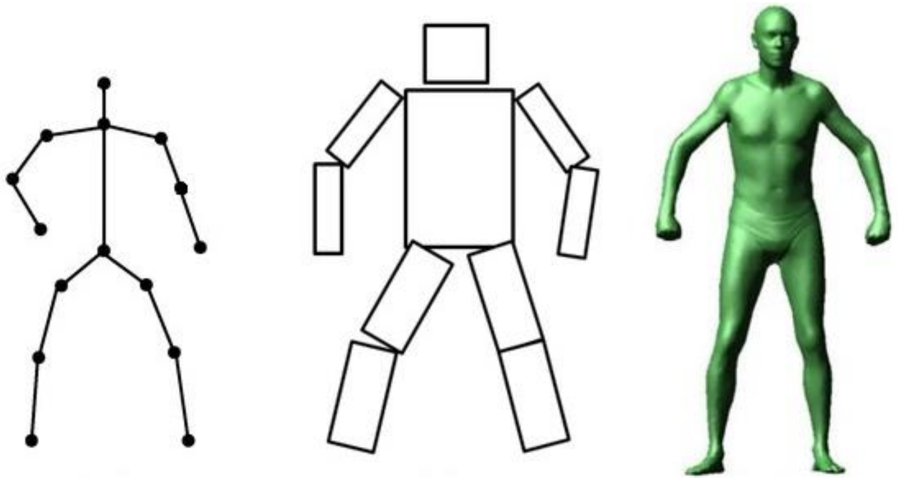
Human Pose Lecture

Dr. Sergio Escalera
University of Barcelona

What is articulated human pose?

- Given body kinematic tree, human pose is defined as the vector of joints locations either in 2D (image coordinate in pixels) or 3D (world coordinate in millimeters).
- Human body pose estimation is a way of representing humans in the images,
- It can be represented and estimated from coarse to fine models.
- Classicaly addressed by pictorial structures and graphical models



Kinematic model Cardboard model Volumetric model

Image credit: https://www.mdpi.com/1424-8220/16/12/1966/htm

Challenges

Image credit: https://www.slideshare.net/plutoyang/human-pose-estimation-by-deep-learning

Human pose regression

- The output of the network is directly a vector of (x,y) joints locations,
- This is a difficult task for the network because
  - Pose vector is a highly nonlinear variable,
  - Network must deal with scale and translation as well.

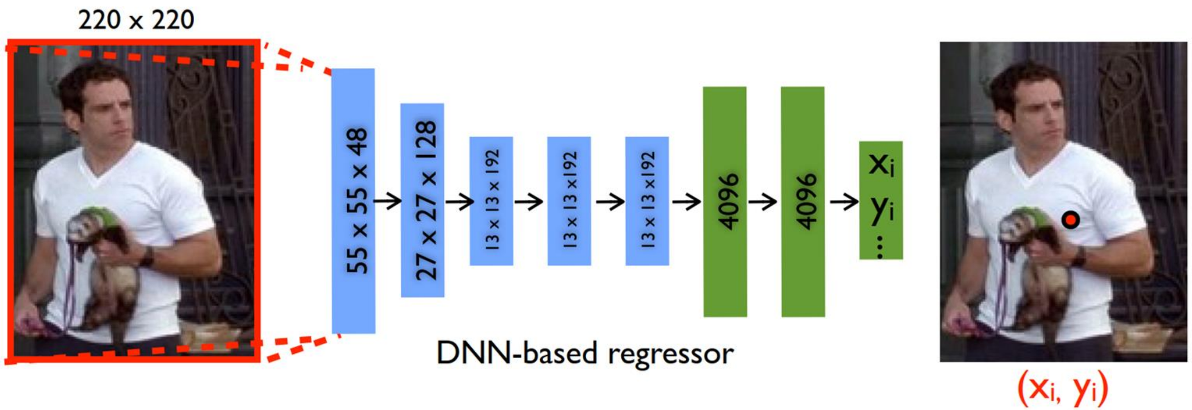


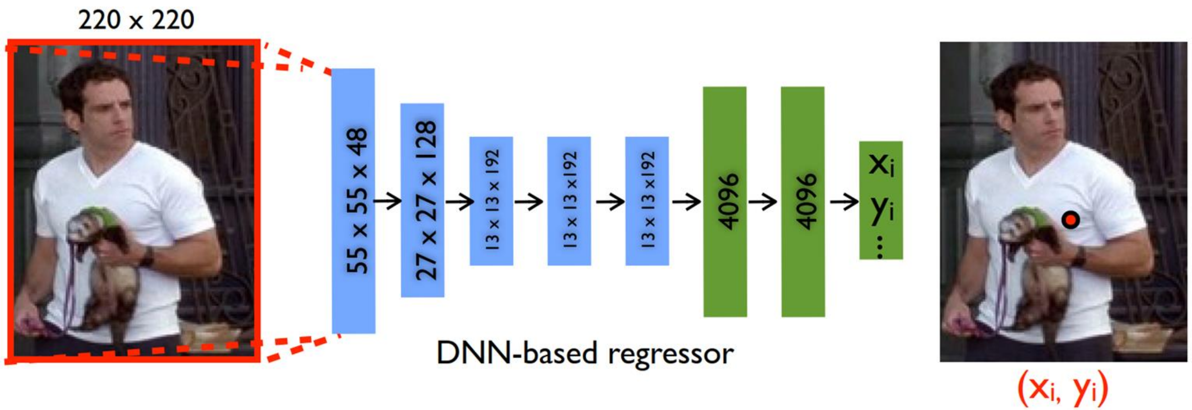
Image credit: https://arxiv.org/pdf/1312.4659.pdf

2D pose estimation

Human pose regression

- The output of the network is directly a vector of (x,y) joints locations,
- This is a difficult task for the network because ⇦ **A better alternative:** a dense probability map for each joint.
    - Pose vector is a highly nonlinear variable, ⇦ **Solution:** cascade of pose regressors.
    - Network must deal with scale and translation as well.

⇨ **Solution:** crop person bounding box as pre-processing.



Loss: L1 or L2

Architecture: AlexNet

Image credit: https://arxiv.org/pdf/1312.4659.pdf

Cascade of pose regressors

- Given an initial estimation of the joints (in holistic view), iteratively refines the error (in local view).

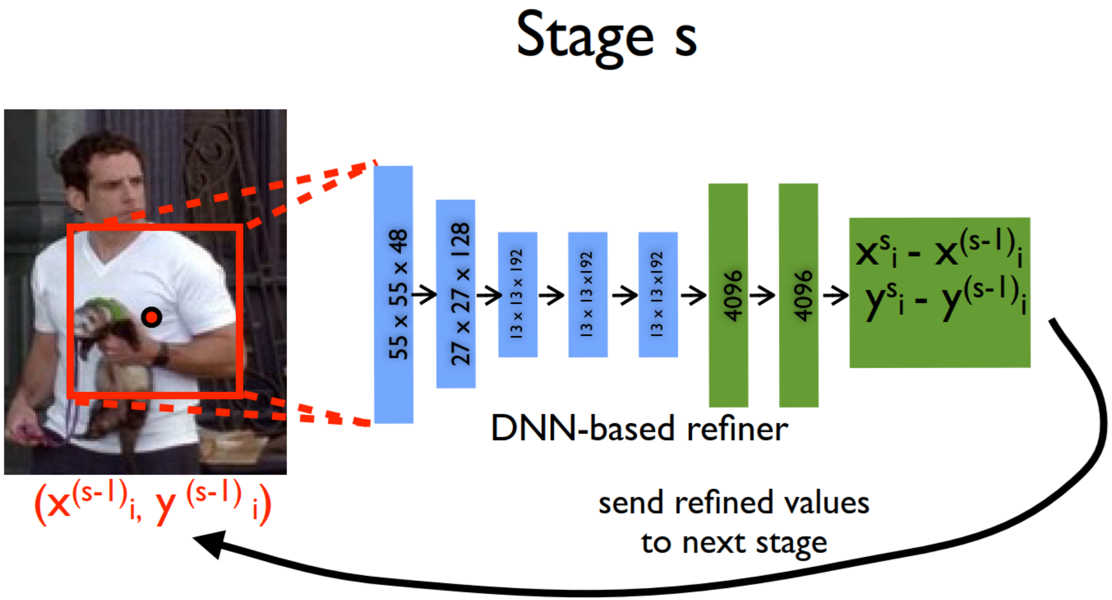


Image credit: https://arxiv.org/pdf/1312.4659.pdf

Cascade of pose regressors

Limitations:

