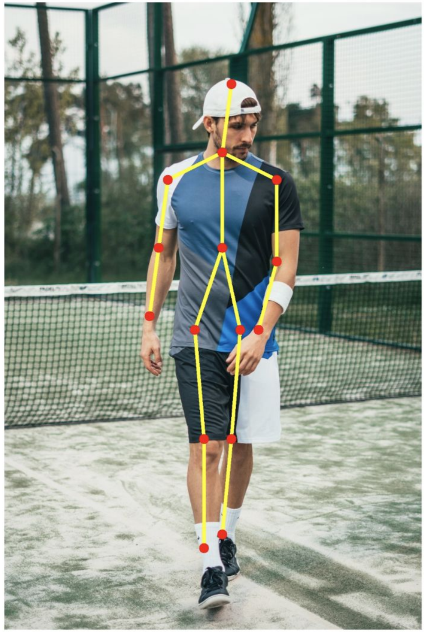
Instructor: Seunghoon Hong
School of Computing, KAIST

Image source: https://github.com/CMU-Perceptual-Computing-Lab/openpose

Human pose estimation

- Goal: identifying and localizing human joints (or key-points)

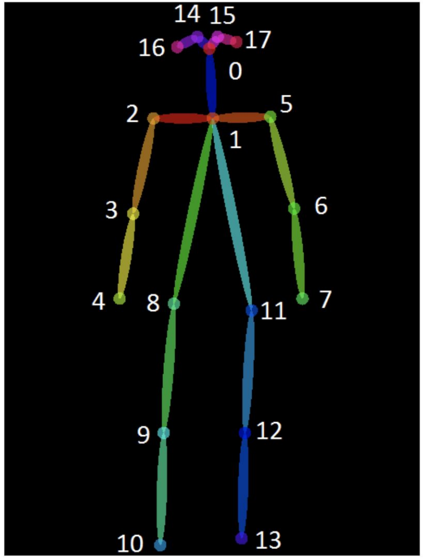




Human pose estimation

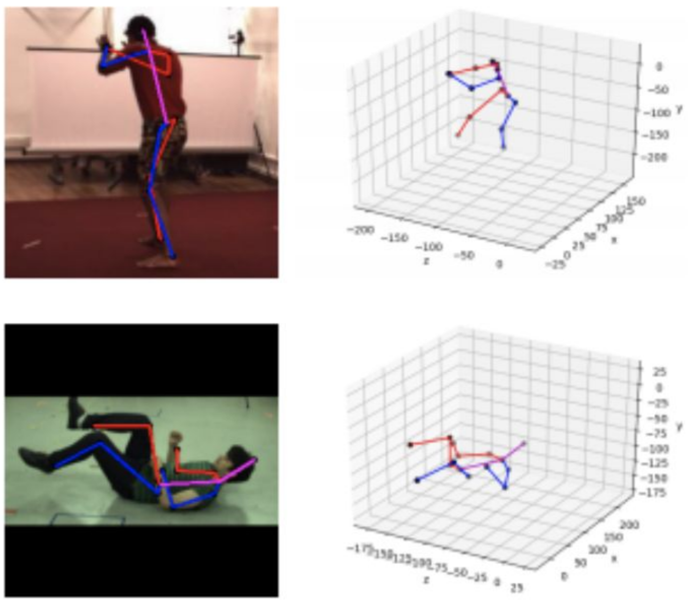
2d pose estimation

Identifying (x,y) position of joints



3d pose estimation

Identifying (x,y,z) position of joints



Human pose estimation

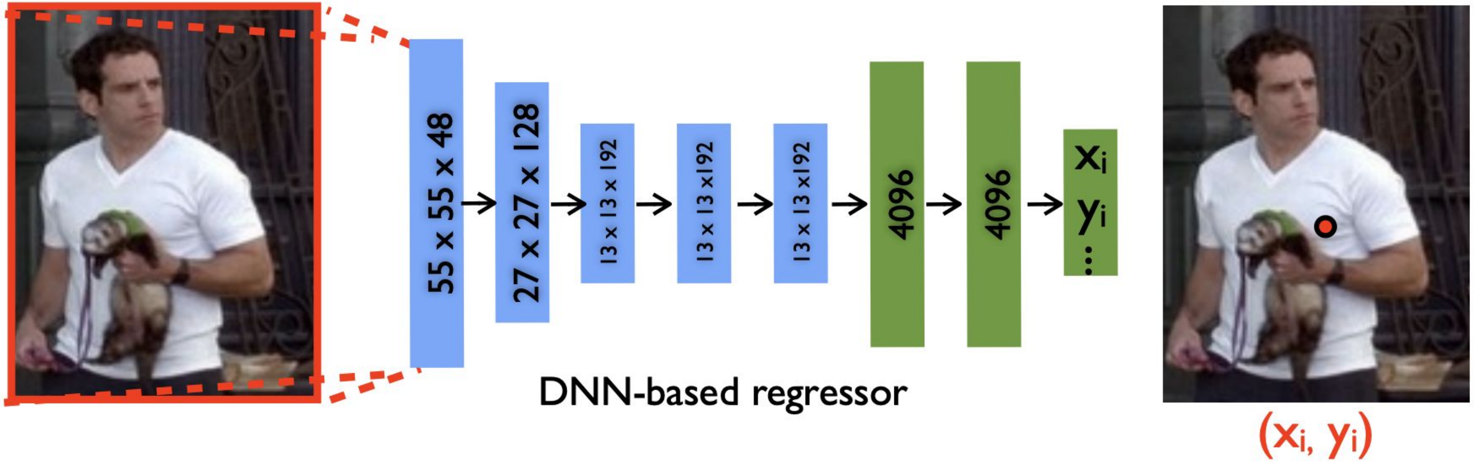
- Challenges
  - Significant pose variations (articulation, deformation)
  - Partial observation (occlusion)
  - Different body-part configuration per person, etc...

CNN for pose estimation

- DeepPose
- Convolutional Pose Machine
- Iterative Error Feedback
- Stacked hourglass

DeepPose: holistic pose estimation using CNN

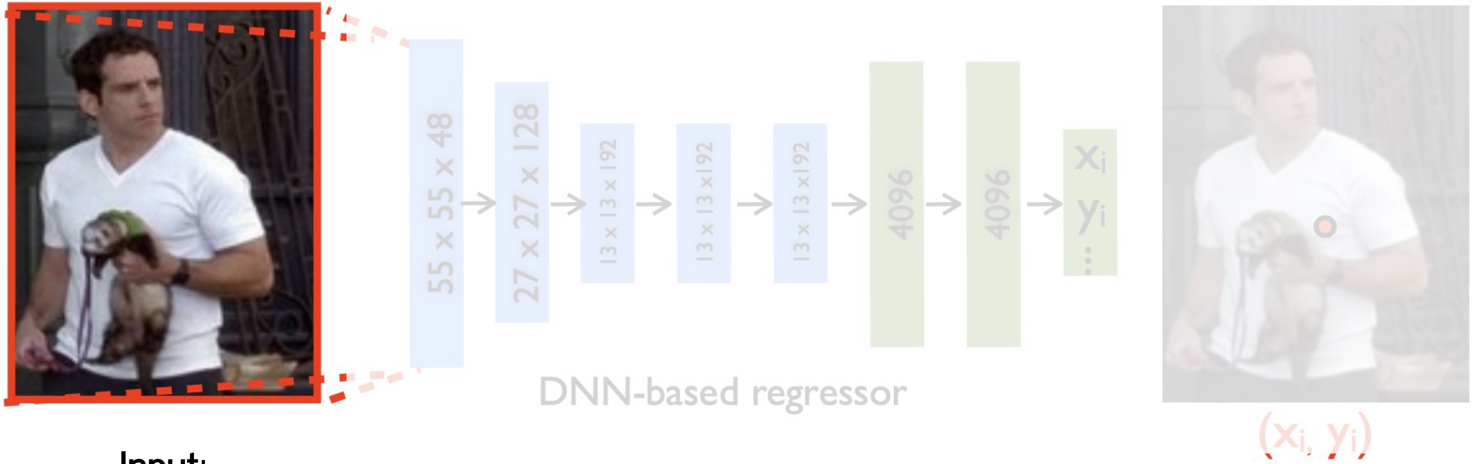
- Predict the human joints from an holistic image observation



Toshev et al., DeepPose: Human Pose Estimation via Deep Neural Networks, In CVPR, 2014

DeepPose: holistic pose estimation using CNN

- Predict the human joints from an holistic image observation

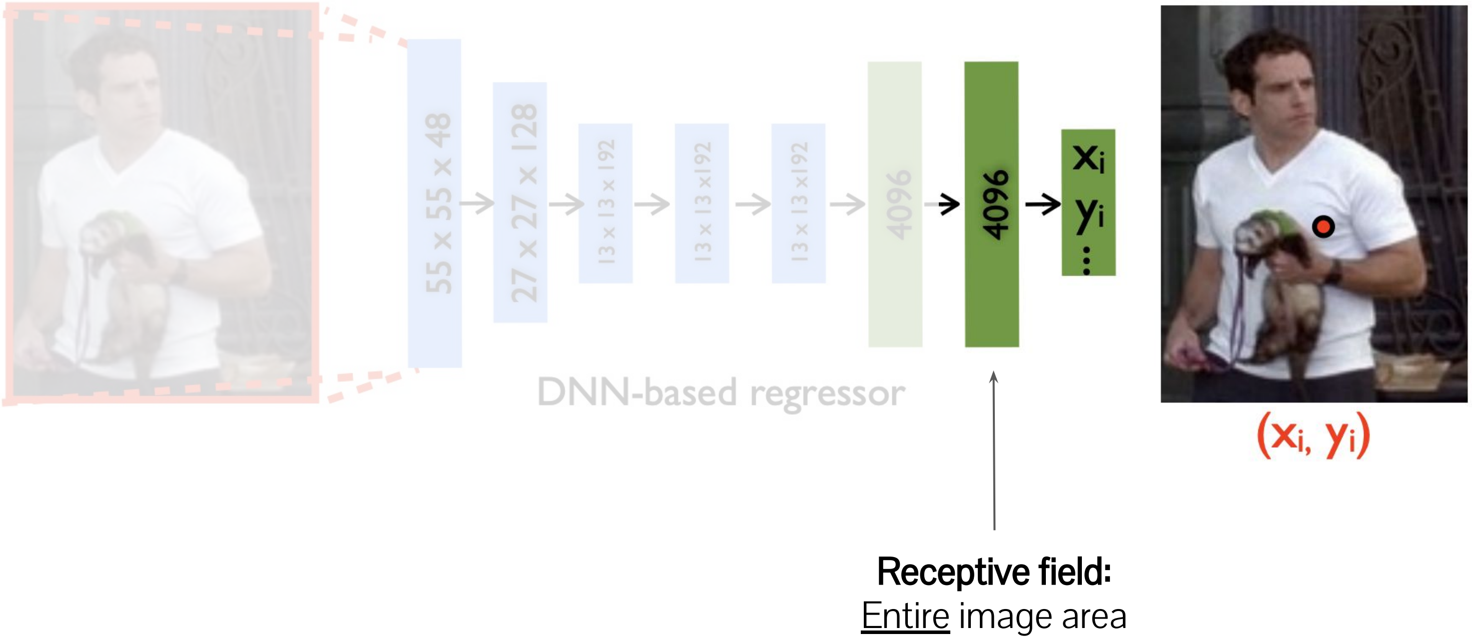


Input:
220x220 image

Toshev et al., DeepPose: Human Pose Estimation via Deep Neural Networks, In CVPR, 2014

DeepPose: holistic pose estimation using CNN

- Predict the human joints from an holistic image observation



Toshev et al., DeepPose: Human Pose Estimation via Deep Neural Networks, In CVPR, 2014

