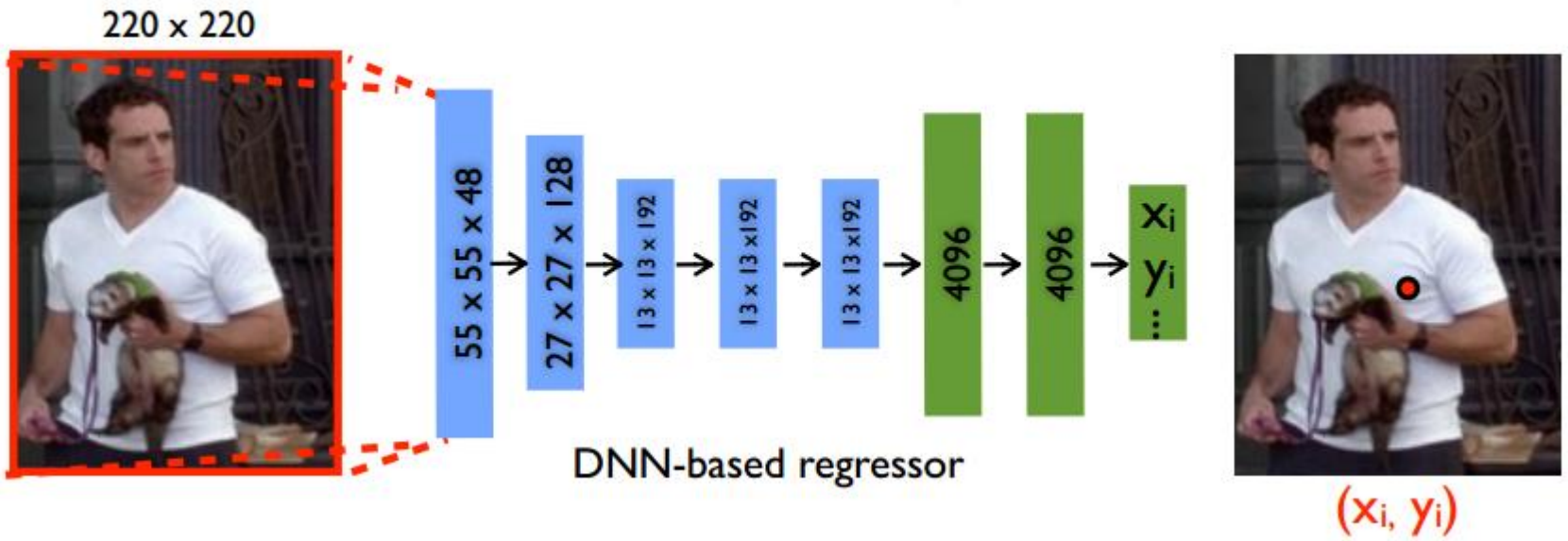
Goal: Localize all joints given input image and bounding box.

Network structure:

C(55 x 55 x 96) – LRN – P – C(27 x 27 x 256) – LRN – P –
C(13 x 13 x 384) – C(13 x 13 x 384) – C(13 x 13 x 256) – P –
F(4096,4096) – F(4096,4096)