- Sensitive to initial estimation,
- Local solution, easily can stick to local minima,
- Lack of structured predictions.

Image credit: https://arxiv.org/pdf/1312.4659.pdf

Results on LSP dataset

Image credit: https://arxiv.org/pdf/1312.4659.pdf

Joints heatmaps instead of pose vector

- It is an easier problem for the network,
- Still can be combined with pose regressors,
- The map can be fed into graphical models to learn higher order joint relationships.

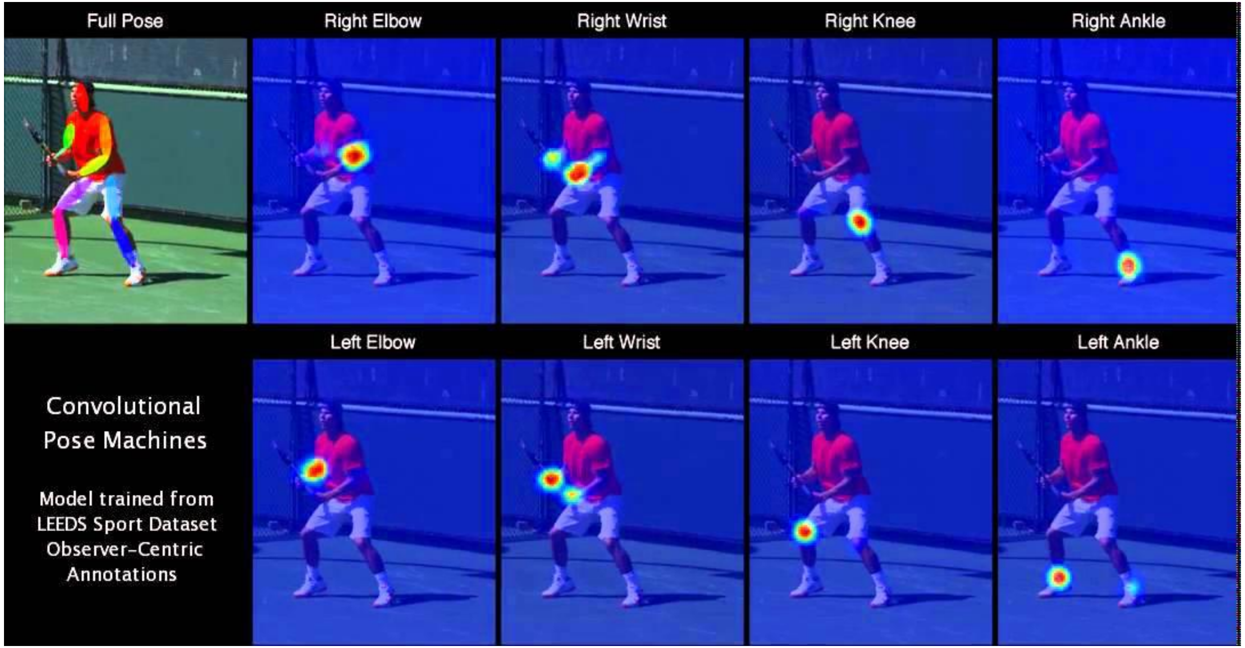


Image credit: https://www.youtube.com/watch?v=hcnBHWo2RGw

Context is important

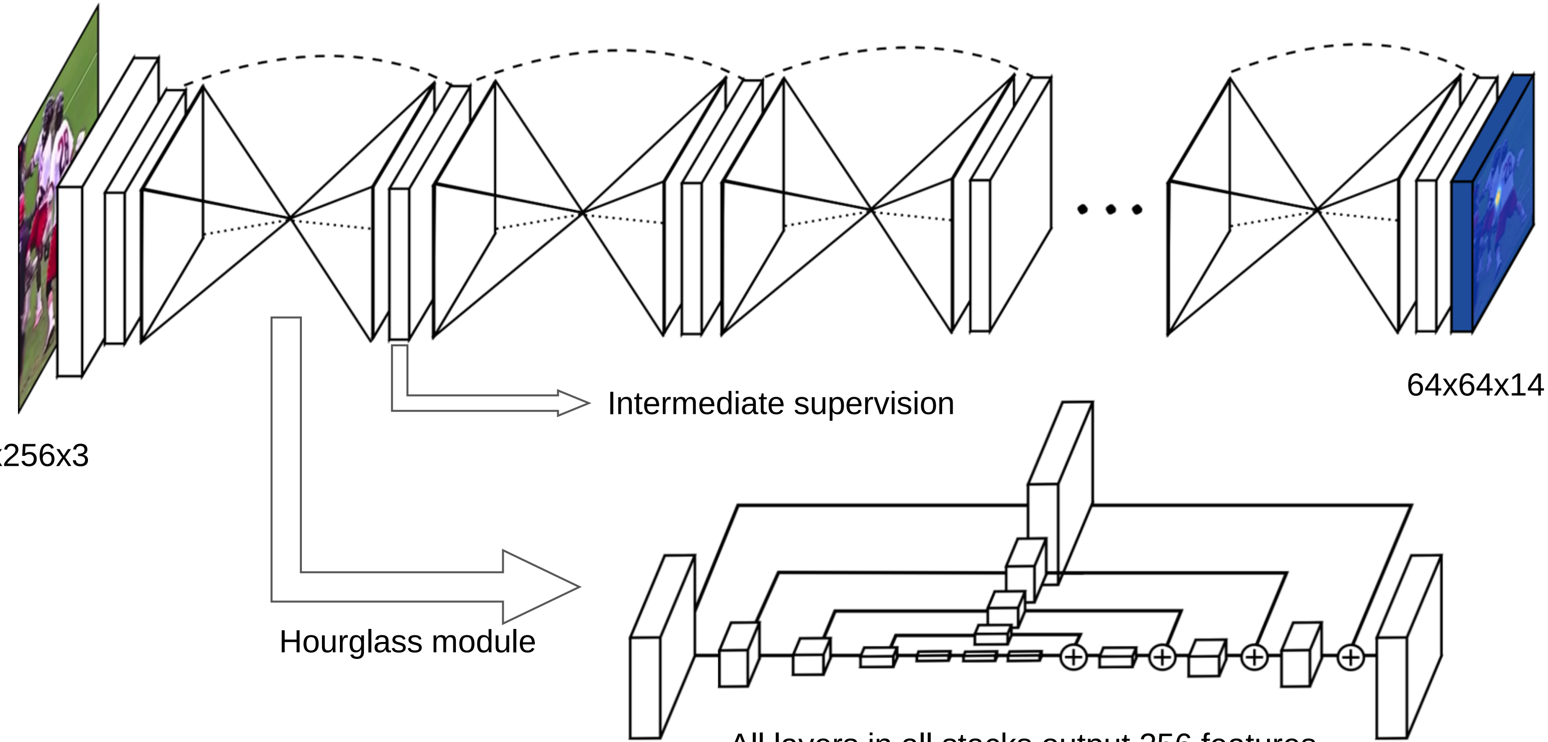
Which part belongs to a human body?