DeepPose: holistic pose estimation using CNN

- Predict the human joints from an holistic image observation

Output: (x,y) coordinate of each object joint

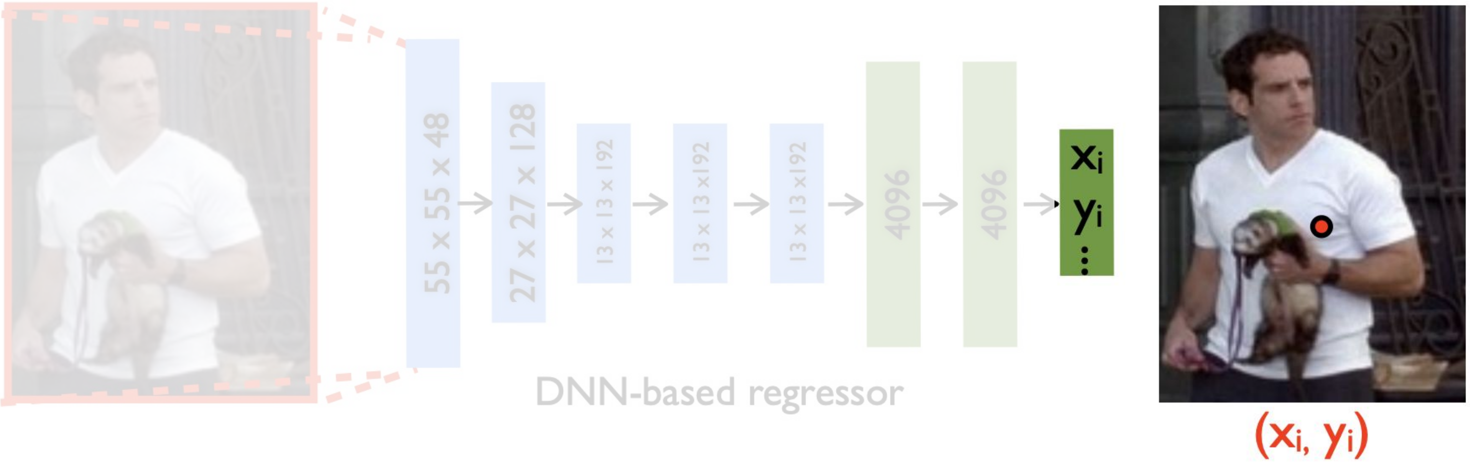
- Total Kx2 outputs (K: # of joints)
- Each joint coordinate is normalized between [0,1]

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \\ x_k & y_k \end{bmatrix} \longrightarrow \text{K x 2 matrix}$$

Toshev et al., DeepPose: Human Pose Estimation via Deep Neural Networks, In CVPR, 2014

DeepPose: holistic pose estimation using CNN

- Predict the human joints from an holistic image observation

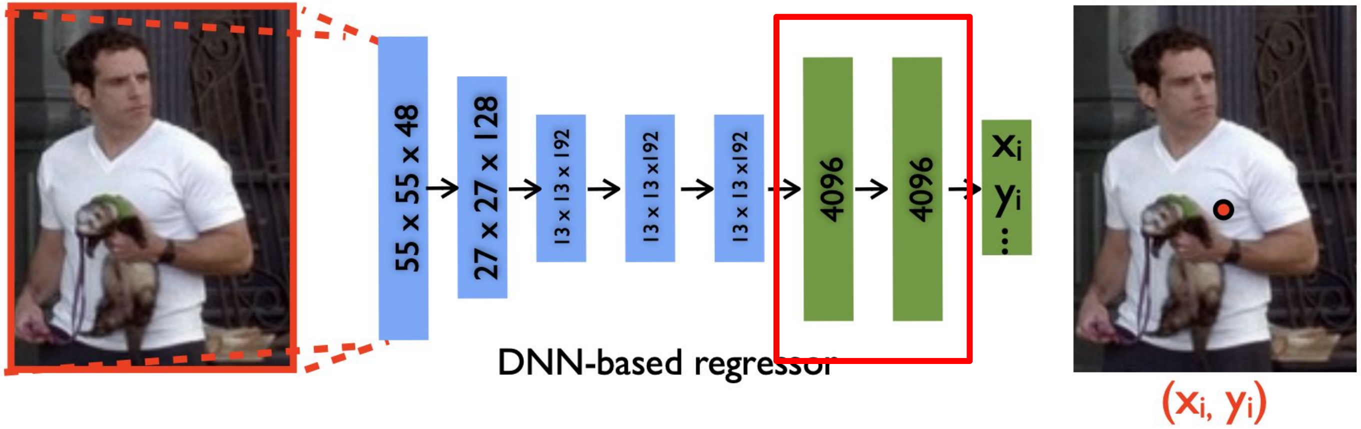


Loss function: reducing a L2 distance between ground-truth and predicted joints

$$\arg\min_{\theta} \sum_{(x,y)\in D_N} \sum_{i=1}^{k} ||\mathbf{y}_i - \psi_i(x;\theta)||_2^2$$

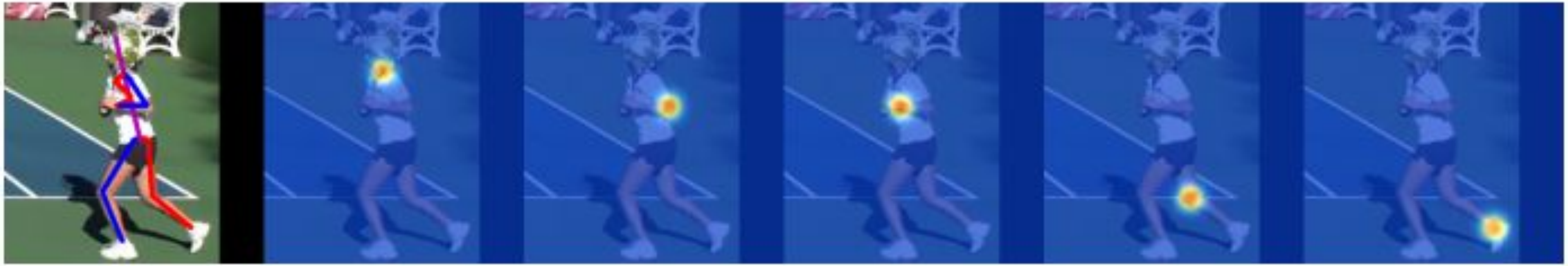
DeepPose: holistic pose estimation using CNN

- Limitations
  - Too much spatial abstraction for capturing holistic information (**224x224**x3 → **1x1**x4096)
  - May not appropriate for accurate localization of joints



From image to human joints

- Instead of directly converting an image to (x,y) coordinates,
  Predict **heat map** (score map) of joints in image coordinate

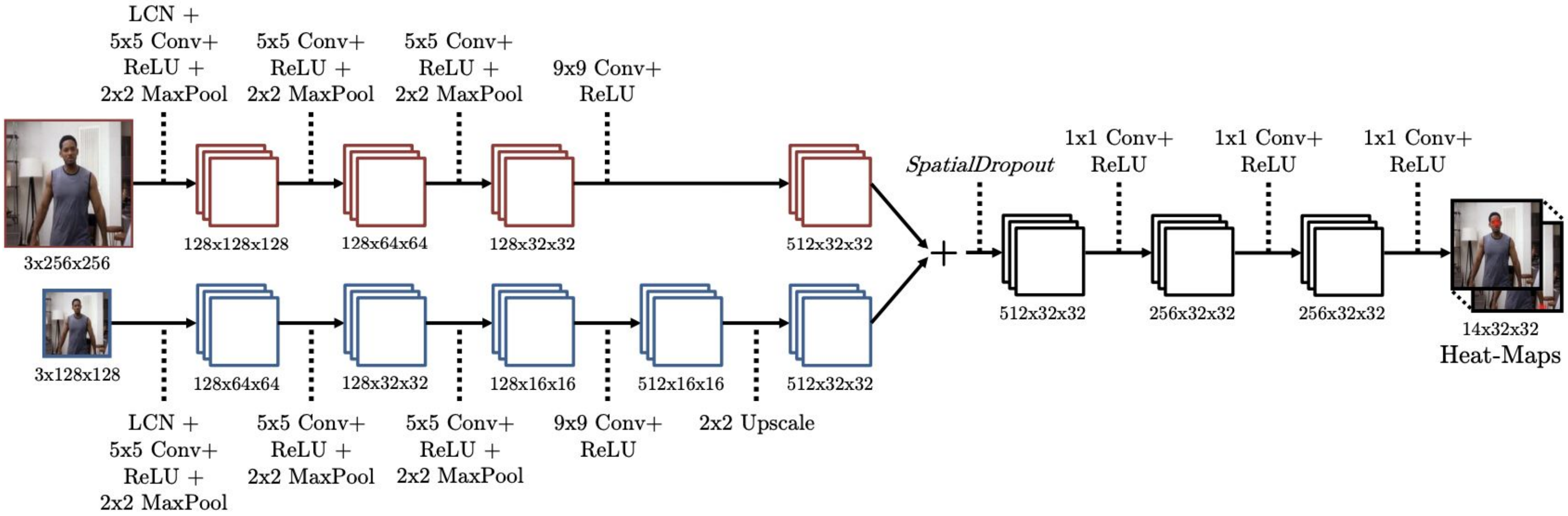


- Benefits:
  - Reducing spatial abstraction
  - Reduce complexity of prediction task → less prone to overfitting

Tompson et al., Efficient Object Localization Using Convolutional Network, In CVPR, 2015

Multi-resolution Heat Map Regressor

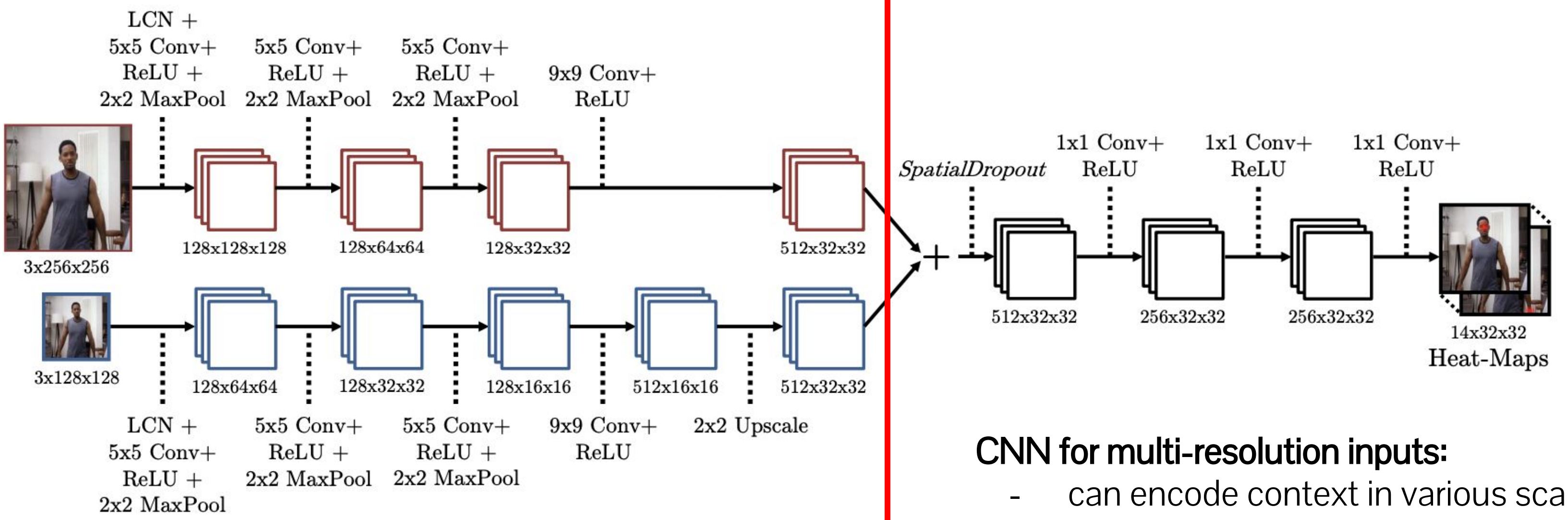
- Instead of directly converting an image to (x,y) coordinates, Predict **heat map** (score map) of joints in image coordinate