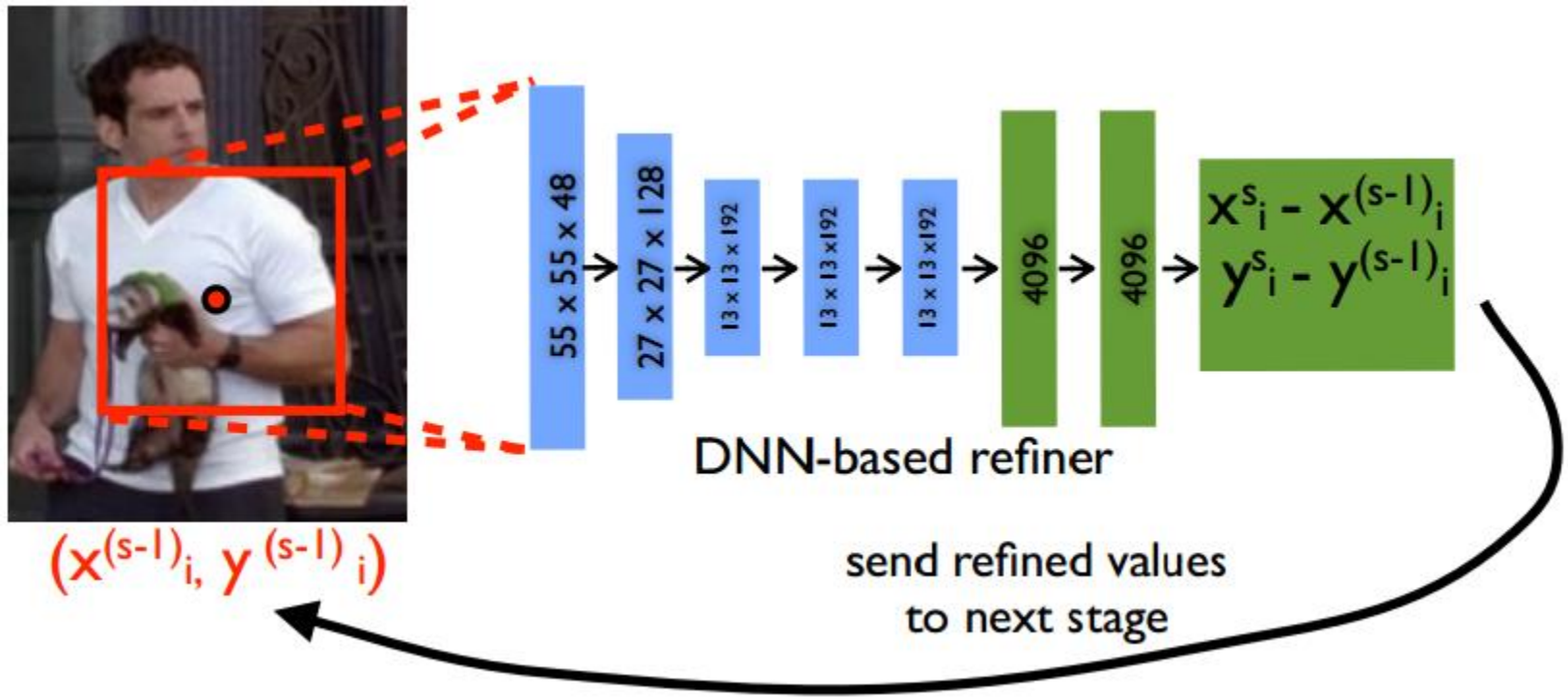


Input image is normalized wrt. bounding box width and height and joint coordinates are also normalized.

Image stride = 4

Network structure:

C(55 x 55 x 96) – LRN – P – C(27 x 27 x 256) – LRN – P –
C(13 x 13 x 384) – C(13 x 13 x 384) – C(13 x 13 x 256) – P –
F(4096,4096) – F(4096,4096)



Joint location is refined S times (S=3).
Same network structure as before.
Tries to estimate the residual from the previous step using a patch centered at the previous joint estimate.

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

i : joint index
yi : normalized coordinate of i:th joint
x : normalized image patch
theta : network parameters (approx. 40M)

In the first stage normalization is done wrt. the person bounding box.
In the refinement stages, normalization is wrt. a bounding box centered on the previous esimate.
In training the refinement stages, input data is augmented by random perturbations of the ground truth joint locations.

Method	Arm		Leg		Ave.
	Upper	Lower	Upper	Lower	
DeepPose-st1	0.5	0.27	0.74	0.65	0.54
DeepPose-st2	**0.56**	0.36	**0.78**	0.70	0.60
DeepPose-st3	**0.56**	**0.38**	0.77	**0.71**	**0.61**
Dantone et al. [2]	0.45	0.25	0.65	0.61	0.49
Tian et al. [21]	0.52	0.33	0.70	0.60	0.56
Johnson et al. [11]	0.54	**0.38**	0.75	0.66	0.58
Wang et al. [22]	**0.565**	0.37	0.76	0.68	0.59
Pishchulin [15]	0.49	0.32	0.74	0.70	0.56

Table 1. Percentage of Correct Parts (PCP) at 0.5 on LSP for DeepPose as well as five state-of-art approaches.

