Computer Vision; Pose Estimation

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

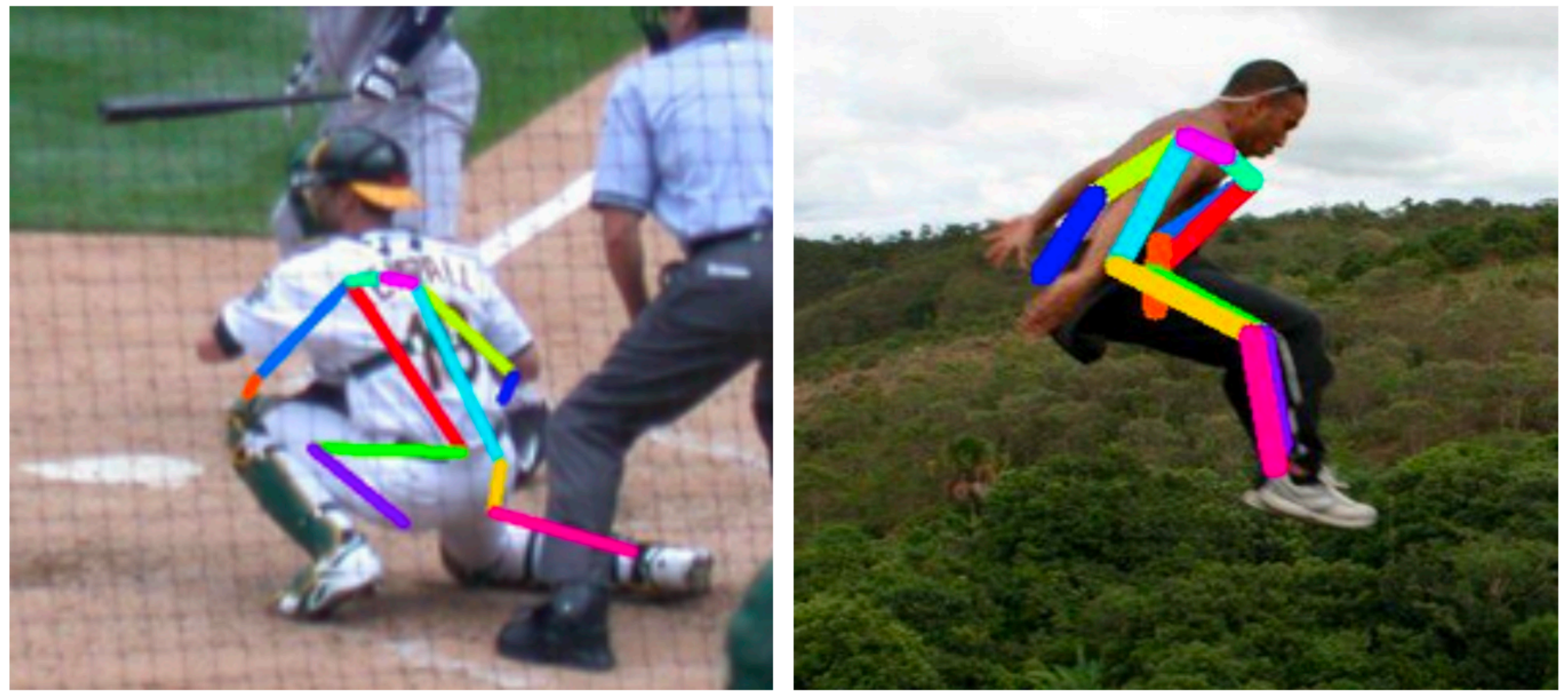
University of Colorado Boulder

maziar.raissi@colorado.edu

DeepPose: Human Pose Estimation via Deep Neural Networks

Boulder

YouTube Video



– small and barely visible joints

– occlusions and the need to capture the context

$k \rightarrow$ number of body joints

$y = (y_1, \ldots, y_k) \rightarrow$ pose vector

$y_i \in \mathbb{R}^2 \rightarrow$ coordinates of the i-th joint

$(x, y) \rightarrow$ image-label pair

$b = (b_c, b_w, b_h) \rightarrow$ box bounding the human body or parts of it (could be the full image)