Tompson et al., Efficient Object Localization Using Convolutional Network, In CVPR, 2015

Multi-resolution Heat Map Regressor

- Instead of directly converting an image to (x,y) coordinates, Predict **heat map** (score map) of joints in image coordinate



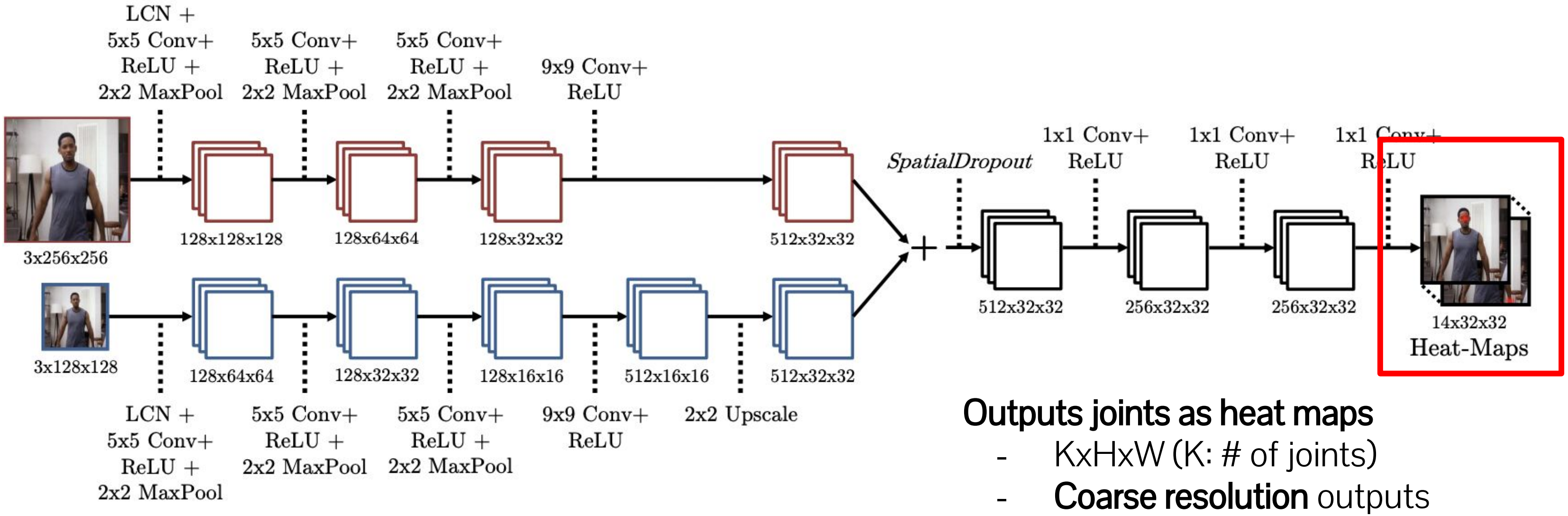
CNN for multi-resolution inputs:

- can encode context in various scales

Tompson et al., Efficient Object Localization Using Convolutional Network, In CVPR, 2015

Multi-resolution Heat Map Regressor

- Instead of directly converting an image to (x,y) coordinates, Predict **heat map** (score map) of joints in image coordinate



Outputs joints as heat maps

- KxHxW (K: # of joints)
- **Coarse resolution** outputs (256x256 → 32x32)

Tompson et al., Efficient Object Localization Using Convolutional Network, In CVPR, 2015

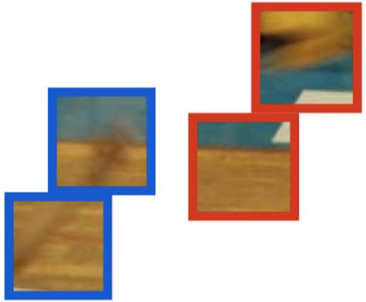
CNN for pose estimation

- DeepPose
- **Convolutional Pose Machine**
- Iterative Error Feedback
- Stacked hourglass

Which patch corresponds to a body part?

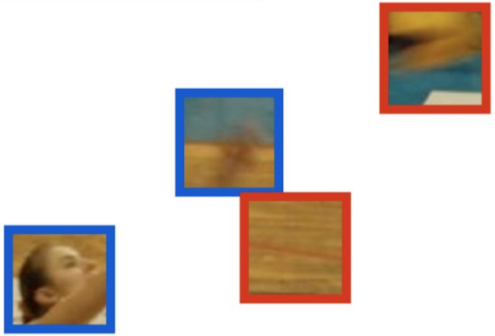
Slide credit: Varun Ramakrishna

Which patch corresponds to a body part?



Slide credit: Varun Ramakrishna

Which patch corresponds to a body part?



Slide credit: Varun Ramakrishna

Which patch corresponds to a body part?

Slide credit: Varun Ramakrishna

Which patch corresponds to a body part?

- **Local evidence is weak**

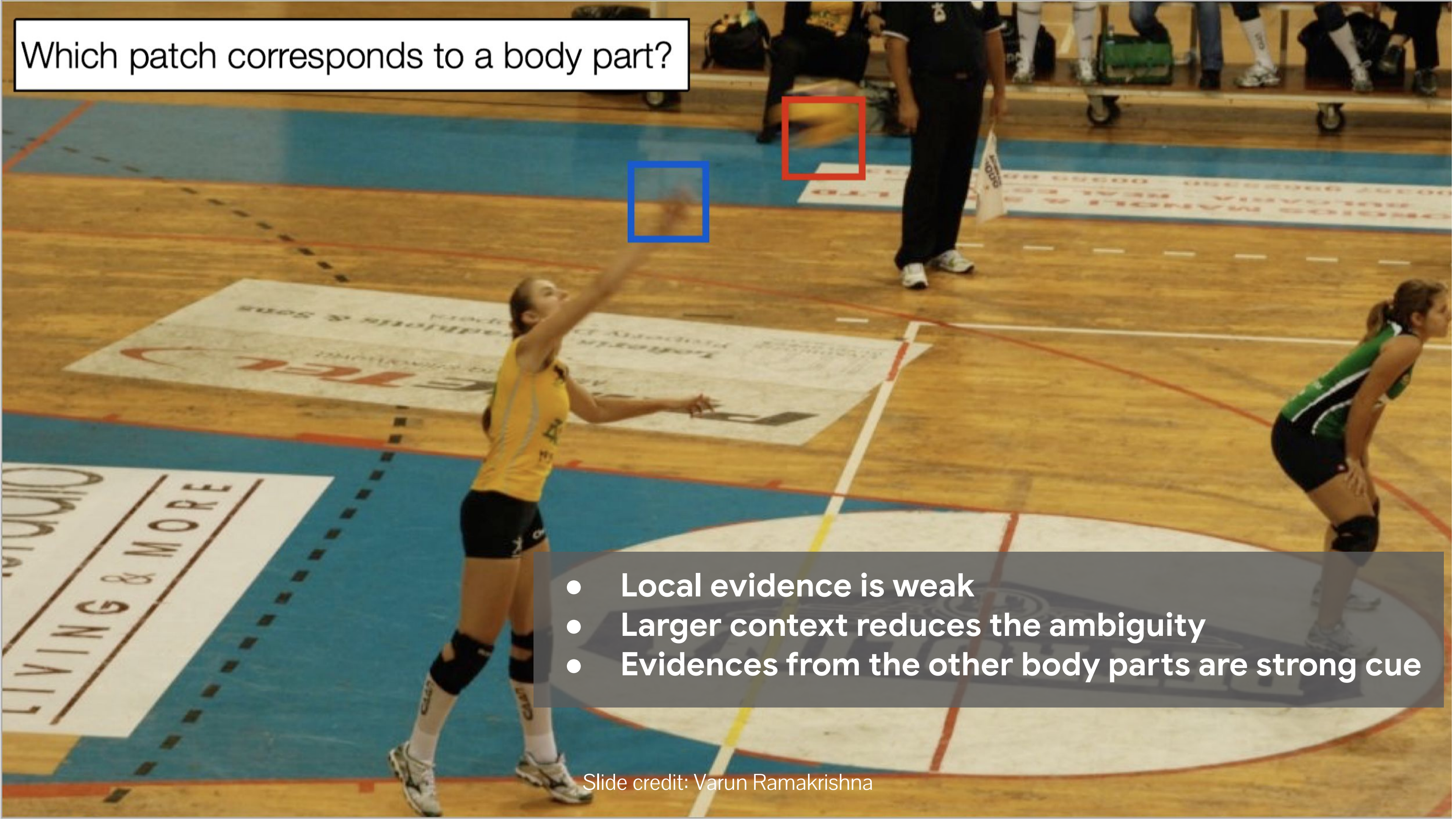
Slide credit: Varun Ramakrishna

Which patch corresponds to a body part?

- Local evidence is weak
- Larger context reduces the ambiguity

Slide credit: Varun Ramakrishna

Which patch corresponds to a body part?

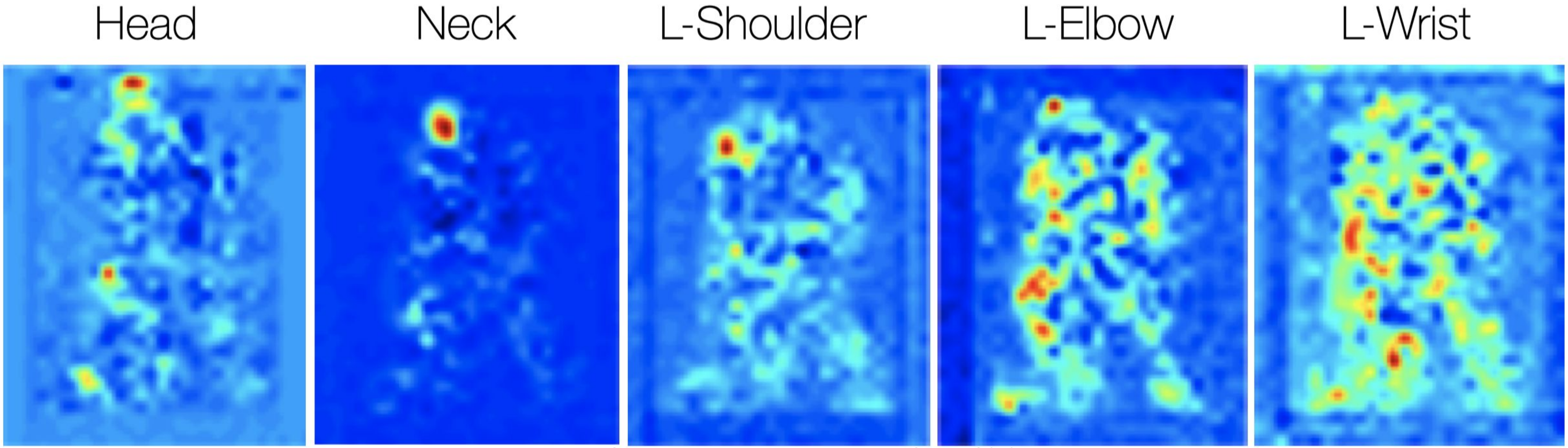


- Local evidence is weak
- Larger context reduces the ambiguity
- Evidences from the other body parts are strong cue

