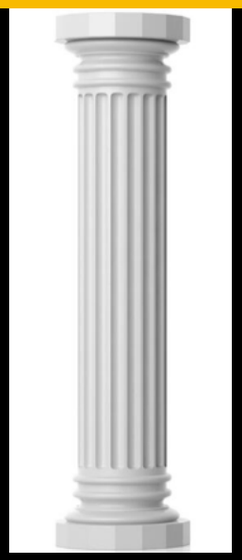
Supervised Learning: Pose Estimation

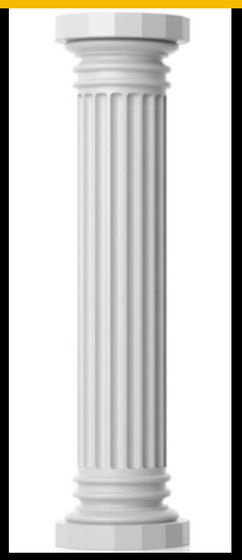
Earl Wong

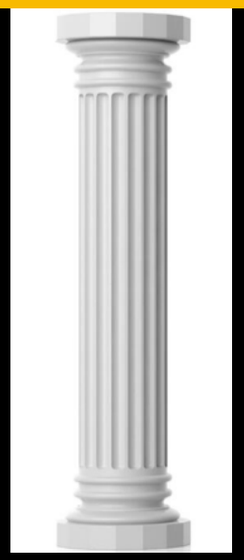
Labeled Dataset

Algorithm

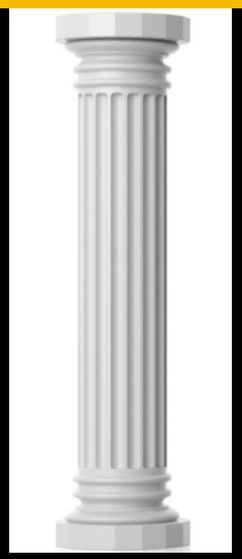
Loss Function







Labeled Dataset



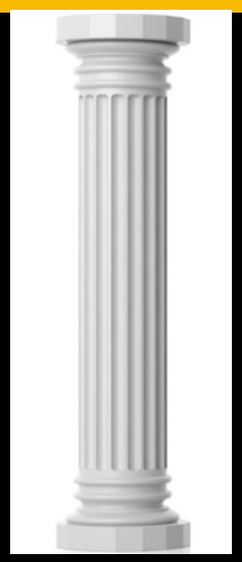
Labeled Dataset

Dataset	LSP Extended	JHMDB	MPII Single Human Pose	Poselet
Poses	10K	31K	26K	153K
Activities	11 (sports)	21	800->(revised to 491)	Diverse

To increase the difficulty of a dataset, more activities were added.

In addition, images containing challenging body poses, different torso viewpoints / torso rotations, individuals dressed in loose clothing and background clutter were also added.