$b_c \in \mathbb{R}^2 \rightarrow$ center of the box

$b_w, b_h \in \mathbb{R}^2 \rightarrow$ width & height of the box

$$N(y_i; b) = \frac{y_i - b_c}{(b_w, b_h)} \rightarrow \text{normalization by } b$$

$N(y; b) = (N(y_1; b), \ldots, N(y_k; b)) \rightarrow$ normalized pose vector

$N(x; b) \rightarrow$ crop of the image x by b

Pose Estimation as DNN-based Regression

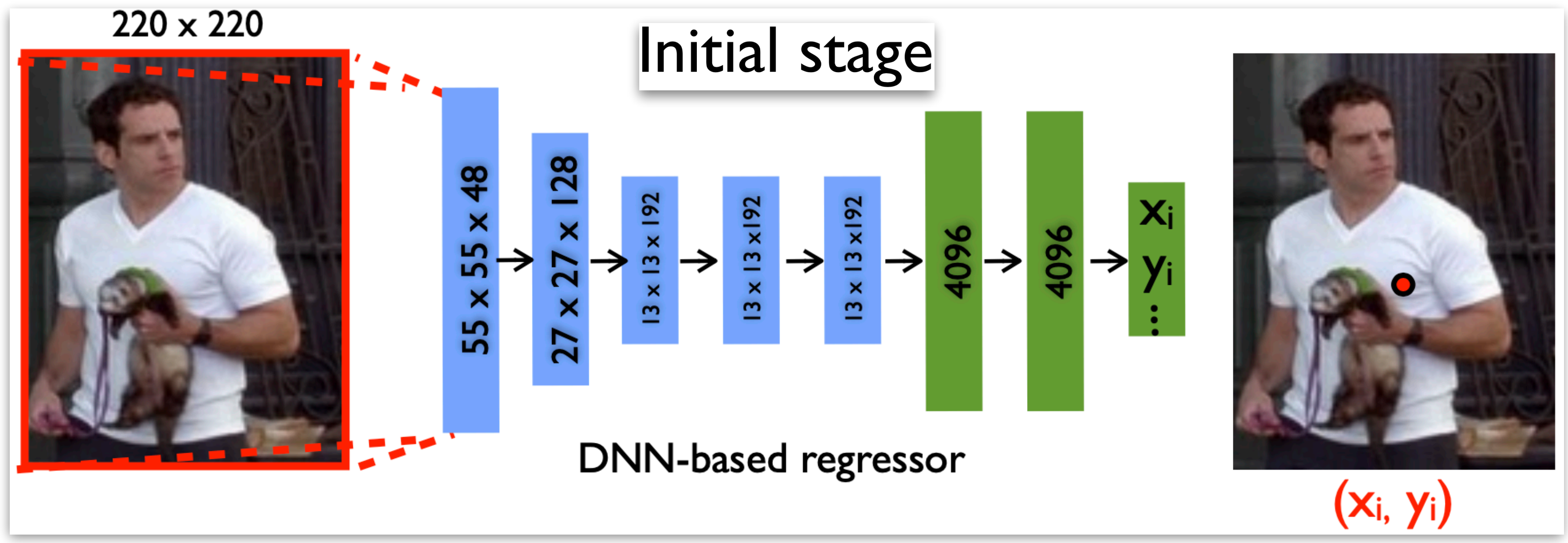
$\psi(x; \theta) \in \mathbb{R}^{2k} \rightarrow$ CNN (AlexNet)

$y^* = N^{-1}(\psi(N(x); \theta^*)) \rightarrow$ pose prediction in absolute image coordinates