Slide credit: Varun Ramakrishna

Limitations in local evidences

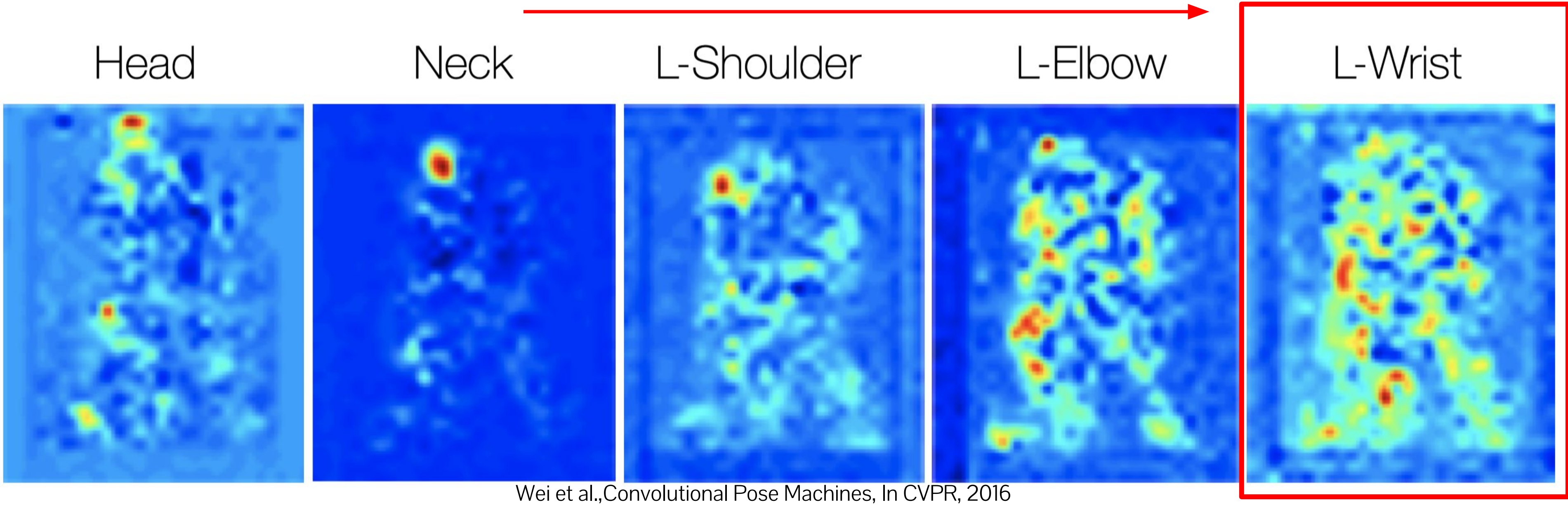
- Local evidences are sometimes ambiguous



Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Limitations in local evidences

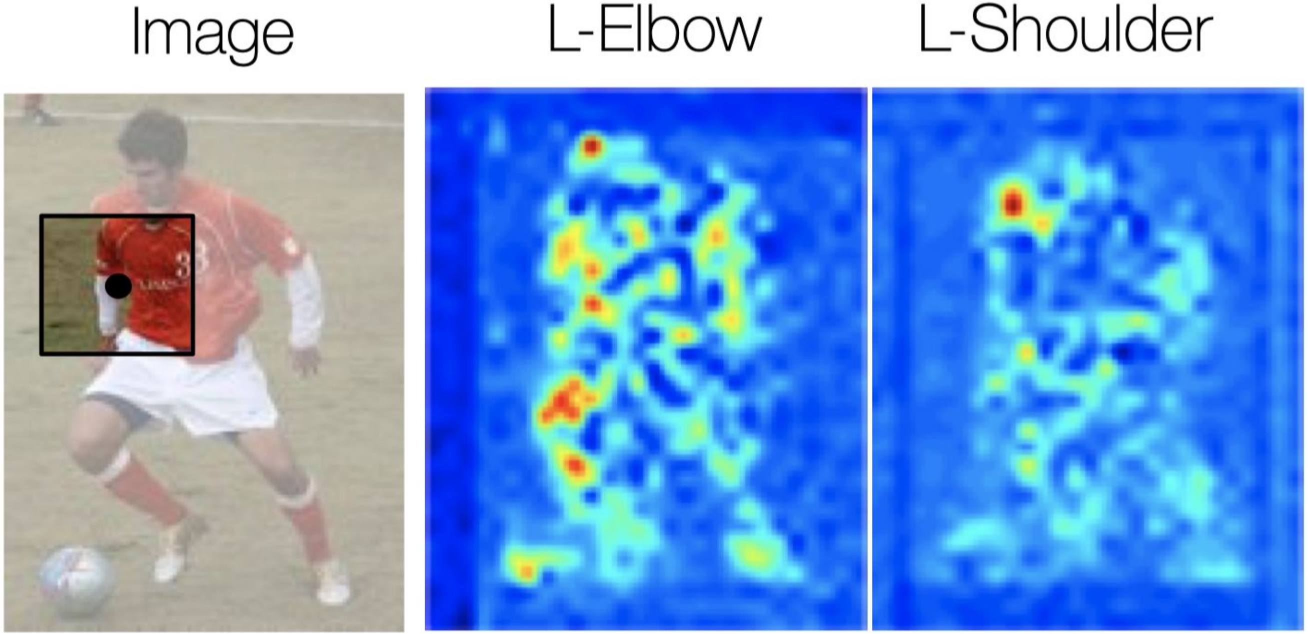
- Local evidences are sometimes ambiguous
- Some body parts are more difficult to detect than others
  (e.g. more deformable parts are difficult to detects)



Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Incorporating a larger context

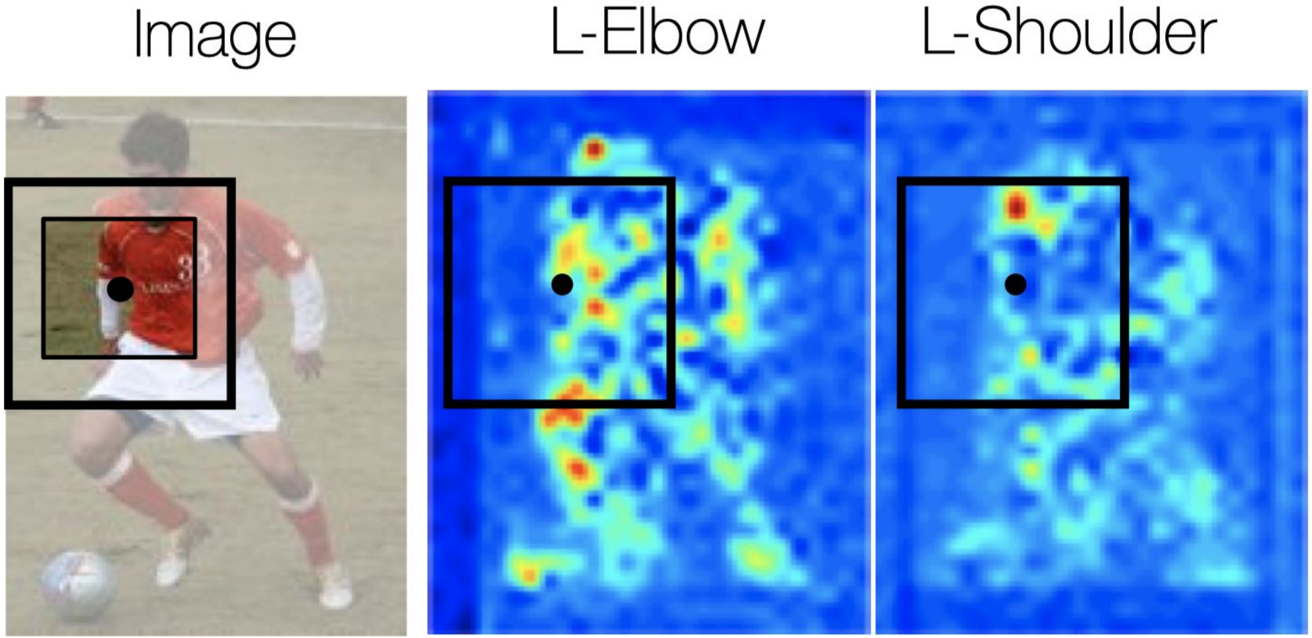
- Reducing the ambiguity of part-wise detection by the evidence of other body parts



Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Incorporating a larger context

- Reducing the ambiguity of part-wise detection by the evidence of other body parts



Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Incorporating a larger context

How to incorporate a larger context into a part localization?

Incorporating a larger context

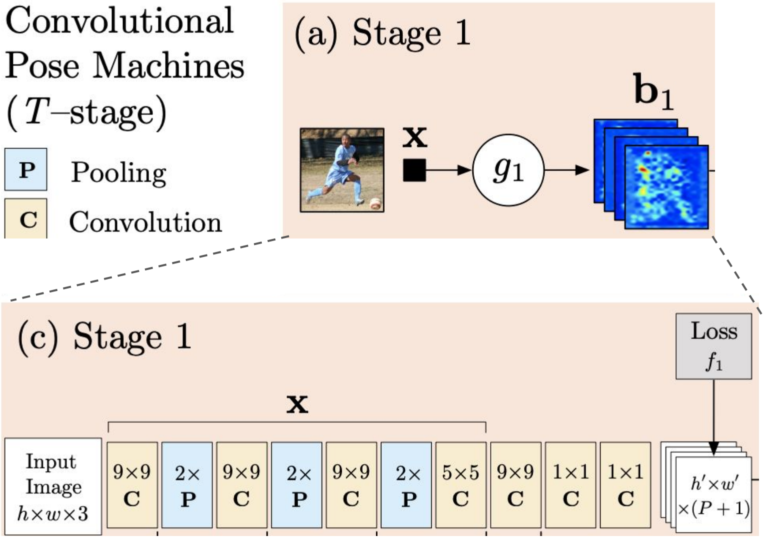
How to incorporate a larger context into a part localization?

→ Increasing the receptive field!

- Increase the pooling window (or number of poolings)
- Increase the convolutional filter size
- Spatial pyramid pooling, multiple image resolution, etc....

Convolutional Pose Machine

- Iterative refinement of part localization by increasing receptive field

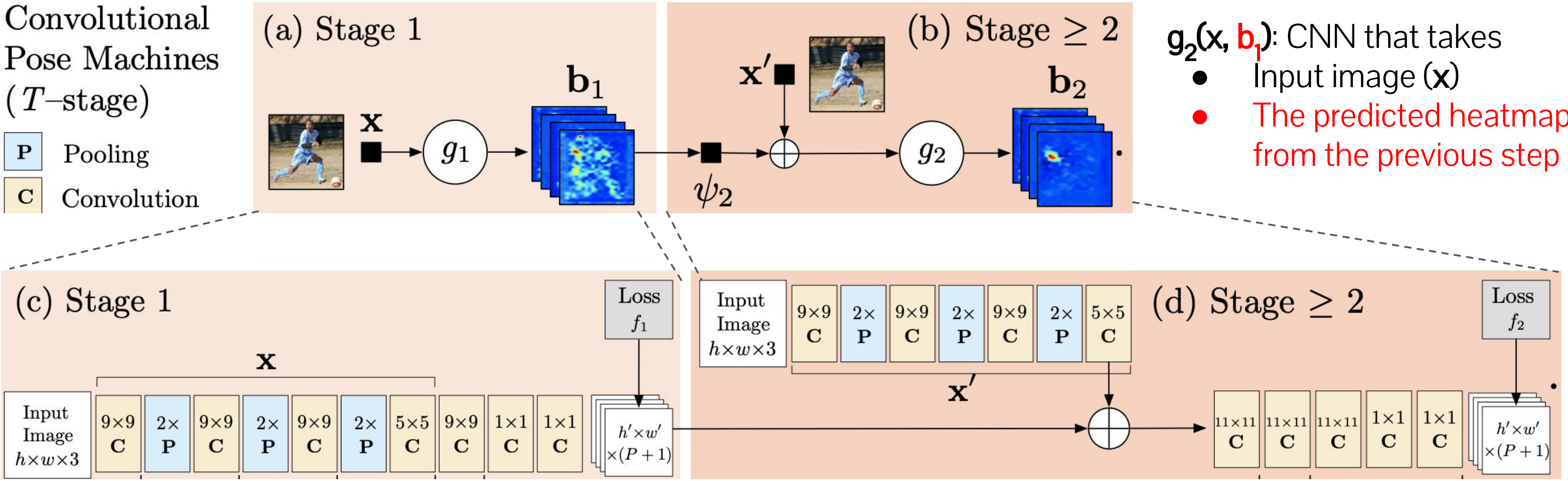


- **g(x)**: CNN that produces heat map of body parts (**b**)
- The predicted heat maps are ambiguous

Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Convolutional Pose Machine

- Iterative refinement of part localization by increasing receptive field

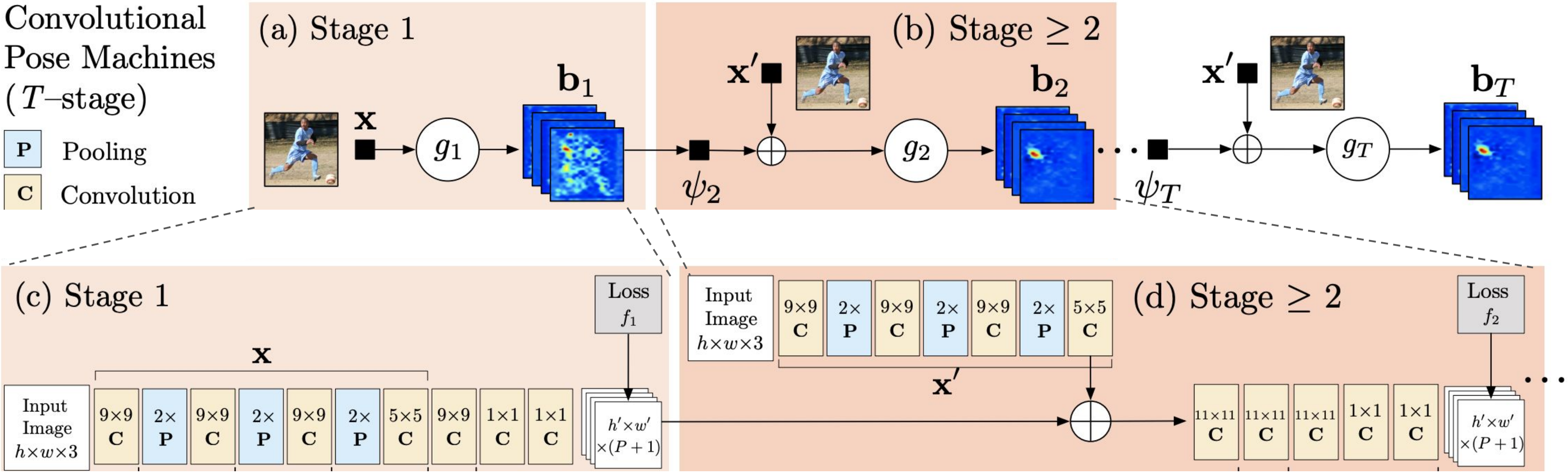


$g_2(\mathbf{x}, \mathbf{b}_1)$: CNN that takes
- Input image ($\mathbf{x}$)
- The predicted heatmaps from the previous step ($\mathbf{b}_1$)

Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Convolutional Pose Machine

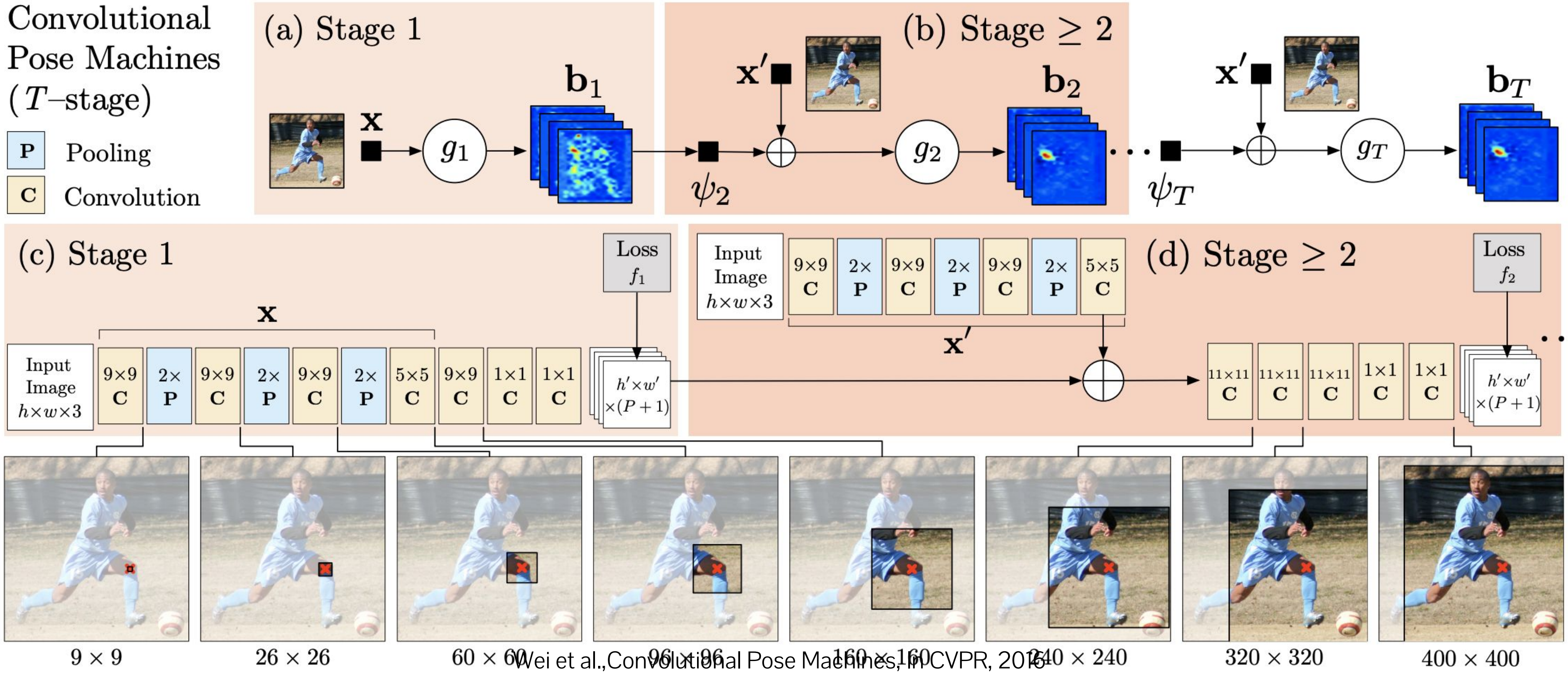
- Iterative refinement of part localization by increasing receptive field



Wei et al.,Convolutional Pose Machines, In CVPR, 2016

Convolutional Pose Machine

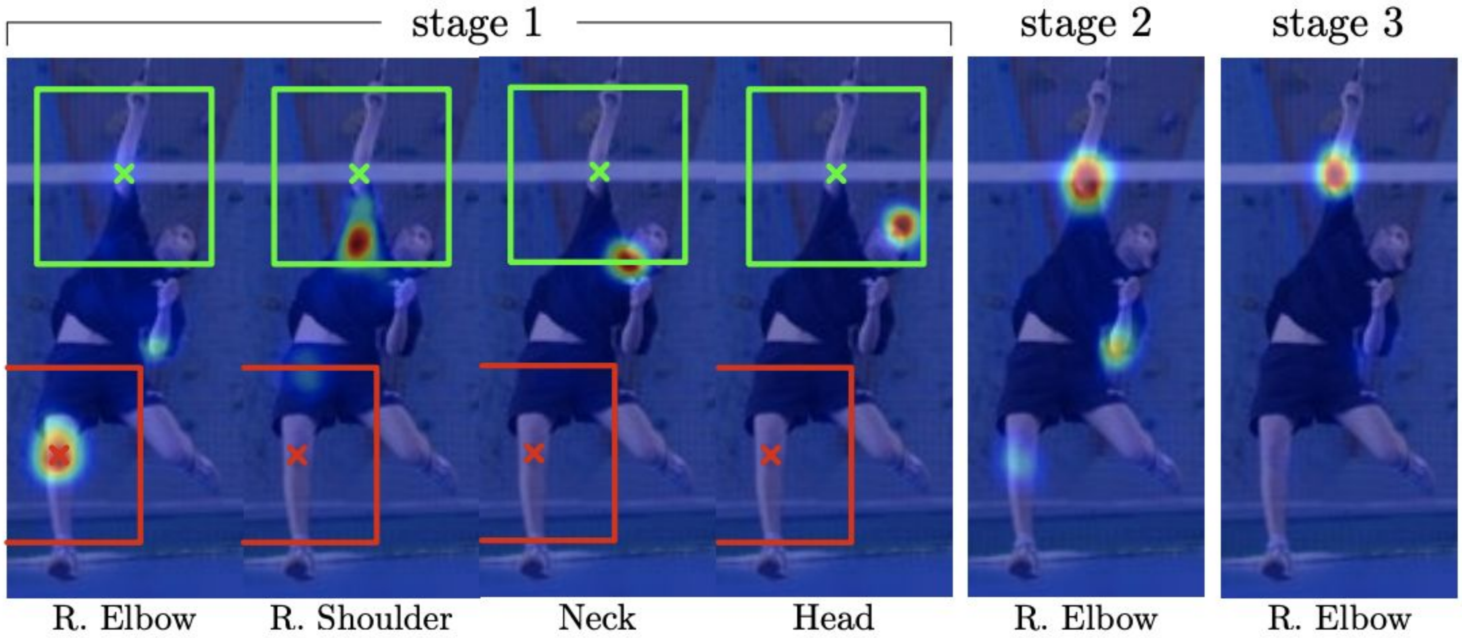
- Iterative refinement of part localization by increasing receptive field



Wei et al., Convolutional Pose Machines, In CVPR, 2016

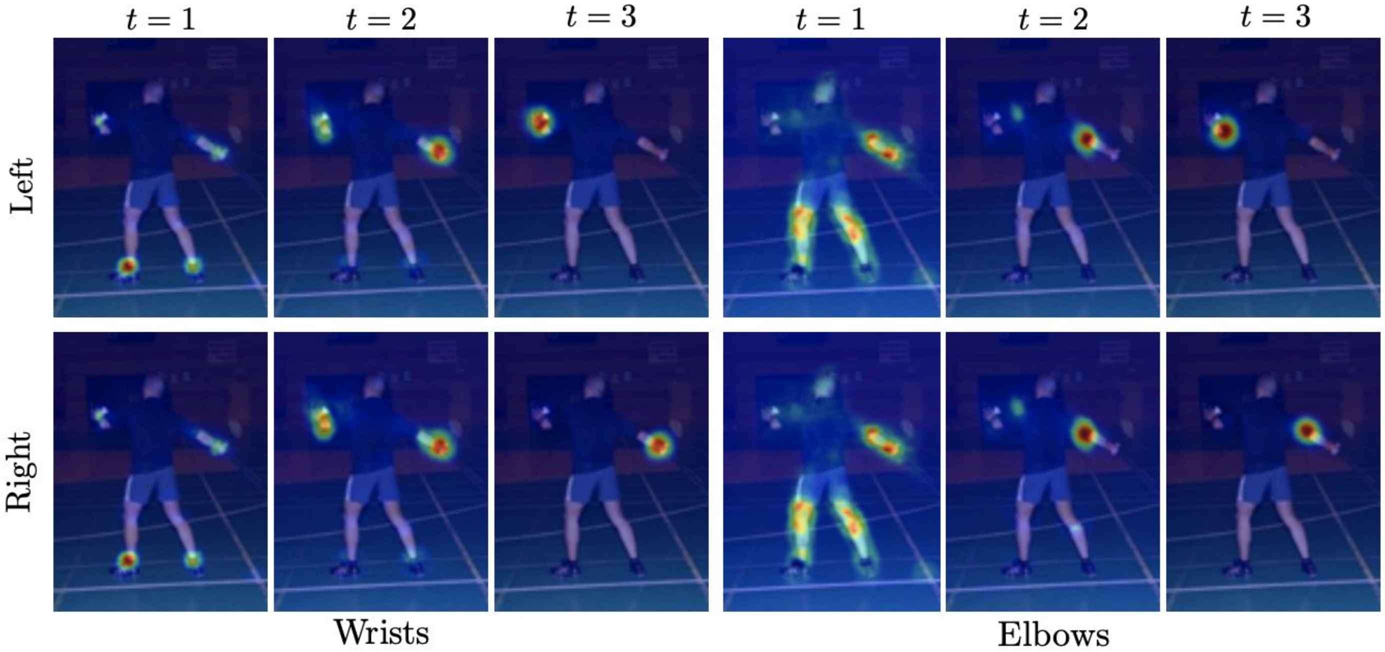
Convolutional Pose Machine

- Iterative refinement leads to elimination of noisy predictions and the discovery of missed body parts