Algorithm

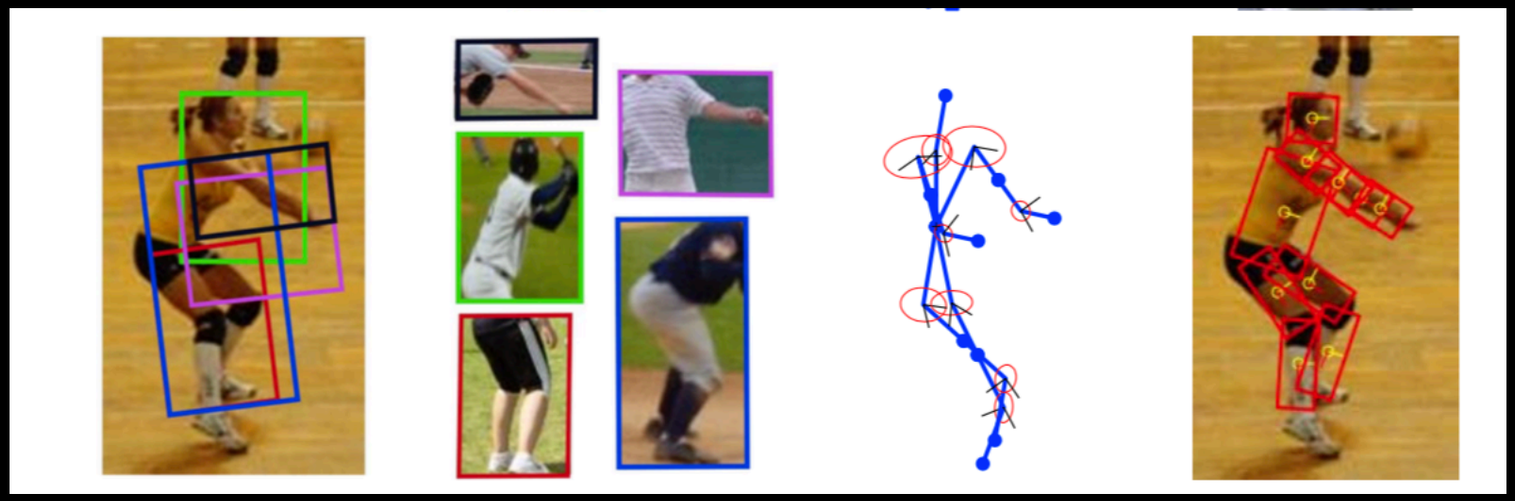


Historical

In the pre-AlexNet era, the state of the art in pose estimation was:

Poselet Conditioned Pictorial Structures, Pischulin et. al., CVPR 2013

Poselets



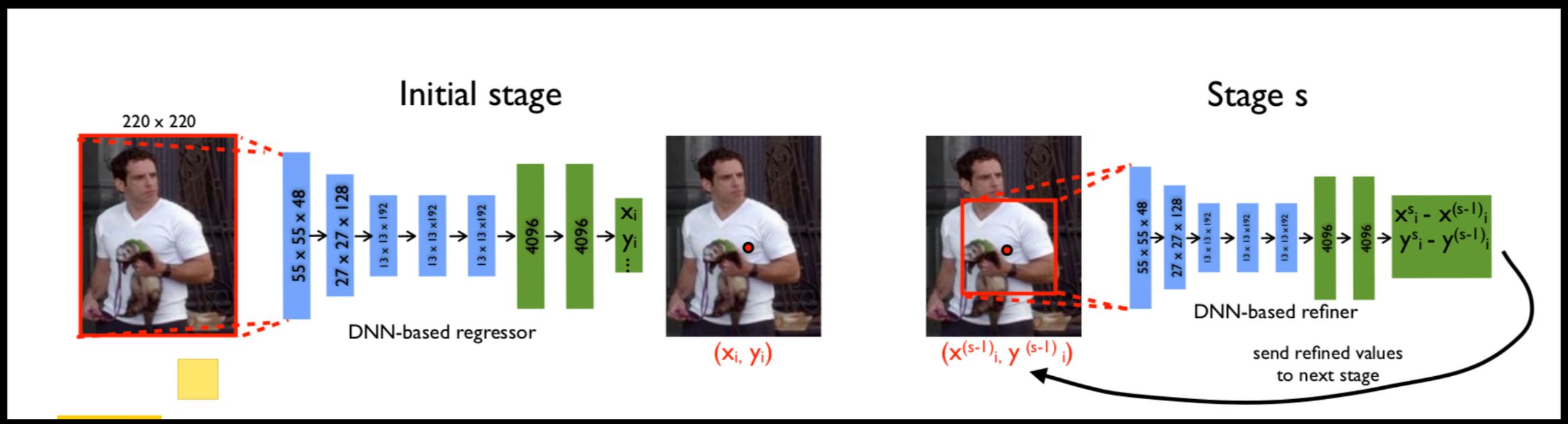
Poselet Detections | **Associated Poselet Cluster Models** | **PS Model** | **Improved Outcome**

1) A standard Pictorial Structures (PS) model captured the rigid, pairwise relationship between adjacent body parts. (i.e. The left wrist was connected to the left elbow. The left elbow was connected to the left shoulder, etc.)
2) A tree based model was then inferred from this information.
3) However, important higher order dependencies also existed between non-adjacent body parts. (i.e. The left wrist, the left elbow and the left shoulder might form a specific pose for a given image.)
4) When this mid-level information was captured in the new model (poselet = the captured configuration of multiple body parts jointly), a more robust tree model was inferred. (i.e. The left wrist was also connected to the left shoulder.)