Context is important

Which part belongs to a human body?

- Local evidence is weak,
- Larger receptive field = more context,
- Recover from failures by cascading.

Stacked Hourglass Network



All layers in all stacks output 256 features.

Image credit: https://arxiv.org/pdf/1603.06937.pdf

Stacked Hourglass Network: intermediate supervision

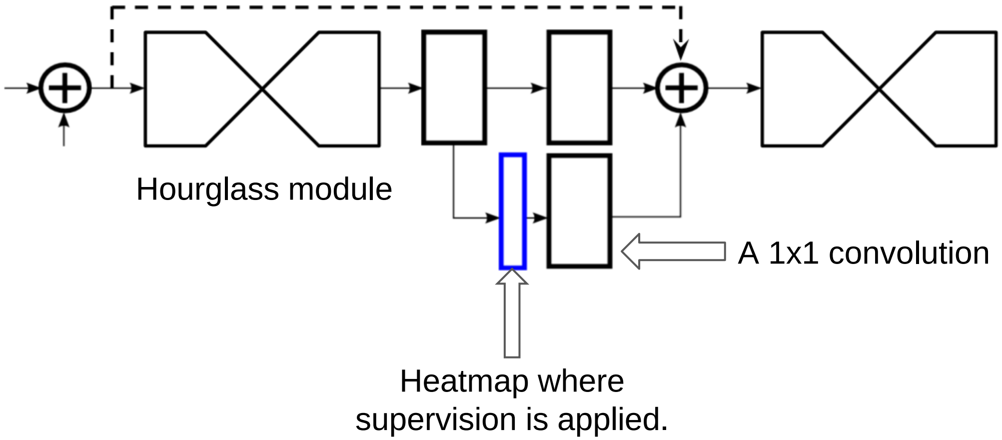


Image credit: https://arxiv.org/pdf/1603.06937.pdf

Stacked Hourglass Network

Cons:

- ➢ Quite heavy in GPU,
- ➢ Joints may be confused with background,
- ➢ Still does not explicitly deal with structure.

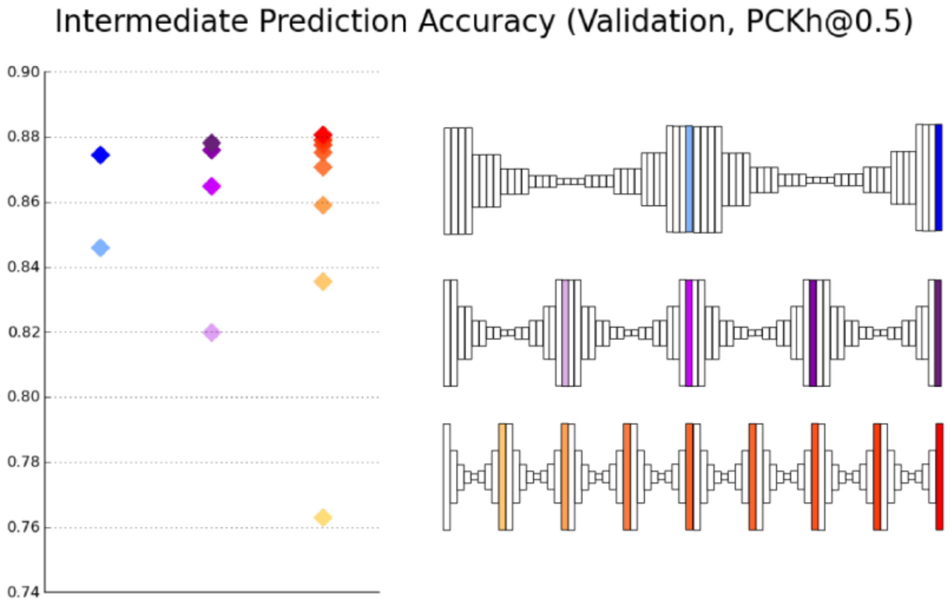


Image credit: https://arxiv.org/pdf/1603.06937.pdf

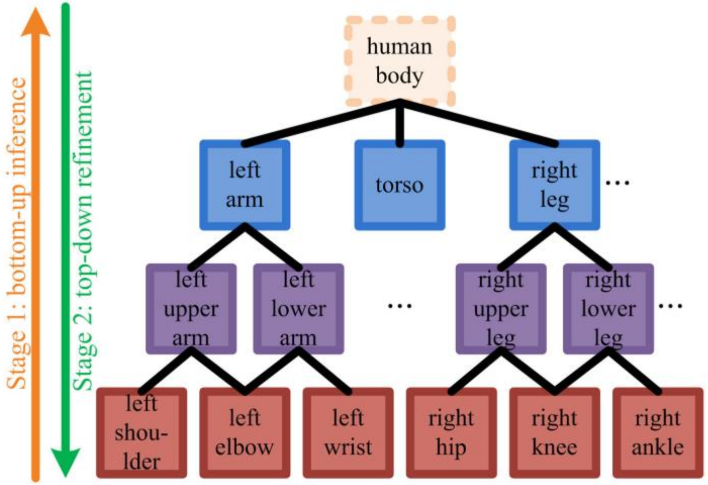
Some results on MPII dataset

Image credit: https://arxiv.org/pdf/1603.06937.pdf

Deeply learned compositional models

1. We have a compositional model of part-subpart relationships,
2. Traditional solutions by modelling the graph (or tree) with Gibbs formulation,
3. We want a bottom-up and top-down inference,

Question: How to model it with CNN models?



$$p(\Omega|\mathbf{I}) = \frac{1}{Z}\exp\{-E(\Omega,\mathbf{I})\}$$

$$S(\Omega) \equiv -E(\Omega,\mathbf{I}) = \sum_{u\in\mathcal{V}^{leaf}} \phi_u^{leaf}(w_u,\mathbf{I}) + \sum_{u\in\mathcal{V}^{and}} \sum_{v\in ch(u)} \phi_{u,v}^{and}(w_u,w_v)$$

Score to be minimized

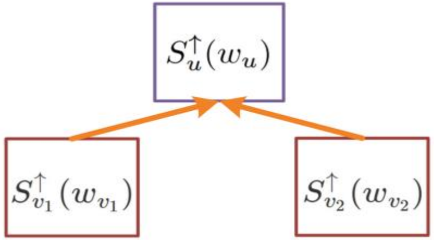
Unary term

Pairwise term

Image credit: Tang et al., "Deeply learned compositional models for human pose estimation." *ECCV*. 2018.