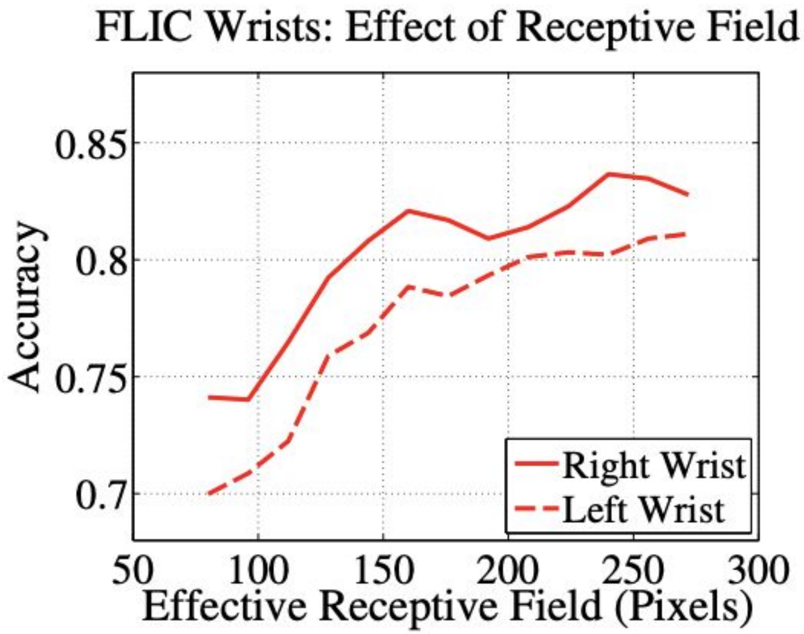
Convolutional Pose Machine

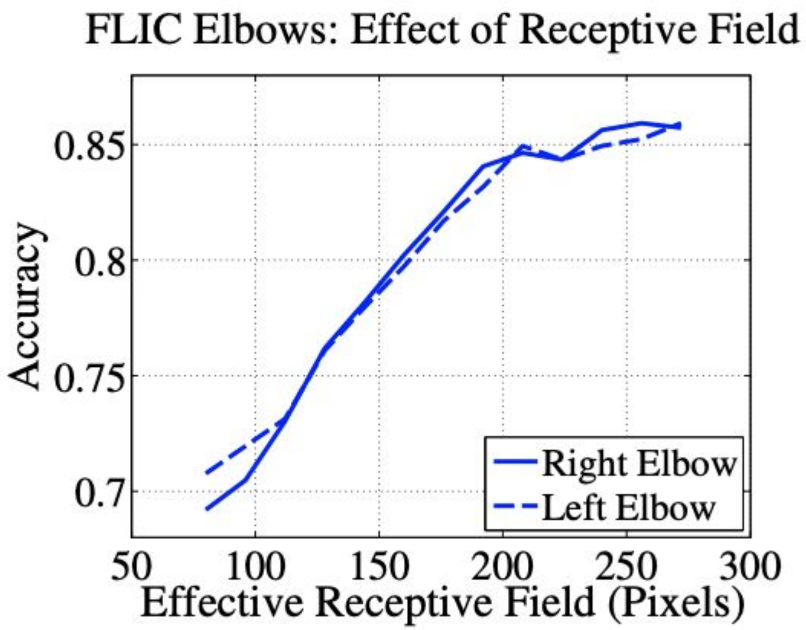
- Iterative refinement leads to elimination of noisy predictions and the discovery of missed body parts



Results

The localization accuracy increases with a larger receptive field



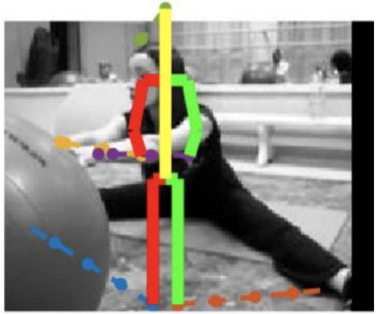


Summary: convolutional pose machine

- CNN with iterative refinement
- Local evidences are weak
- Context does matter
  - Larger receptive field allows correction of mispredictions
  - Evidences of the other body parts reduces ambiguities in localization
- Would be there other types of iterative refinement?

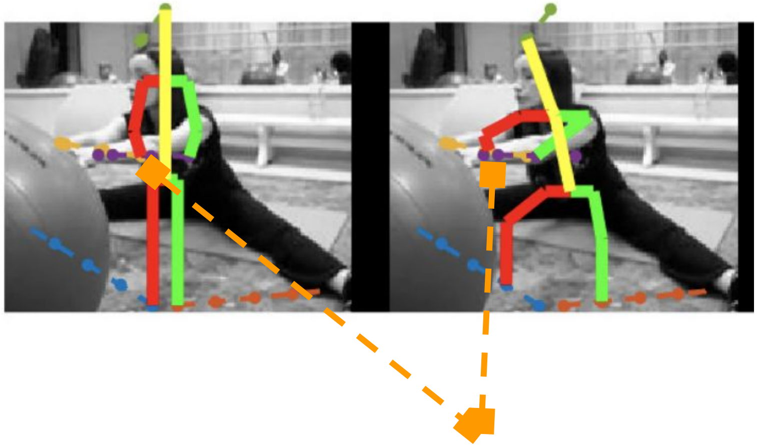
Iterative update of prediction

- Convolutional pose machine produces refined prediction every step
- Another idea: directly learns a **refinement procedure**



Iterative update of prediction

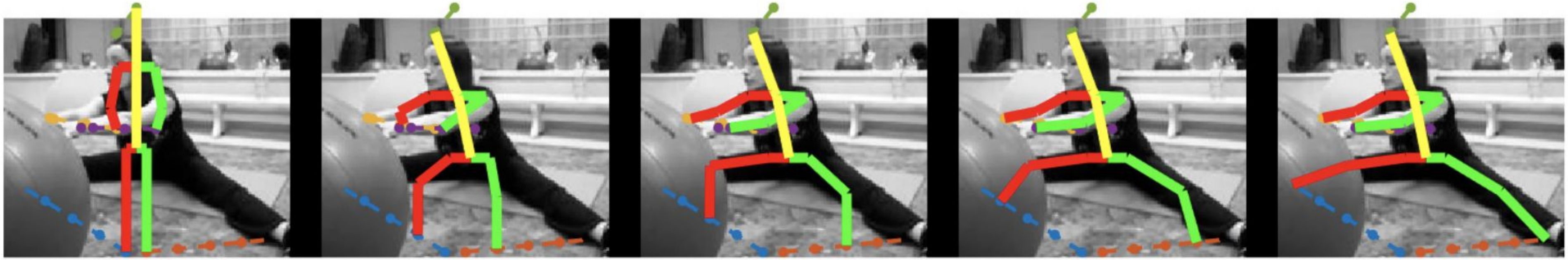
- Convolutional pose machine produces refined prediction every step
- Another idea: directly learns a **refinement procedure**



Move the body parts within σ towards actual location

Iterative update of prediction

- Convolutional pose machine produces refined prediction every step
- Another idea: directly learns a **refinement procedure**



Iteratively refine the pose over T steps!

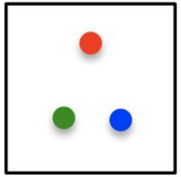
Iterative Error Feedback

- Iteratively predict a correction (error feedback) to refine the localization

I

Input image

y_0

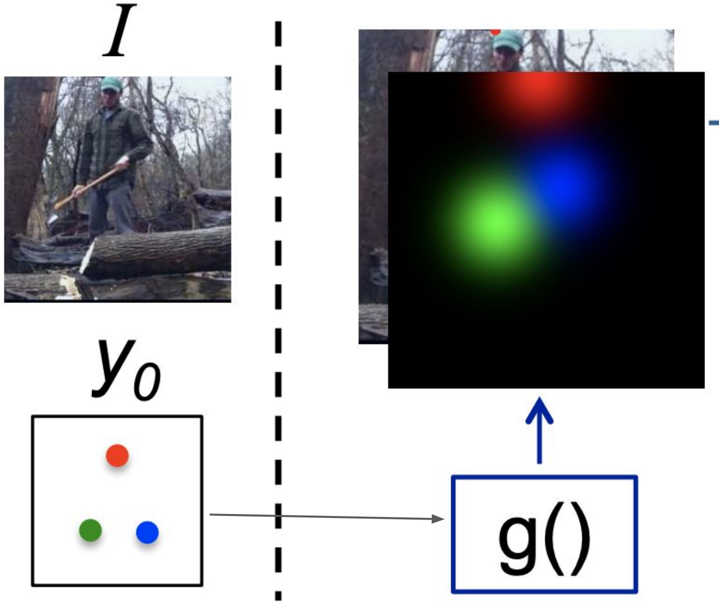


Initial body part predictions (K=3)

Carreira et al.,Human Pose Estimation with Iterative Error Feedback, In CVPR, 2016

Iterative Error Feedback

- Iteratively predict a correction (error feedback) to refine the localization



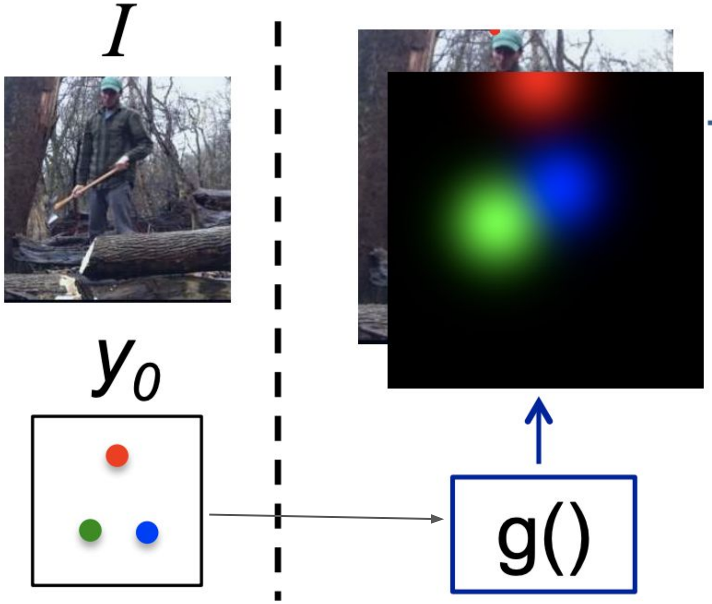
Concatenate the predicted pose and image