Deep Network Era

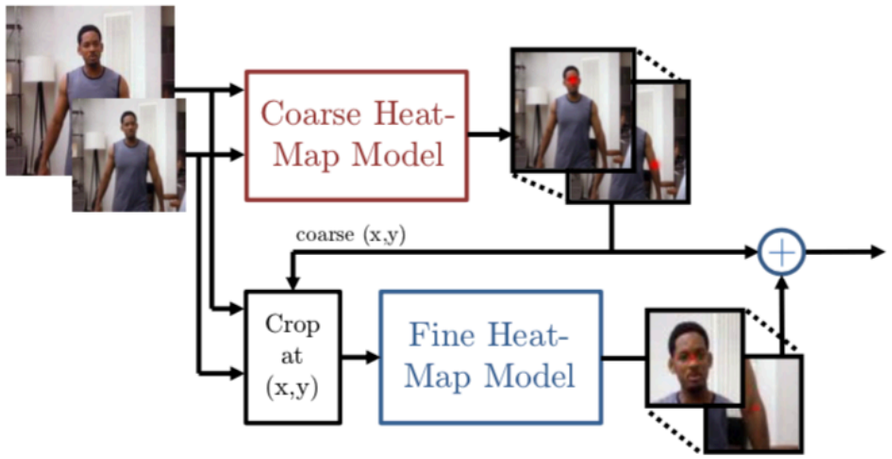
- DeepPose
- Efficient Object Localization using CNNs
- Stacked Hourglass Networks
- DeepCut
- Part Affinity Fields
- Toward Accurate Multi-person Pose Estimation In the Wild
- Deep High Resolution Representation Learning for Human Pose Estimation

DeepPose

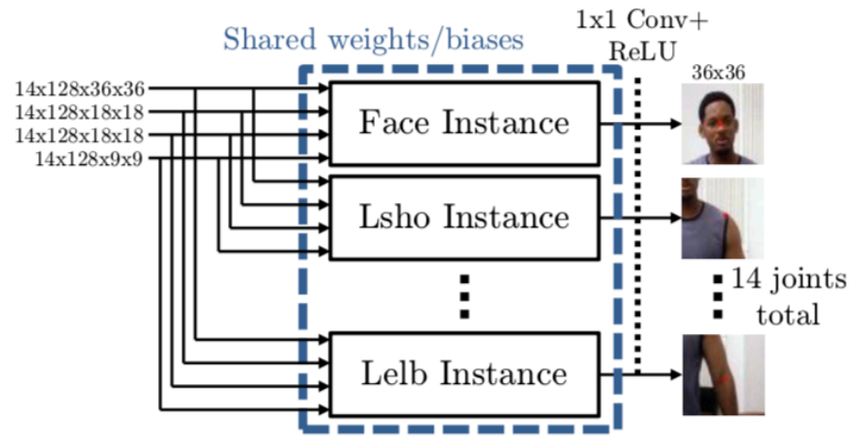


1) Two CNN's were trained.
2) The first CNN acted as a coarse regressor, outputting the initial coordinates of the various body parts.
3) A cropped region around the body part was then taken.
4) The cropped region was fed into a second CNN that refined the accuracy of the original coordinates (x_head, y_head), etc.