First define updating rule

Bottom-up inference:

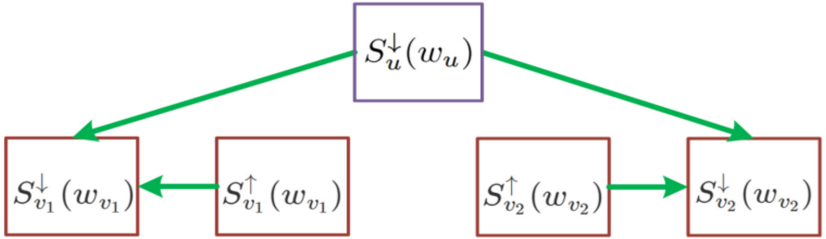


(a)

$$\text{(Leaf)}\ S_u^{\uparrow}(w_u) = \phi_u^{leaf}(w_u, \mathbf{I})$$

$$\text{(And)}\ S_u^{\uparrow}(w_u) = \sum_{v \in ch(u)} \max_{w_v}[\phi_{u,v}^{and}(w_u, w_v) + S_v^{\uparrow}(w_v)]$$

Top-down inference:



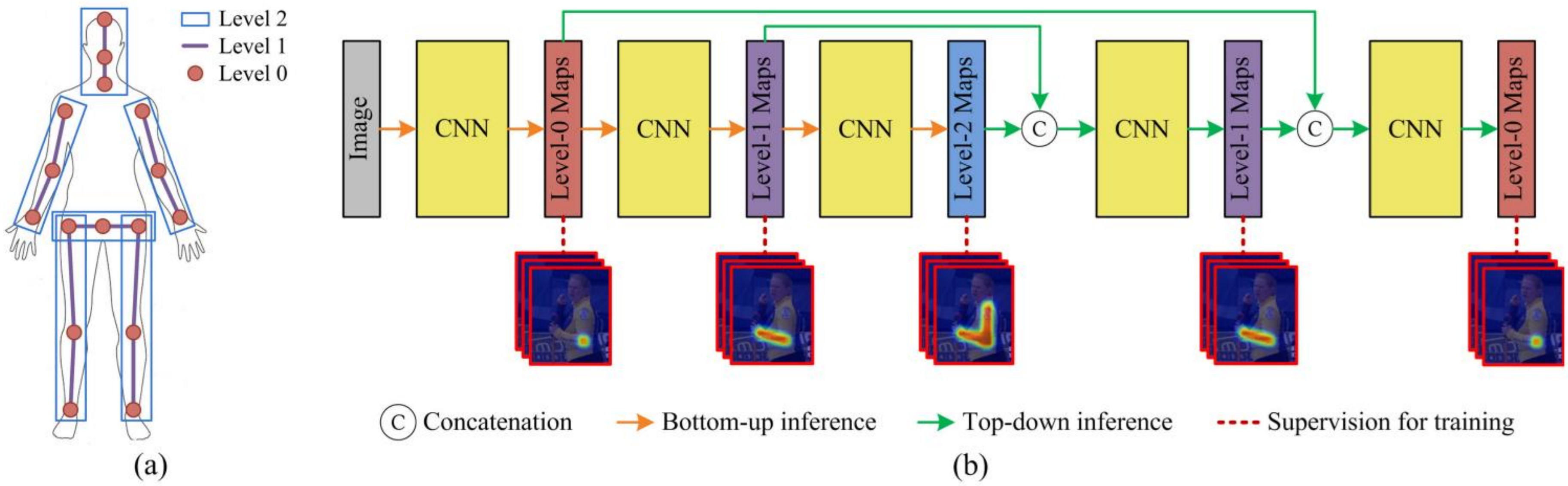
(b)

$$\text{(Root)}\qquad w_u^* = \text{argmax}_{w_u} S_u^{\downarrow}(w_u) \equiv \text{argmax}_{w_u} S_u^{\uparrow}(w_u)$$

$$\text{(Non-root)}\ w_v^* = \text{argmax}_{w_v} S_v^{\downarrow}(w_v) \equiv \text{argmax}_{w_v}[\phi_{u,v}^{and}(w_u^*, w_v) + S_v^{\uparrow}(w_v)]$$

Image credit: Tang et al., "Deeply learned compositional models for human pose estimation." *ECCV*. 2018.

Revisit stacked hourglass network



Heatmaps of level *i* (*i*>0) are composed of heatmaps of level 0 and level *i*.
CNN -> Hourglass Module

Image credit: Tang et al., "Deeply learned compositional models for human pose estimation." *ECCV*. 2018.

Some results comparing to 8-stack hourglass

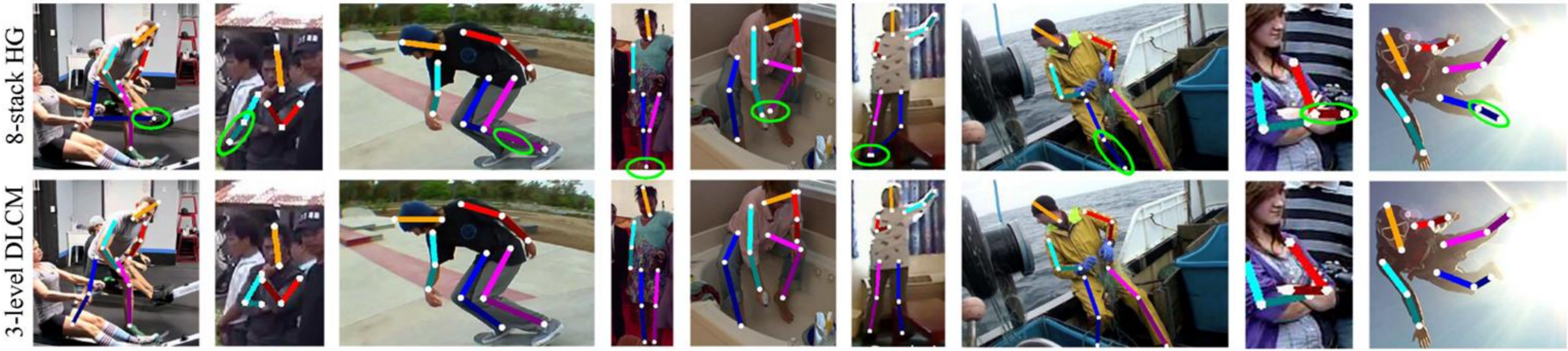


Image credit: Tang et al., "Deeply learned compositional models for human pose estimation." *ECCV*. 2018.

3D human body pose

- 3D pose is the vector of body joints in 3D space,
- 2D pose is the projection of 3D pose to image plane,
- 3D pose is defined with the same models as 2D pose,
- 3D pose estimation is a challenging task since depth dimension is lost during projection to RGB image.