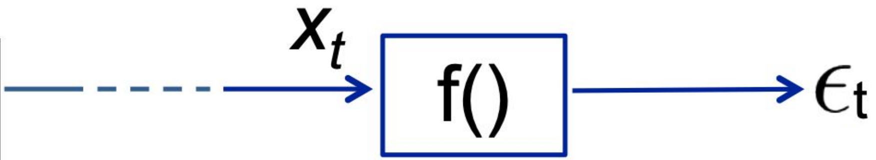
- **g(y)**: convert the predicted pose **y** (vector) to heat maps

Carreira et al.,Human Pose Estimation with Iterative Error Feedback, In CVPR, 2016

Iterative Error Feedback

- Iteratively predict a correction (error feedback) to refine the localization





Predict **displacement** ϵ of joints that corrects the misprediction

- **f(x)**: function that produces the error for each body part

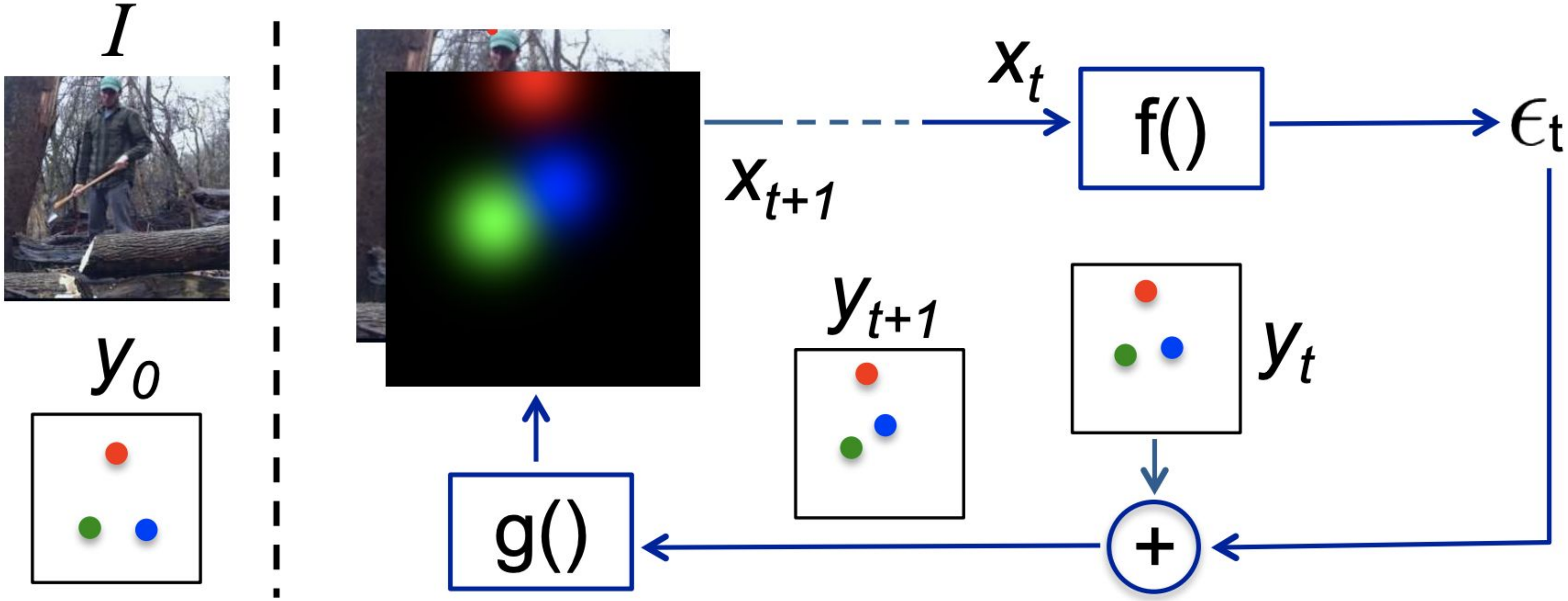
$$||y^* - y_0|| > ||y^* - (y_0 + \epsilon)||$$

- **y***: ground-truth part locations

Carreira et al.,Human Pose Estimation with Iterative Error Feedback, In CVPR, 2016

Iterative Error Feedback

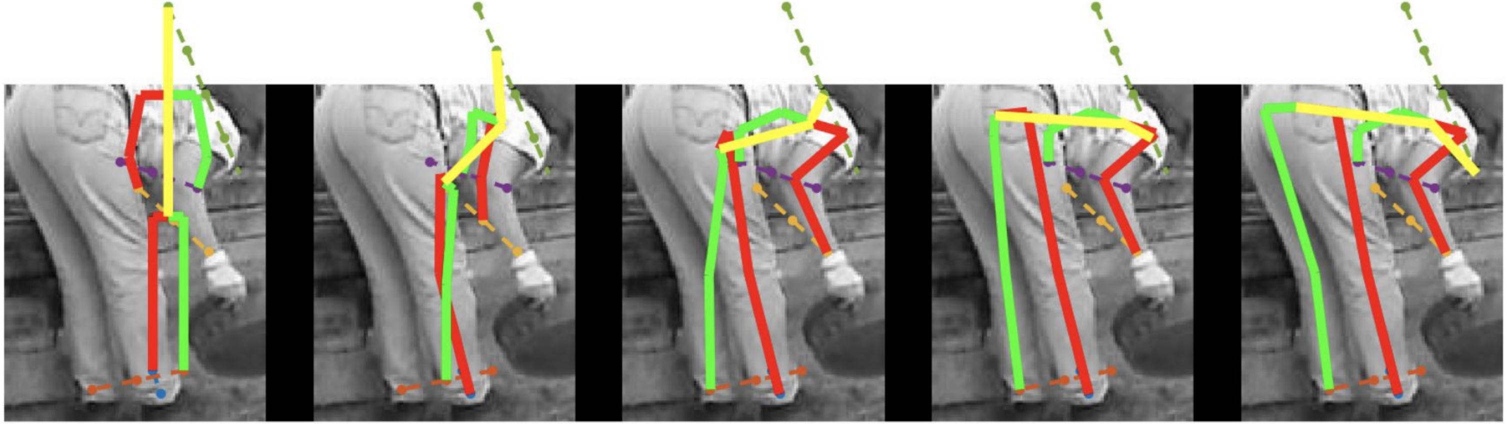
- Iteratively predict a correction (error feedback) to refine the localization

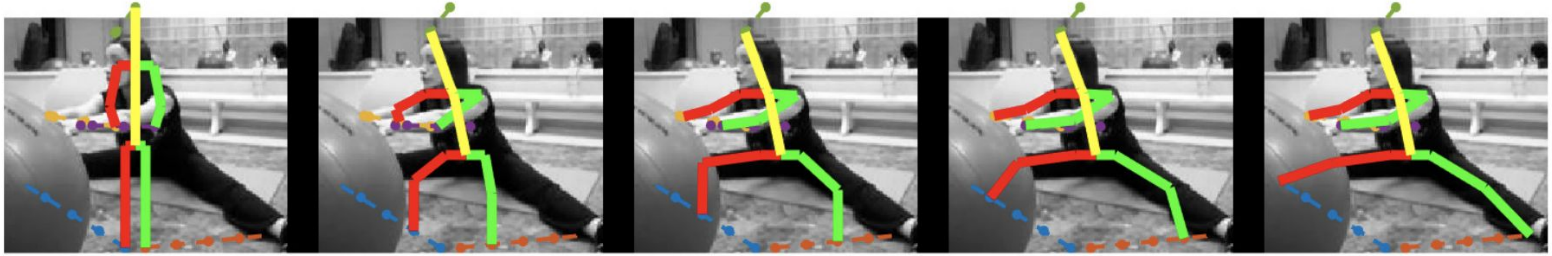


Carreira et al.,Human Pose Estimation with Iterative Error Feedback, In CVPR, 2016

Results

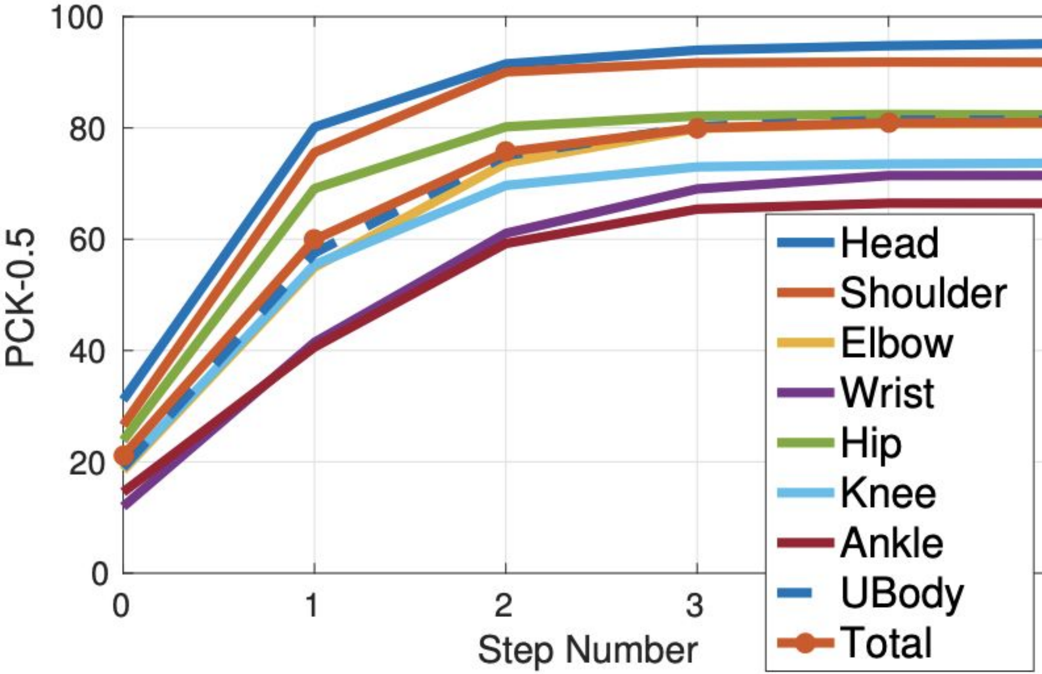
- The pose is refined through error-feedback.





Results

- The pose is refined through error-feedback.

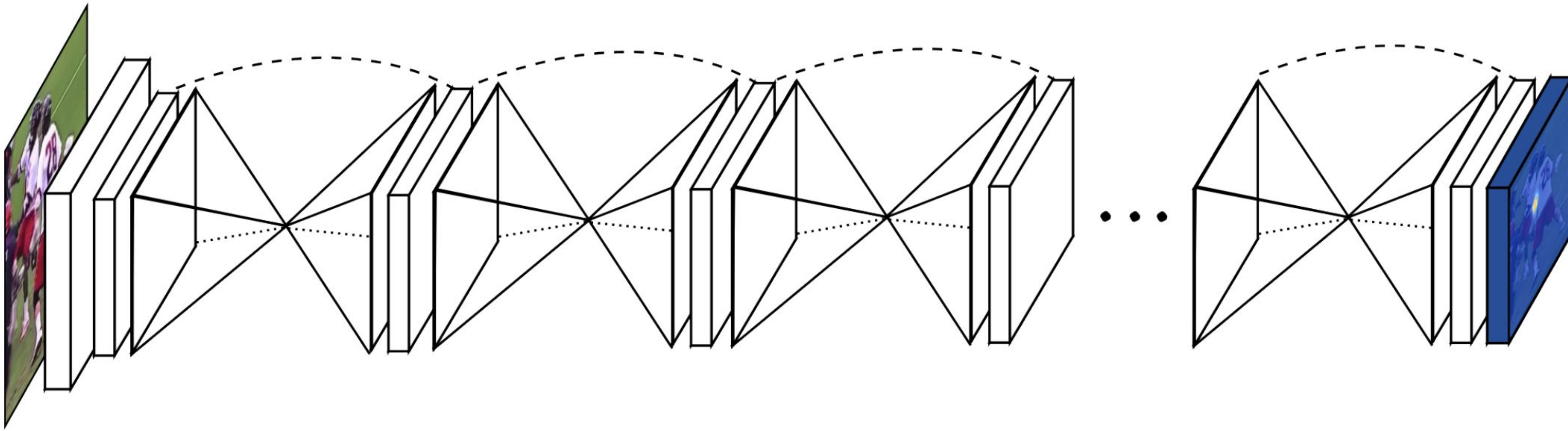


Summary: incorporating context

- So far, we learned that
    - Local evidences are sometimes ambiguous
    - incorporating context improves the accuracy
    - Iterative refinement improves the accuracy
- On the other hand,
    - Too large receptive field damages the localization accuracy
    - Iterative updates is computationally expensive
    - Would there other way to incorporate multi-scale prediction?