Efficient Object Localization Using CNNs



Typical Coarse to Fine Architecture strategy used at that time

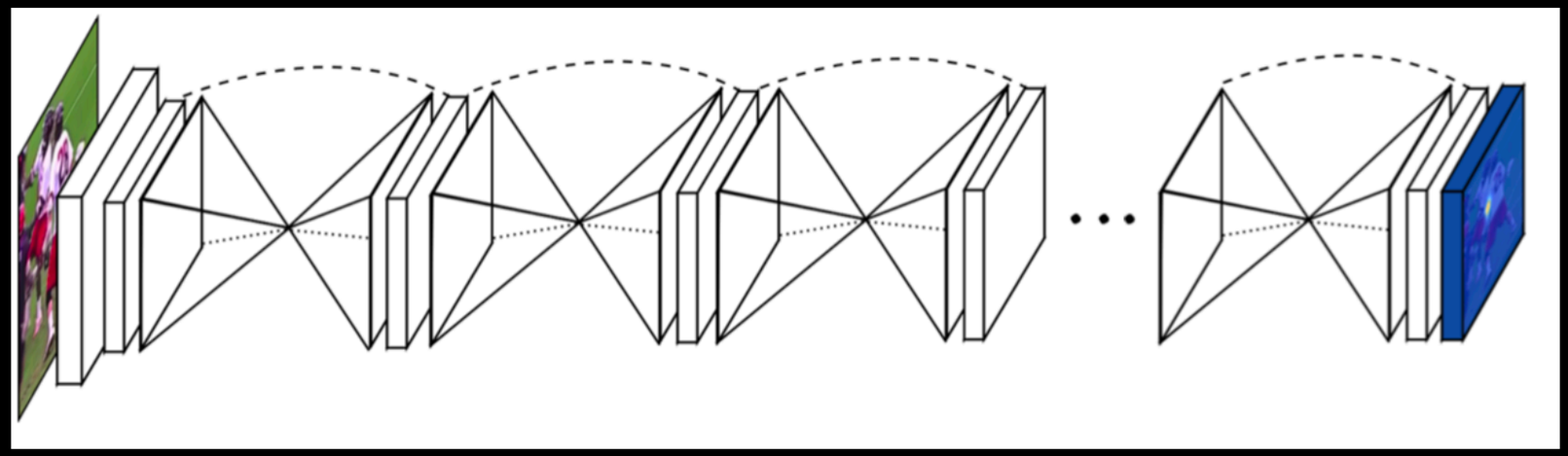


14 siamese CNN's associated with the fine heat map generation

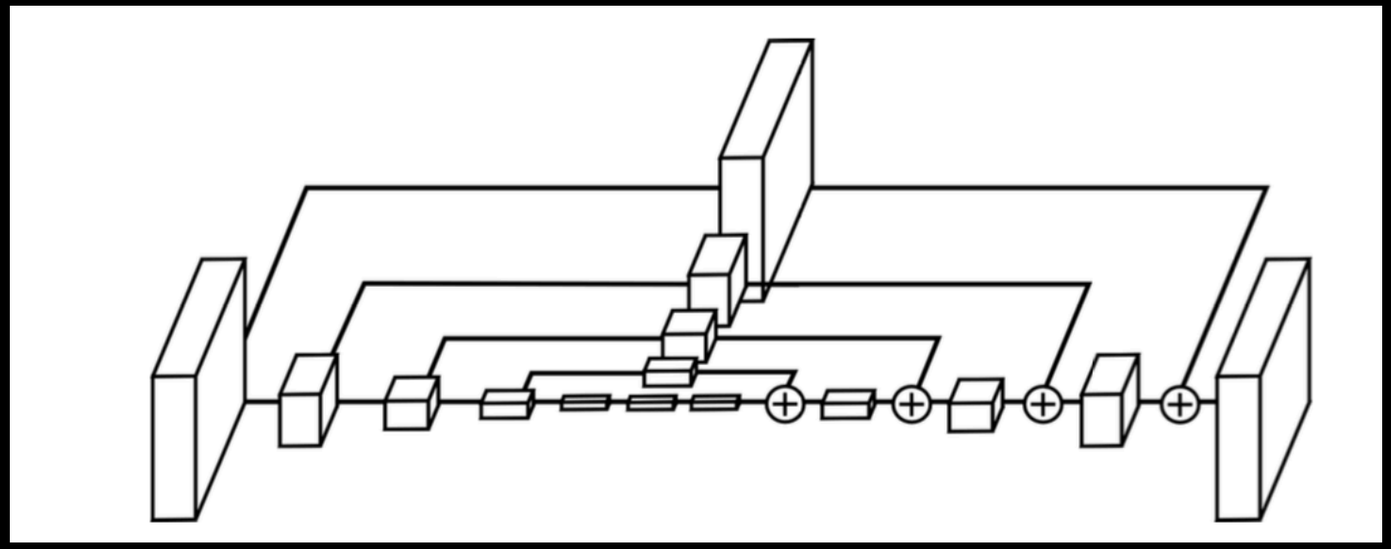
1) What happens if you have multiple people in a scene? The regression approach fails.
2) Hence, the previous 28 element regression vector was replaced by 14 MxN heat maps.
3) First, multi-resolution input images were fed into a coarse CNN to generate the coarse heat maps. (Note: Resolution acuity is compromised from max pooling.)
4) Next, using the information in the 14 coarse heat maps, region crops were taken and fed into a second CNN to generate finer resolution heat maps.
5) Since there were 14 heat maps (for 14 joints), 14 siamese CNN's were used to generate the finer heat maps.

Stacked Hourglass Network

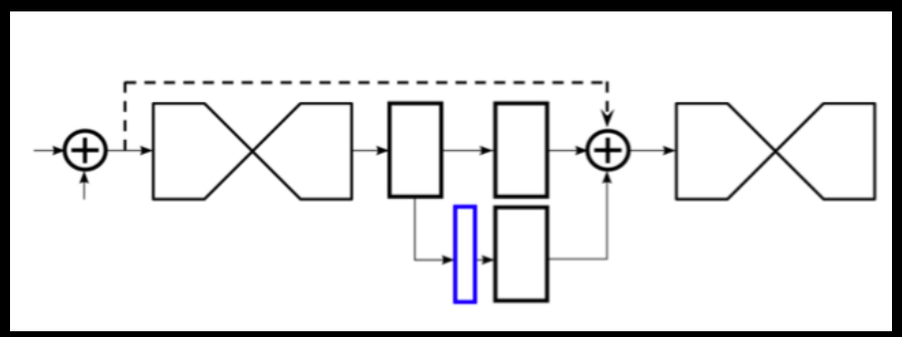
1)



2)



3)



Stacked Hourglass Network

- 1) Consecutive Encoder - Decoder networks were stacked to “enhance” the pose estimation accuracy.