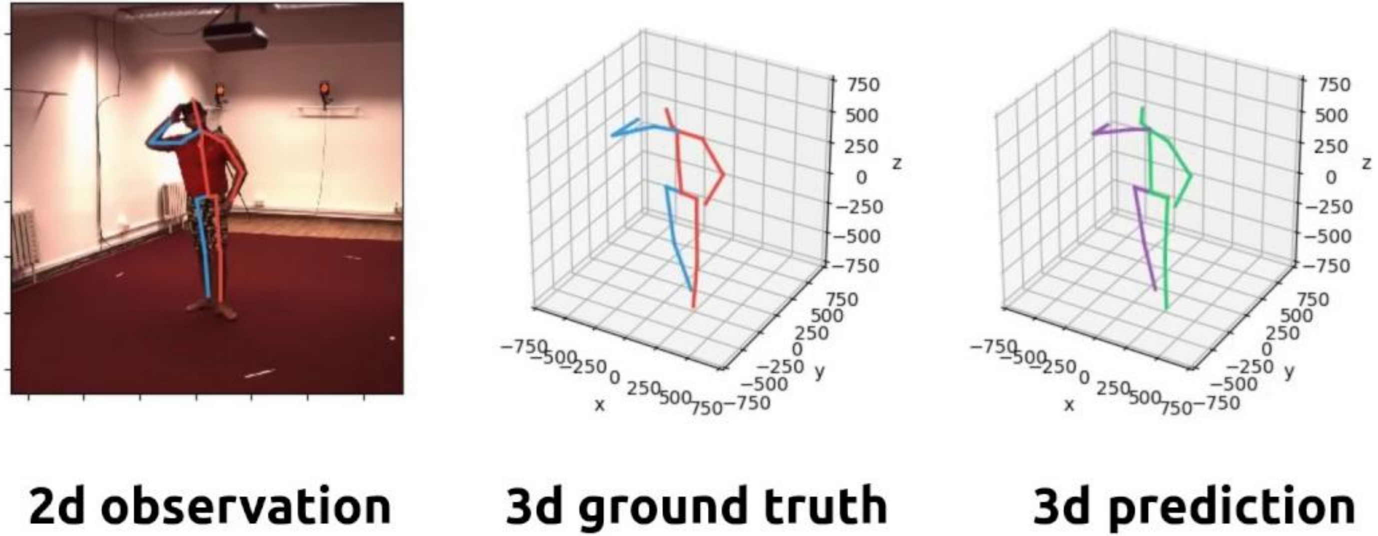


Image credit: https://www.youtube.com/watch?v=Hmi3Pd9x1BE

3D human body pose estimation

Solutions:

- Directly regress 3D pose from RGB image,
- Estimate 2D pose and then regress 3D pose from 2D,
- Apply volumetric heatmaps and estimate 3D pose,
- Apply multi-task learning by the use of different modalities.

3D human body pose estimation

Solutions:

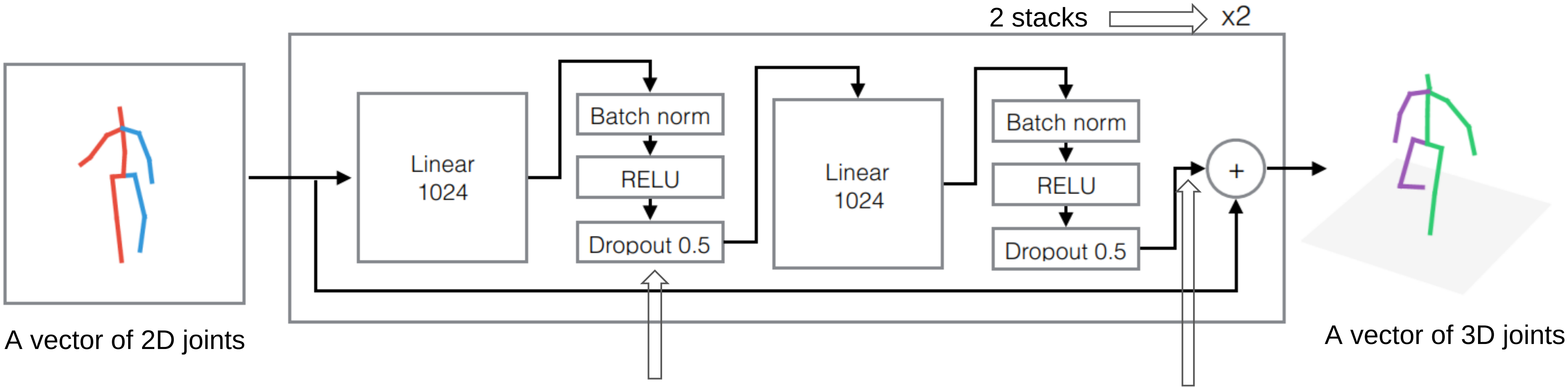
- Directly regress 3D pose from RGB image, ⇐ A difficult problem to the network similar to 2D pose.
- Estimate 2D pose and then regress 3D pose from 2D,
- Apply volumetric heatmaps and estimate 3D pose,
- Apply multi-task learning by the use of different modalities. ⇐ Annotating data for all modalities is not a trivial task for many datasets.

3D human body pose estimation

Solutions:

- Directly regress 3D pose from RGB image,
- Estimate 2D pose and then regress 3D pose from 2D,
- Apply volumetric heatmaps and estimate 3D pose,
- Apply multi-task learning by the use of different modalities.

Lifting 2D joints to 3D: A simple solution



Batch norm and dropout improve network robustness against input noise

Network predicts joint offsets

Image credit: https://arxiv.org/pdf/1705.03098.pdf

Volumetric heatmaps for 3D pose estimation

- A volumetric heatmap is defined as a tensor in the form of (*#Width x #Height x #Depth x #Joints),*
- A joint's depth is a continuous value which is discretized into several bins, i.e. *#Depth,*
- Ground truth heatmap can be defined by a Gaussian.

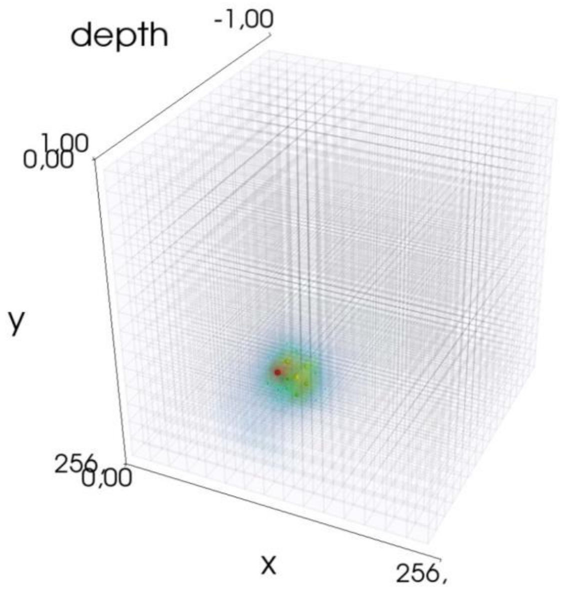


Image credit: https://www.vision.rwth-aachen.de/media/papers/synthetic_occl_augm/posetrack3d-sarandi-poster.pdf

A simple solution

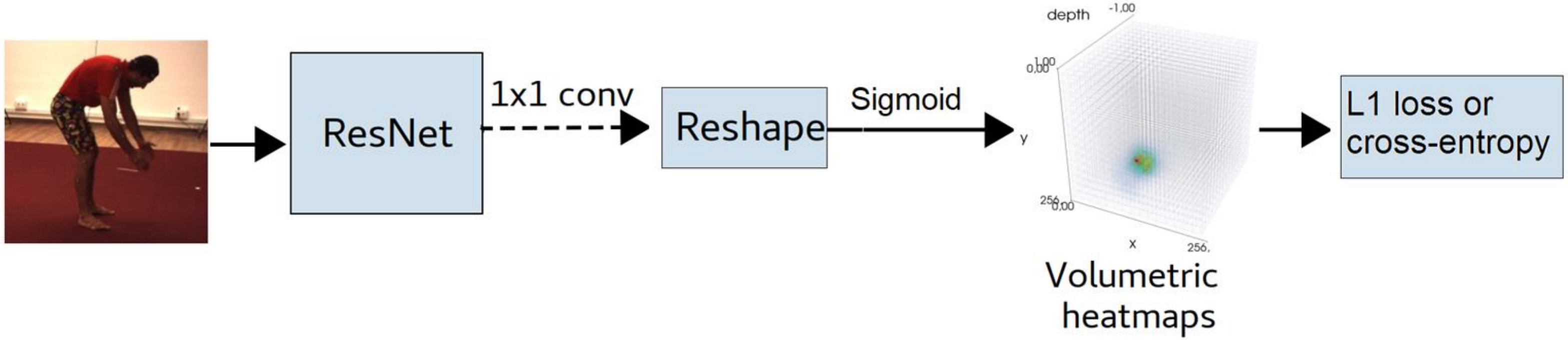


Image credit with modification: https://www.vision.rwth-aachen.de/media/papers/synthetic_occl_augm/posetrack3d-sarandi-poster.pdf