$D_N = \{(N(x), N(y)) : (x, y) \in D\}$

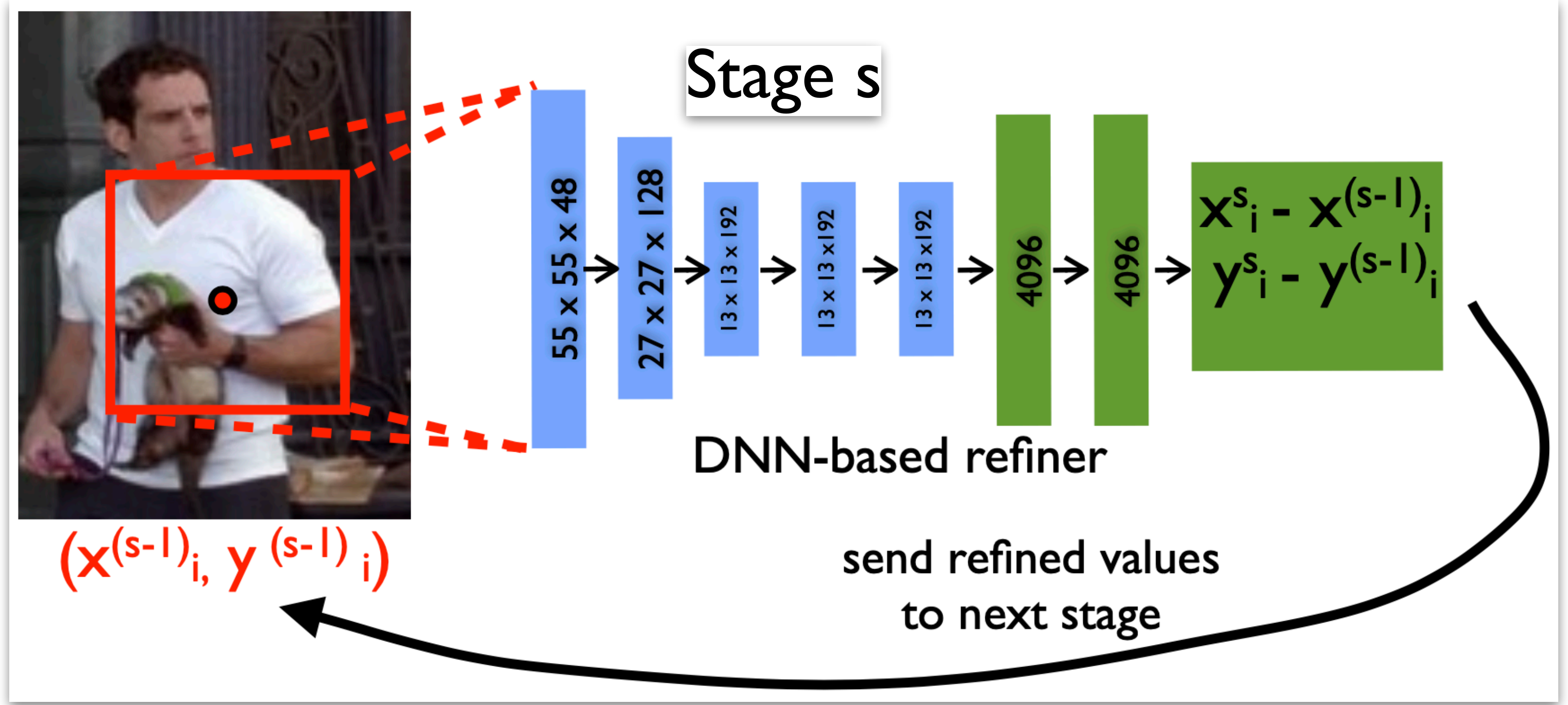
$\hookrightarrow$ normalized training set

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



Cascade of Pose Regressors

Each subsequent stage refines the currently predicted pose by regressing towards a refinement displacement $y_i^s - y_i^{(s-1)}$ on the sub image defined by $b_i^{(s-1)}$ from the previous stage.



Datasets:

– Frames Labeled In Cinema (FLIC)

– Leeds Sports Dataset (LSP)

Evaluation Metrics

<u>Percentage of Correct Parts (PCP)</u> measures detection rate of limbs, where a limb is considered detected if the distance between the two predicted joint locations and the true limb joint locations is at most half of the limb length.

<u>Percent of Detected Joints (PDJ)</u>: A joint is considered detected if the distance between the predicted and the true joint is within a certain fraction of the torso (e.g., left shoulder and right hip) diameter.

Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

Convolutional Pose Machines

YouTube Video

Pose Machines

$Y_p \in \mathcal{Z} \in \mathbb{R}^2 \rightarrow$ pixel location of the p-th $\overbrace{\text{anatomical landmark}}^{\text{part}}$

$\mathcal{Z} \rightarrow$ set of all locations $z = (u, v)$ in an image

$Y = (Y_1, Y_2, \ldots, Y_P) \rightarrow$ image locations for all P parts (to be predicted)

$g_t(\cdot), t = 1, \ldots, T \rightarrow$ sequence of multi-class predictors

$t \rightarrow$ stage

$g_t \rightarrow$ predicts beliefs for assigning a location to each part (i.e., $Y_p = z, \forall z$)

$x_z \in \mathbb{R}^d \rightarrow$ features extracted from the image at location z

$g_1 : x_z \mapsto \{\underbrace{b_1^p(Y_p = z)}_{\text{score for assigning the } p\text{-th part at image location } z \text{ in the first stage}}\}_{p=1,\ldots,P+1}$ background

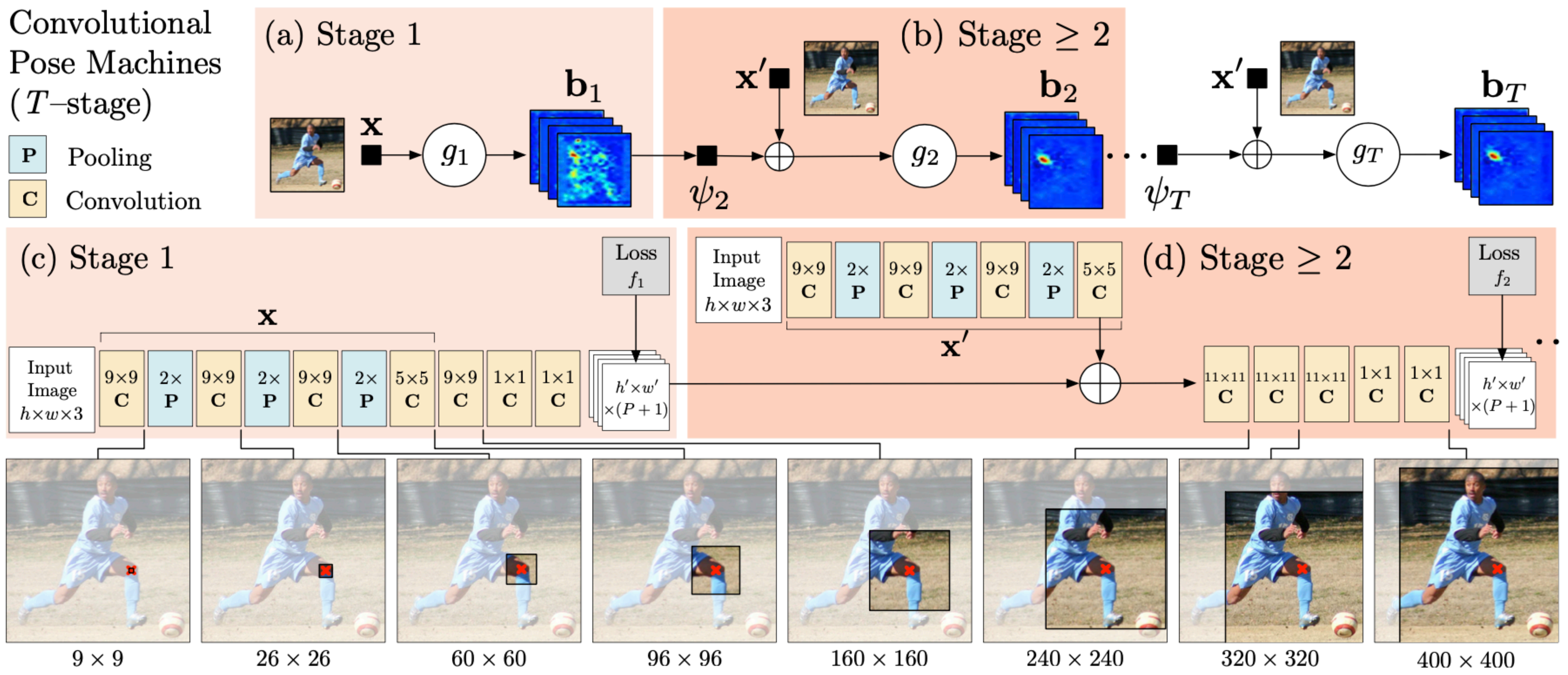
$b_t^p \in \mathbb{R}^{w \times h}, b_t^p \underbrace{(u, v)}_{z} = b_t^p(Y_p = z)$