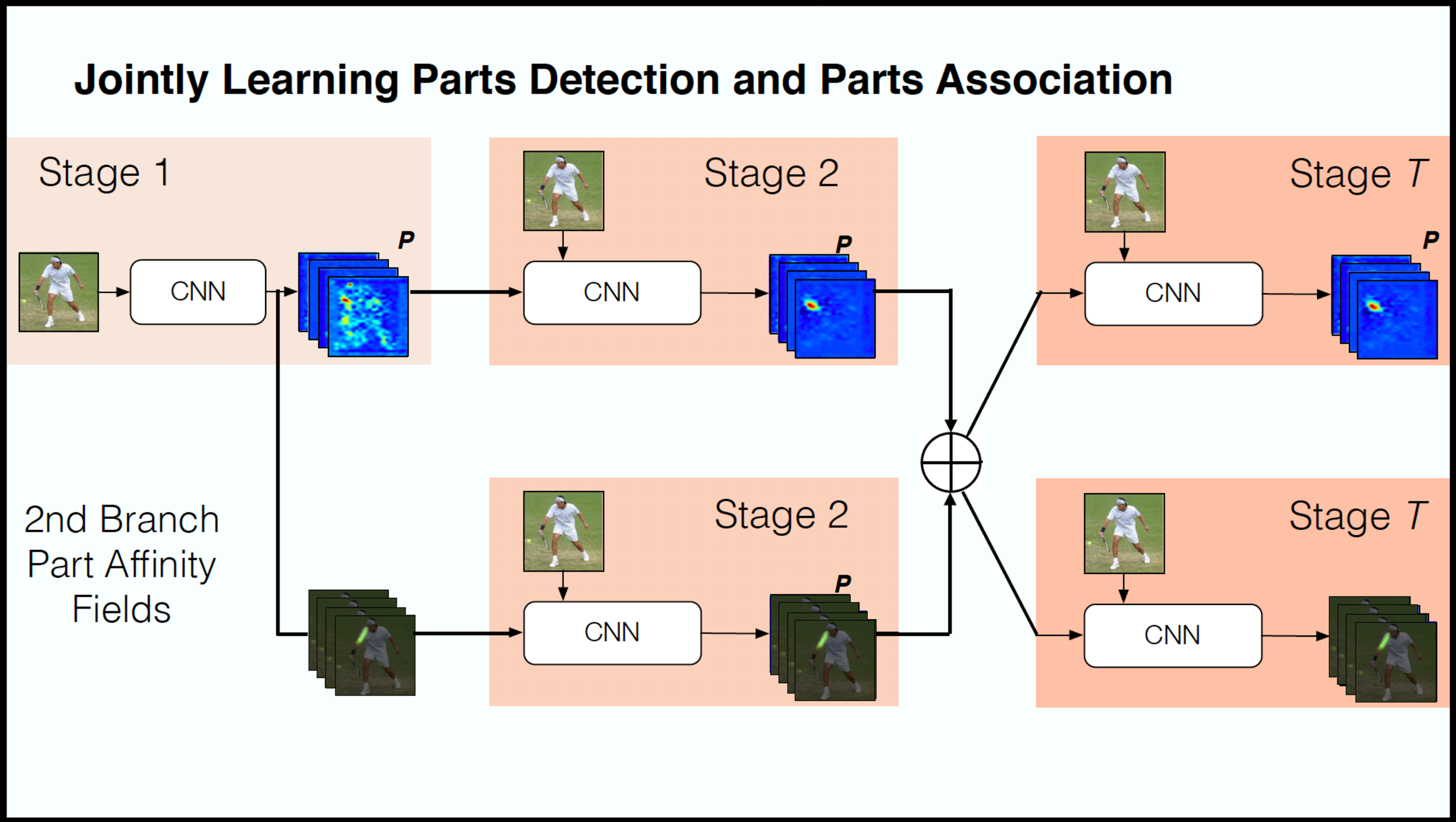
- Enhanced accuracy was accentuated by training the network using “intermediate supervision” instead of end to end back propagation. i.e. The entire network was “tuned” in (hourglass) stages.

- 2) Within each Encoder - Decoder block, the corresponding encoder and decoder layers were connected using residual module blocks. (See the paper for the intricate details.)

- 3) Consecutive Hourglass networks were then connected as shown. (The heat map block is shown in blue.)

  The authors claimed that this network helped capture information at every scale.