A simple solution - data augmentation

2638 occluder objects from Pascal VOC

Filter out ‘person’, ‘truncated’, ‘difficult’ and small object segments

Augmented inputs with pasted occluders

Applied with 50% probability, 1–8 objects, at random scale, at random position

Image credit: https://www.vision.rwth-aachen.de/media/papers/synthetic_occl_augm/posetrack3d-sarandi-poster.pdf

Applications

Image credit: https://www.sciencedirect.com/science/article/pii/S1047320315001121

3D hand pose estimation based on depth images
(similar strategies applied on RGB images)

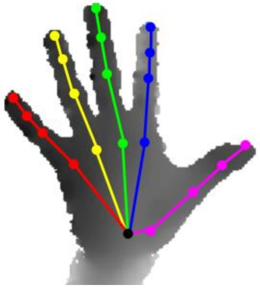


Image credit: http://daariga.github.io/

Applications

- Human-computer interaction,
- Virtual reality,
- Training robot hands,
- Sign language and gesture recognition.

Voxel to voxel predictions

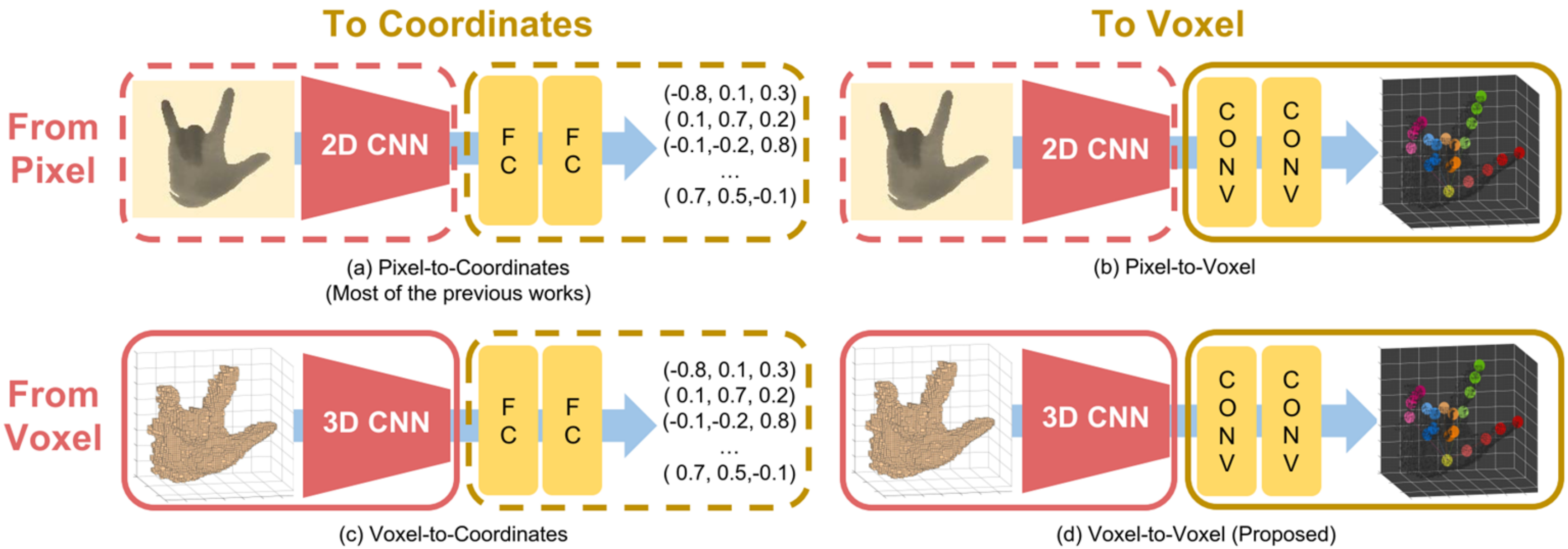


Image credit: https://arxiv.org/pdf/1711.07399

Voxel to voxel predictions

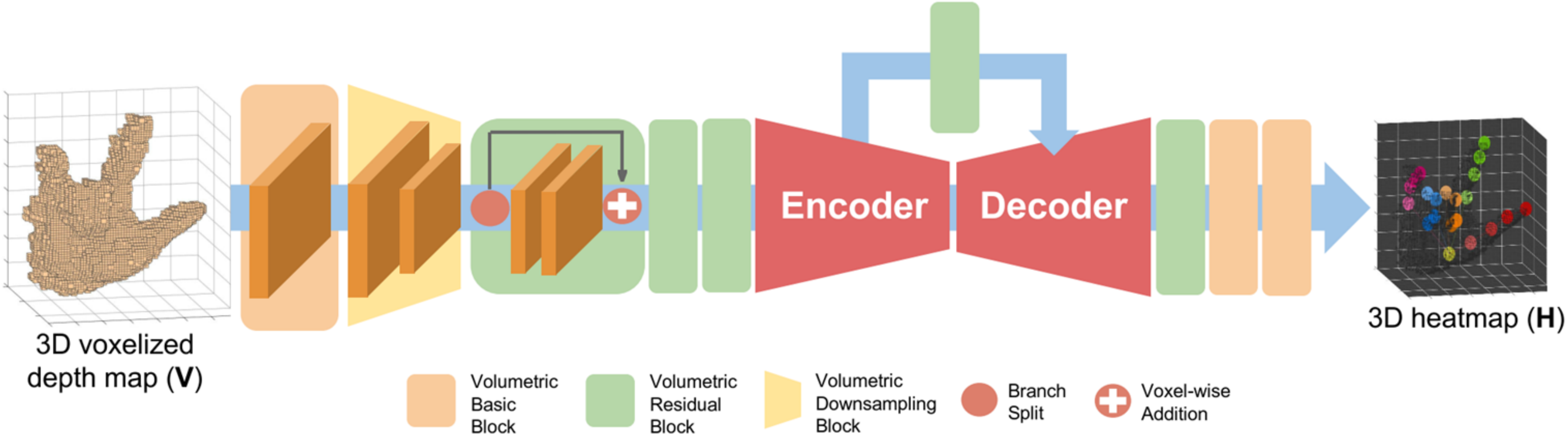


Image credit: https://arxiv.org/pdf/1711.07399

Voxel to voxel predictions - results on NYU dataset

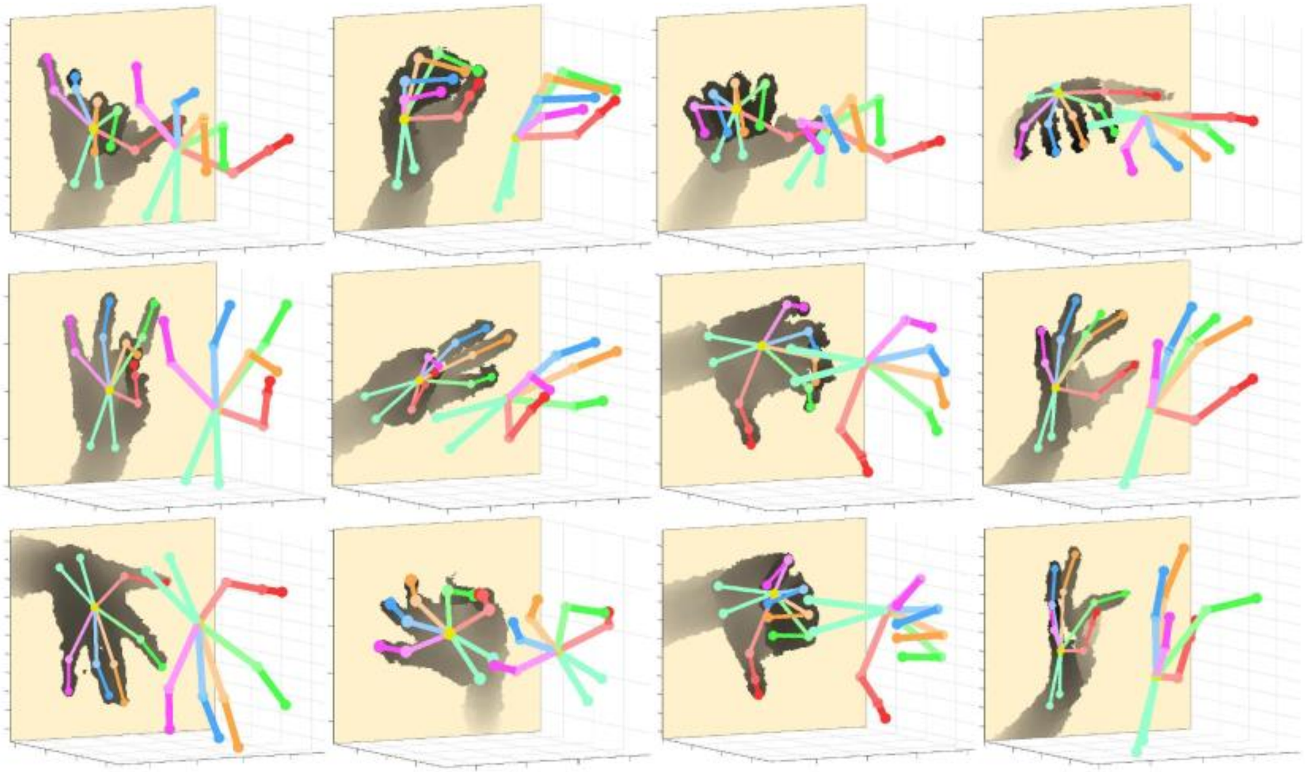
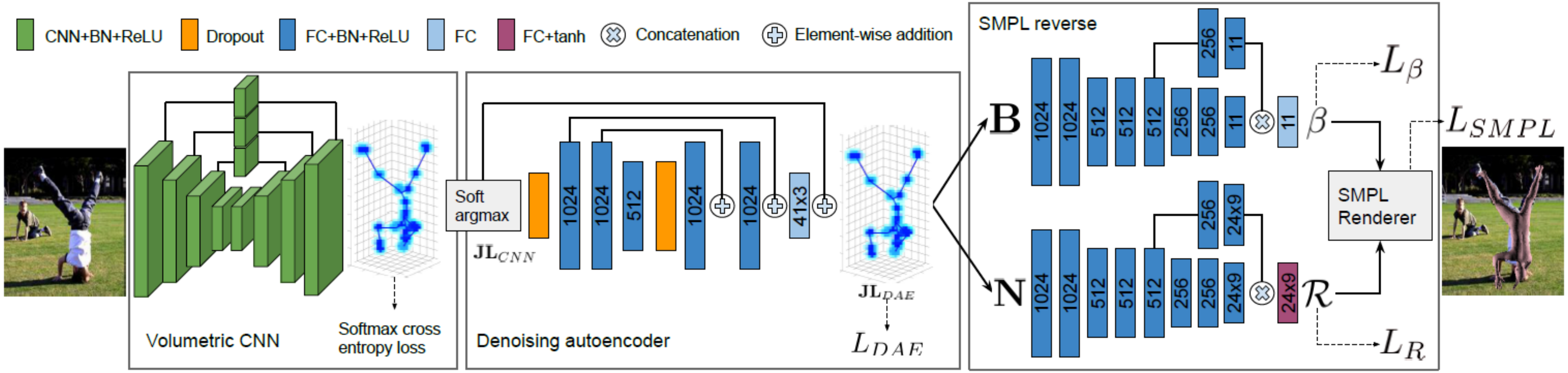


Image credit: https://arxiv.org/pdf/1711.07399

Bonus 1: Volumetric models from 2D (even 3D clothes)



SMPLR: Deep SMPL reverse for 3D human pose and shape recovery
Meysam Madadi, Hugo Bertiche, Sergio Escalera

Bonus 2: Faces

Automatic human face and body analysis