Combining local and global cues

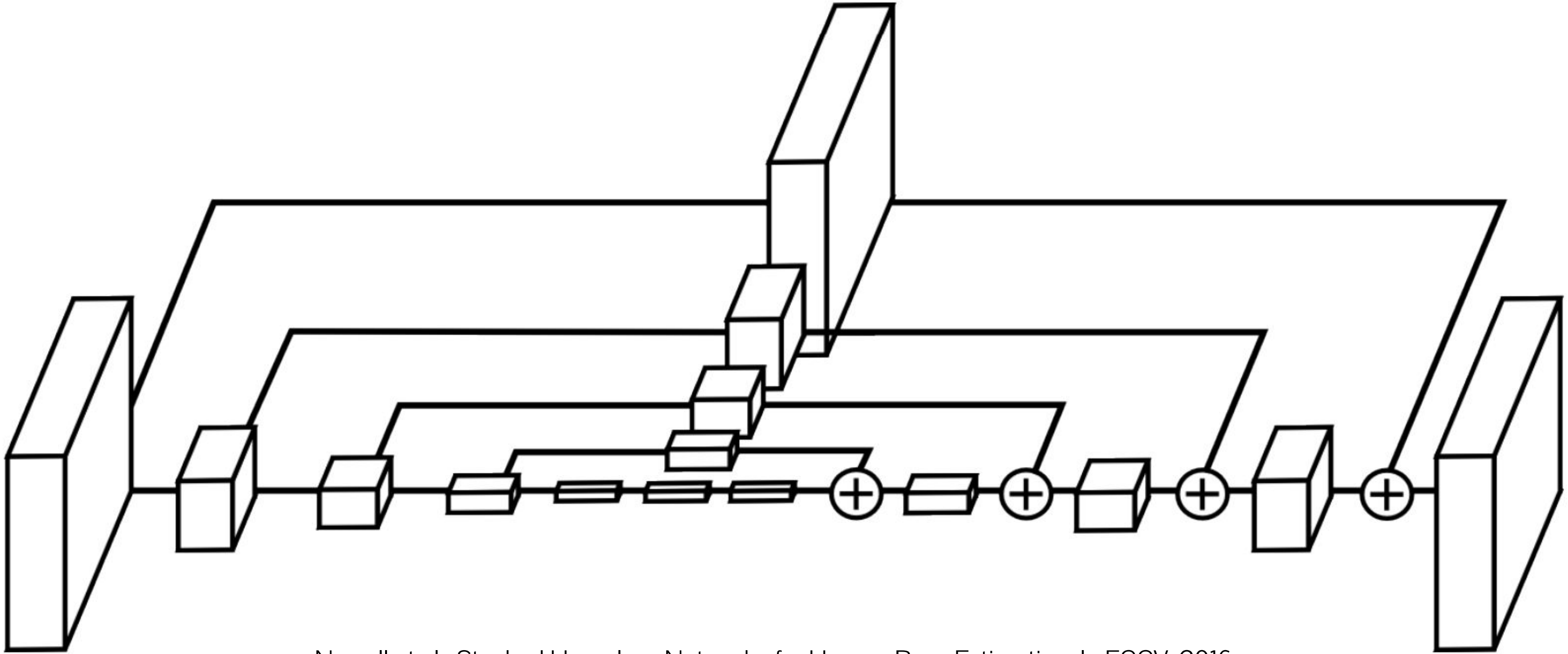
- Incorporating both global and local observations into prediction
- CNN architecture that combines both cues



Newell et al., Stacked Hourglass Networks for Human Pose Estimation, In ECCV, 2016

Stacked Hourglass

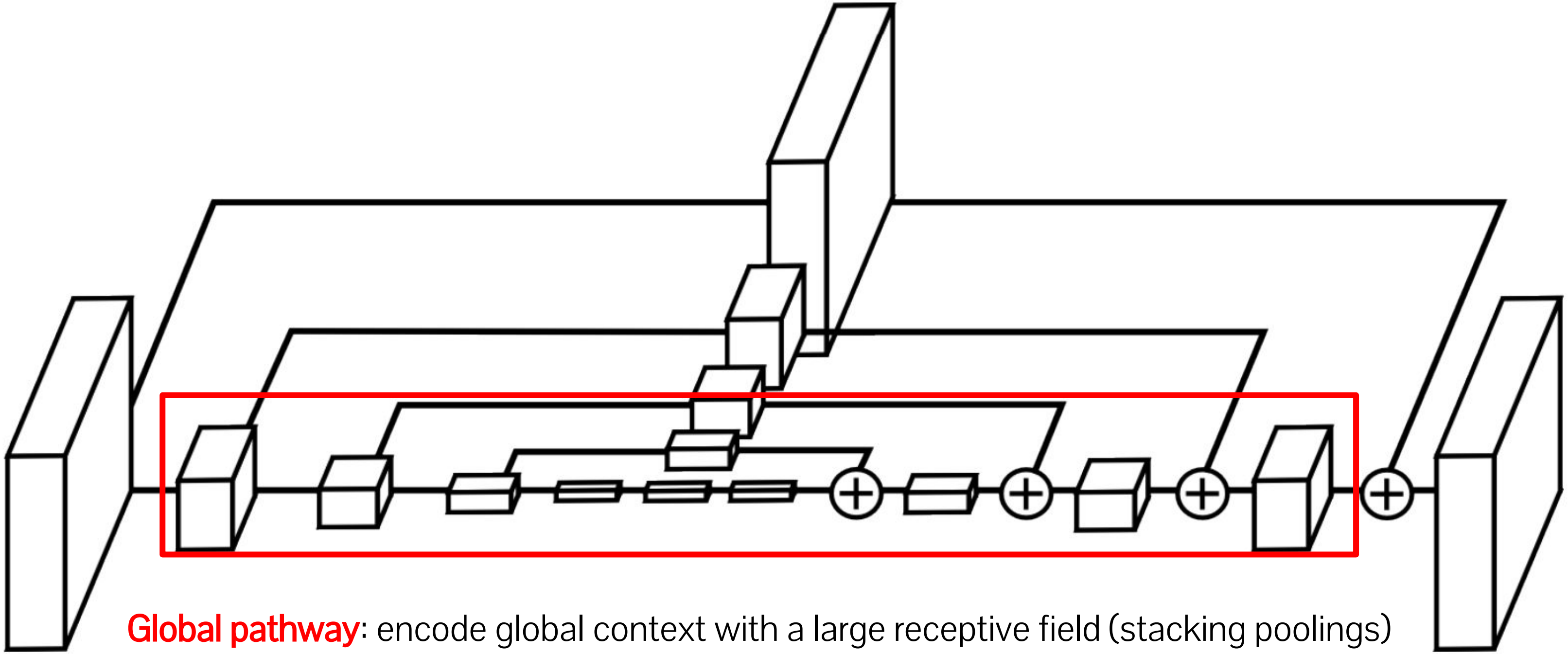
- Hourglass module for multi-scale encoding



Newell et al., Stacked Hourglass Networks for Human Pose Estimation, In ECCV, 2016

Stacked Hourglass

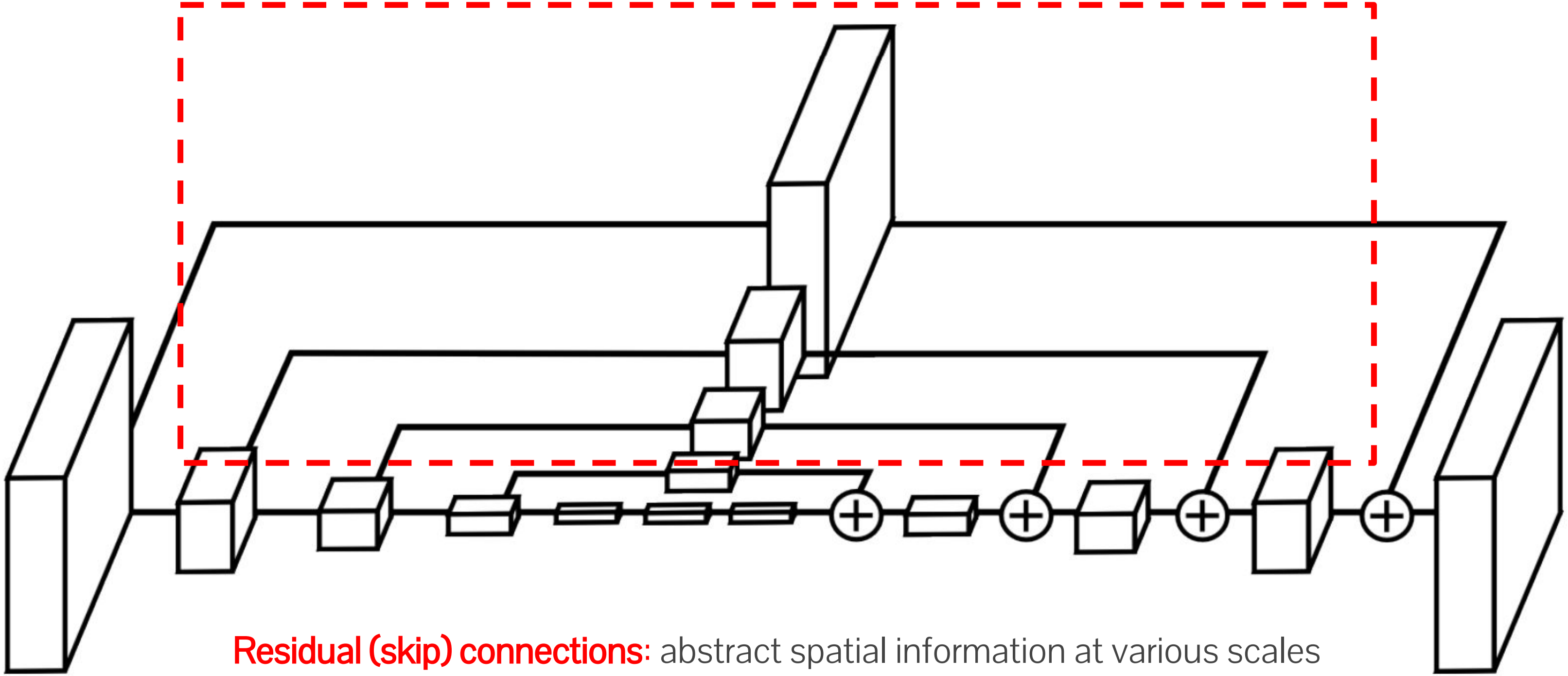
- Hourglass module for multi-scale encoding



Global pathway: encode global context with a large receptive field (stacking poolings)

Stacked Hourglass

- Hourglass module for multi-scale encoding



Residual (skip) connections: abstract spatial information at various scales

Results

- Simple approach, outstanding performance

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al. [16], CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al. [19], CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al. [17], CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al. [27], CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al. [18], CVPR'16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	**98.2**	**96.3**	**91.2**	**87.1**	**90.1**	**87.4**	**83.6**	**90.9**

Table 2. Results on MPII Human Pose (PCKh@0.5)

Summary: stacked hourglass

- Architecture design for incorporating contexts at multiple scales
- Substantial improvement over existing works