Part Affinity Fields



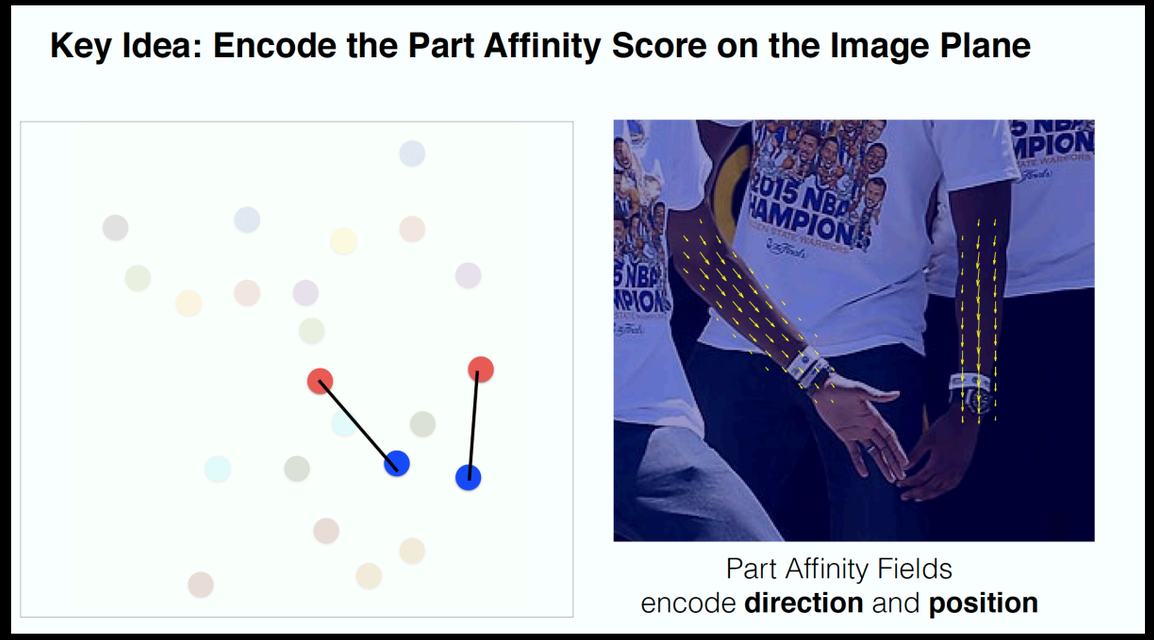
Top branch: CPM body parts estimator. Bottom branch: Part Affinity Fields estimator

Part Affinity Fields

A convolutional pose machine (CPM) consisting of a sequence of CNN's was used to produce 2D belief maps of body part locations.

Each stage of the CPM receives belief map inputs from the previous stage, refining the previous results.

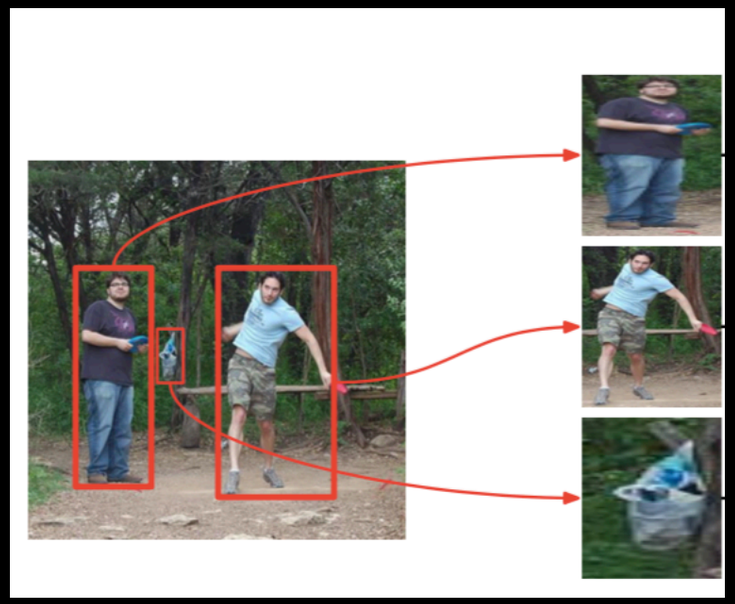
Part Affinity Fields then takes matters one step further, by determining the part to part associations (position and direction) between the body parts of different individuals.