(from Martinez (2019). Context may reveal how you feel. PNAS)

Face
- Identity of the person
- *Perceived* age
- *Perceived* gender
- *Perceived* attractiveness
- Ethnicity/race/skin color
- Head pose
- Gaze direction
- Facial expression of *emotion:* crying (of sadness)
- Other facial features

+ Body
- Body pose
- Hand pose
- Bodily expression of *emotion*: crying (of dispair)
- Other body features

+ Context
- Activity/actions/intentions: concert
- Experienced emotion: crying of joy
- Interaction with other people/objects

Bonus 2: Faces

Automatic human face and body analysis

Applications for good

- Human-Computer/Robot interaction
- Early intervention and medical diagnosis (e.g., ASD, depression)
- Augmented reality
- Image synthesis (e.g., digital avatars)
- Education (measuring engagement, effectiveness)
- Gaming
- Driving assistance
- Market research
- Assistive living (e.g., image captioning, fall detection)

Personalized empathic virtual agents

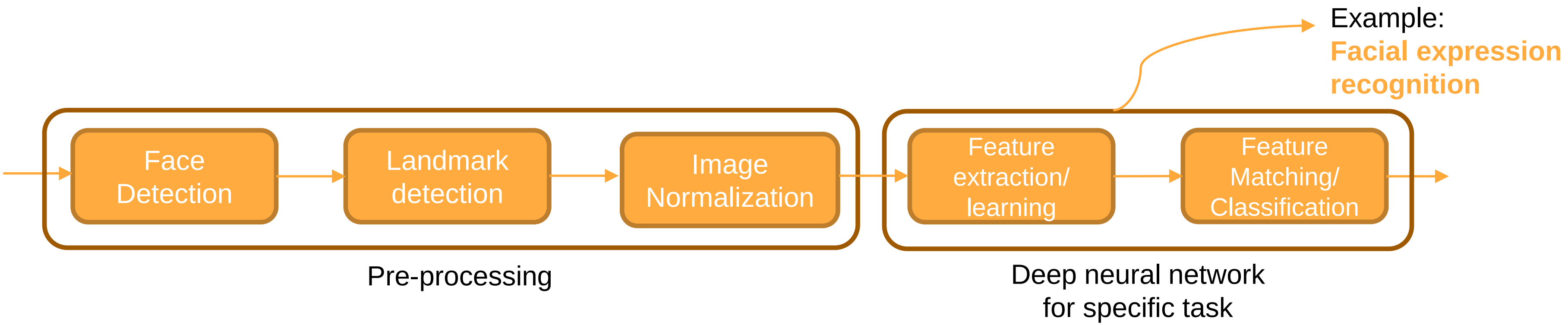
(from Justo et al (2020) Analysis of the interaction between elderly people and a simulated virtual coach. JAIHC)

Digital avatars

(from https://gigazine.net/gsc_news/en/20140813-real-time-facial-tracking/)

Bonus 2: Faces

Face analysis pipeline



<u>Pre-processing:</u>

1. **Face detection**: detect bounding box surrounding the face.
   → Bounding box coordinates regression (xmin, xmax, ymin, ymax).