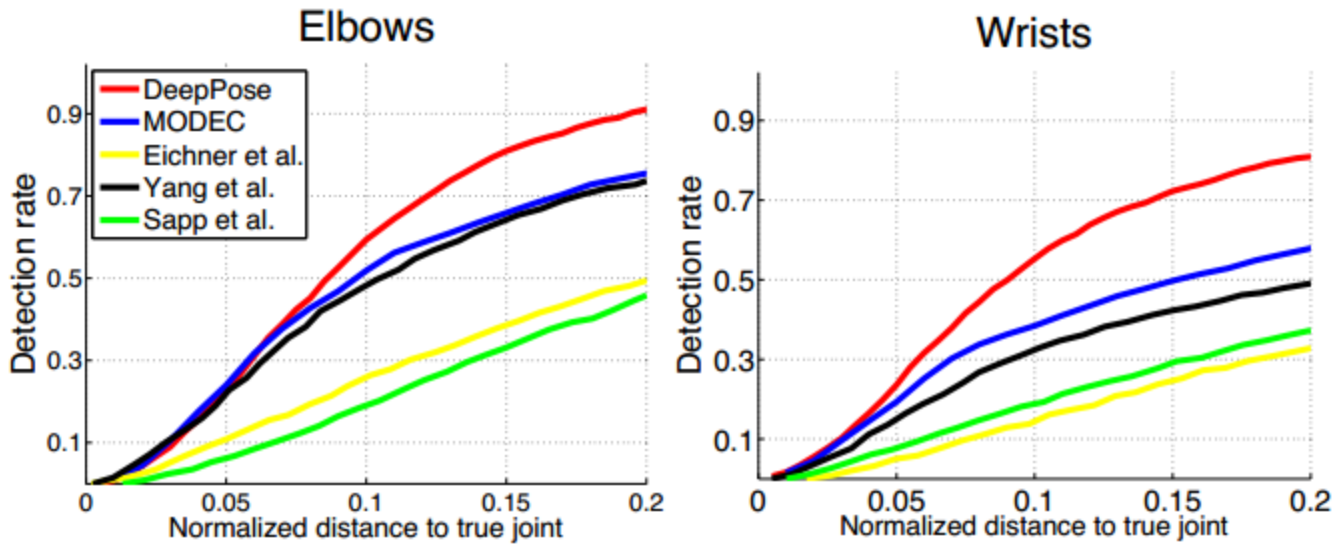


Figure 3. Percentage of detected joints (PDJ) on FLIC for two joints: elbow and wrist. We compare DeepPose, after two cascade stages, with four other approaches.

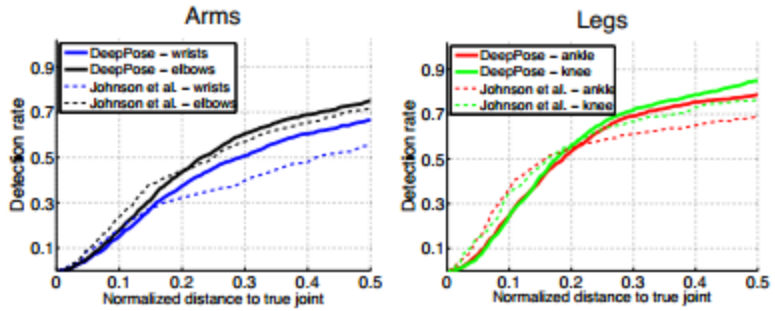


Figure 4. Percentage of detected joints (PDJ) on LSP for four limbs for DeepPose and Johnson et al. [11] over an extended range of distances to true joint: $[0, 0.5]$ of the torso diameter. Results of DeepPose are plotted with solid lines while all the results by [11] are plotted in dashed lines. Results for the same joint from both algorithms are colored with same color.

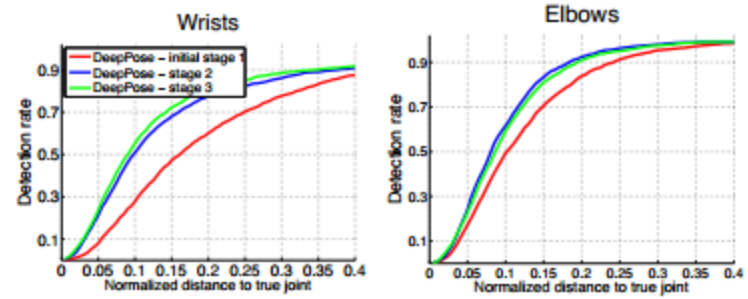


Figure 5. Percent of detected joints (PDJ) on FLIC or the first three stages of the DNN cascade. We present results over larger spectrum of normalized distances between prediction and ground truth.

Figure 6. Predicted poses in red and ground truth poses in green for the first three stages of a cascade for three examples.