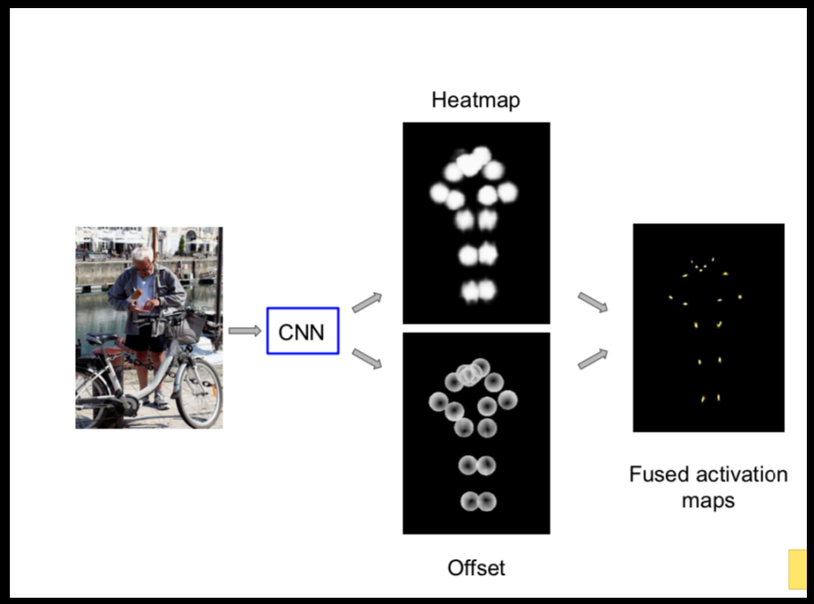
The computed results were then passed to additional stages for enhanced processing.

Using this bottoms up approach, the pose estimation problem was addressed for multiple individuals in a scene in real time.