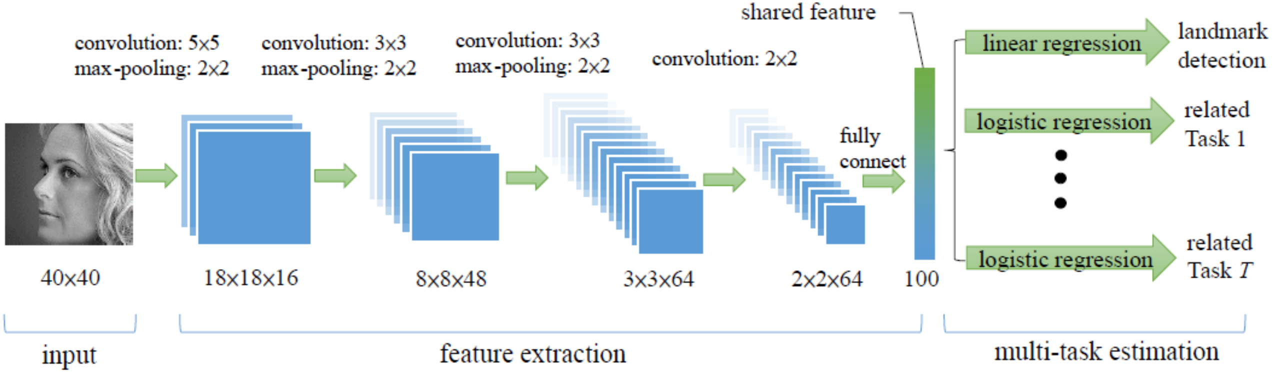
2. **Landmark detection**: detect facial landmarks (fiducial points or keypoints).
   → Keypoint coordinates regression (x,y)

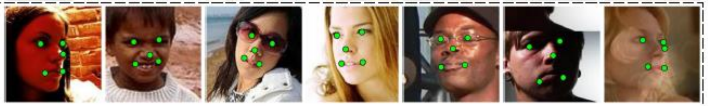
3. **Image normalization**: align and normalize the image to reduce variation in face scale and in-plane rotation.

Bonus 2: Faces

Multi-task learning: *optimizing facial landmark detection (the main task) with related/auxiliary tasks.*

2014 – TCDCN*

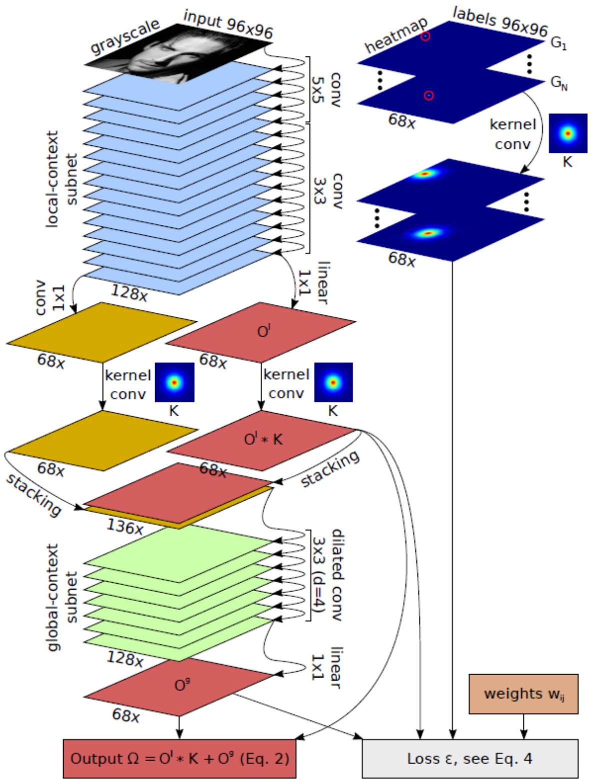


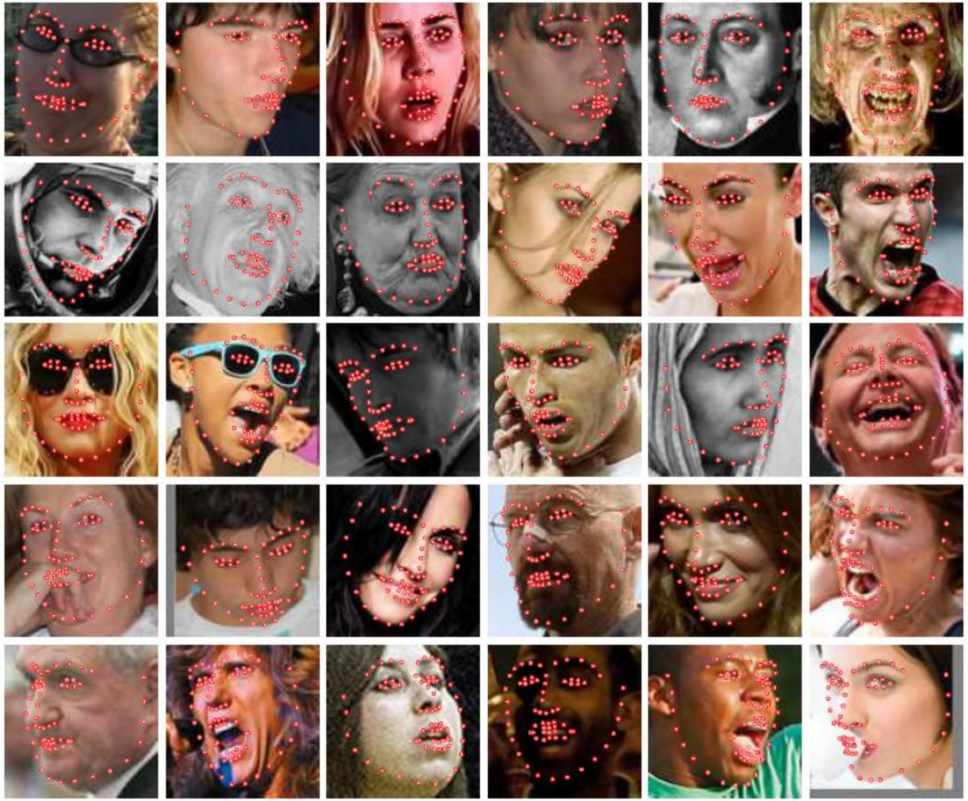


Auxiliary Tasks							
wearing glasses	×	×	√	×	√	×	×
smiling	×	√	×	×	×	×	×
gender	female	male	female	female	male	male	female
pose	right profile	frontal	frontal	left	frontal	frontal	right profile

*Zhang et al. (2014) Facial landmark detection by Deep Multi-task Learning. ECCV

Bonus 2: Faces





Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network, 2019

Human Pose Lecture

Dr. Sergio Escalera
University of Barcelona