Toward Accurate Multi-Person Pose Estimation in the Wild



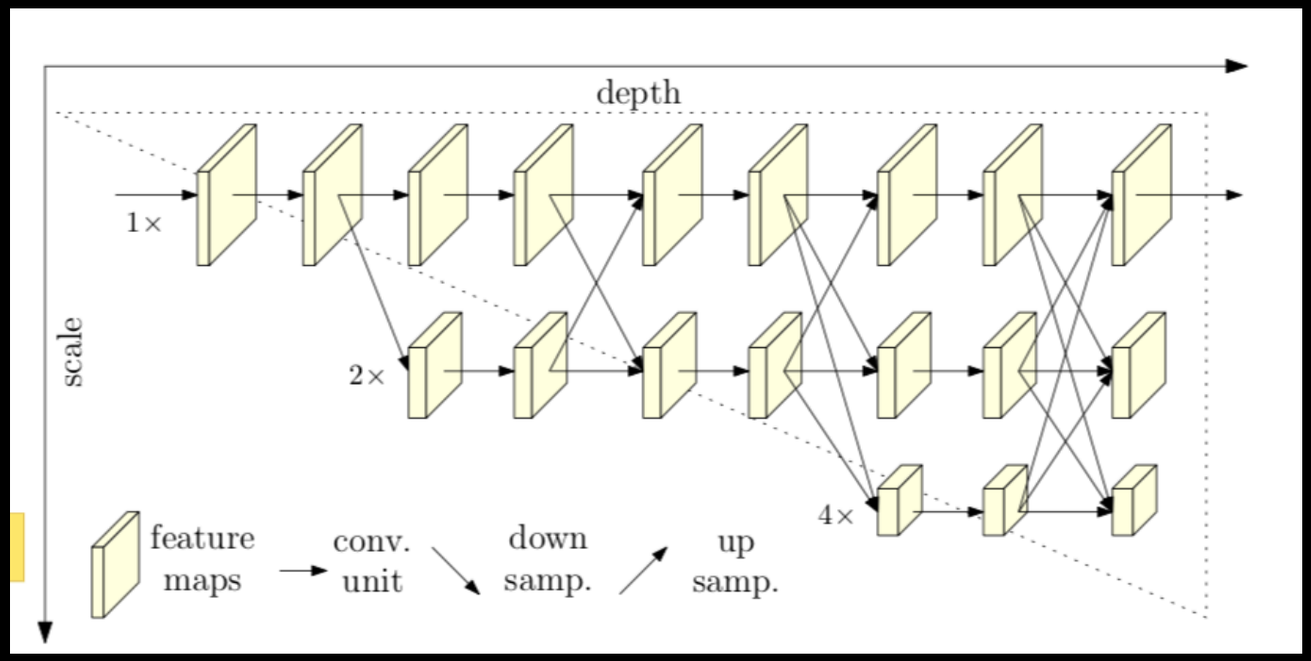
Bounding Box Detector

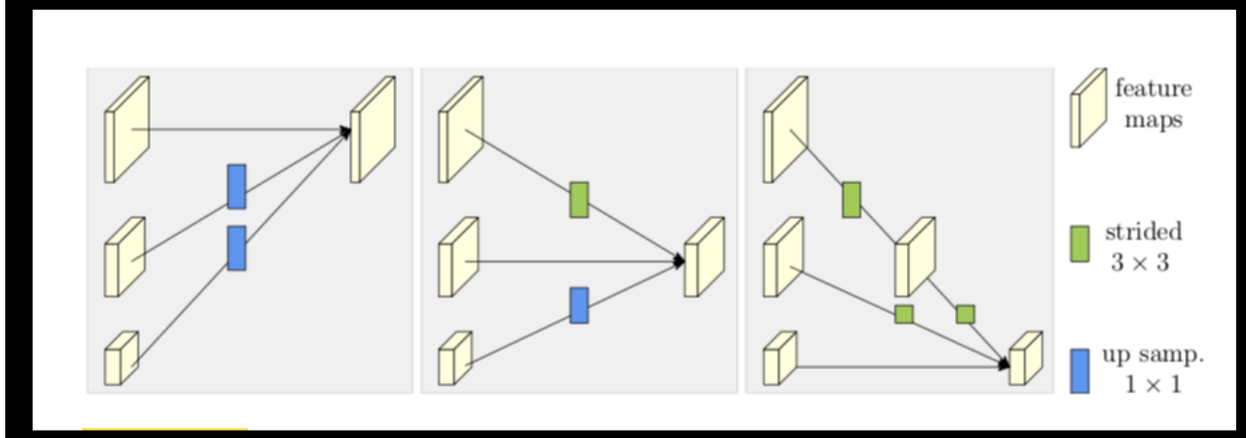


Generate HeapMap and associated coordinate refinement

1) This paper described a 2 stage process for pose estimation.
2) In stage 1, a CNN is used to perform bounding box object detection. (R-CNN)
3) The region of interest is then cropped and rescaled for input into a second CNN.
4) The second CNN produced heat maps for the different joints detections as well as offset heat maps for joint coordinate refinements.

Deep High Resolution Representation Learning for Human Pose Estimation





1) The authors proposed the following network (on the left), to produce high accuracy human pose estimation heat maps.
2) In the top path of the network, the resolution remains fixed at the original input resolution.
3) In downstream paths, the resolution is downsampled.
4) However, at various depth planes in the downstream paths, channels were re-combined with their appropriate resolution counterparts (via concatenation) using one of the three scaling techniques shown in the right image.

Algorithm

- A Vanilla CNN with regressor outputs.
- A Vanilla CNN's with heat map outputs.
- Stacked encoder-decoder networks that are trained with intermediate supervision.
- Two stage networks that first locate people using an object detector, and then produce the body parts heat maps for each individual.
- Bottoms up approach that first finds the body parts, and then determines their associations using directional information.
- Generate all possible resolution outputs from full to low. Then, recombine these outputs (via concatenation) back to a "super" full resolution representation. Then, produce heat maps from the now "super" full resolution output layer.

Comments

- Two different research paths exist for pose estimation: top down (start with the image, find the person and then find the body parts) versus bottoms up (start with a body part and then find other associated body parts).

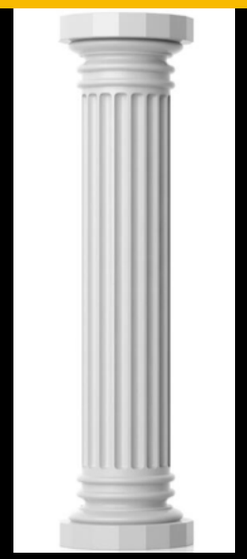
  Open questions and issues include:

- Q: What is the best way to generate the most accurate key points? By refinement? By intermediate supervision of a network comprised of the cascade of identical networks?

- Q: Is there a preferred approach, depending on the number of individuals in a scene?

- Q: How robust are these techniques in the wild? What is their brittleness?

Loss Function



Loss Function

- For heat map outputs, L1, L2 or cross entropy loss was applied.

- For regression / regression offset estimates, an L2 loss was applied.

References

- Poselet Conditioned Pictorial Structures, Pishchulin, et. al., CVPR 2013
- DeepPose: Human Pose Estimation via Deep Neural Networks, Toshev, et. al., December 2013
- Efficient Object Localization using Convolutional Neural Networks, Tompson, et. al., arXiv November 2014
- DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, arXiv November 2015
- Stacked Hourglass Networks for Human Pose Estimation, Newell, et. al., arXiv March 2016
- Real Time Multi Person 2D Pose Estimation using Part Affinity Fields, Cao, et. al., arXiv November 2016
- Toward Accurate Multi Person Pose Estimation In the Wild, Papandreou, et. al., arXiv January 2017
- Deep High Resolution Representation Learning for Human Pose Estimation, Wang, et. al. arXiv August 2019