$b_t \in \mathbb{R}^{w \times h \times (P+1)}$

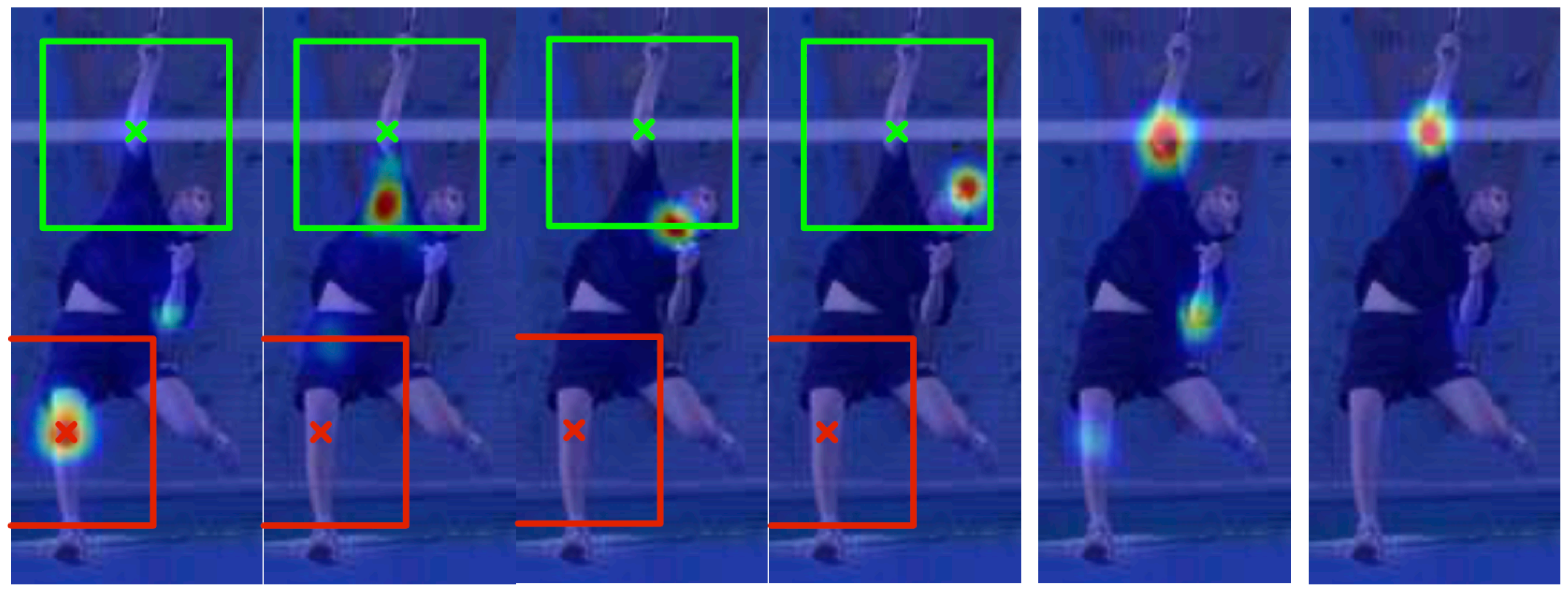
In subsequent stages $(t > 1)$:

$g_t : (x'_z, \underbrace{\psi_t(z, b_{t-1})}_{\text{context features}}) \mapsto \{b_t^p(Y_p = z)\}_{p=1,\ldots,P+1}$

boosted random forest and hand-crafted image & context features!



(e) Effective Receptive Field



Convolutional Pose Machines

$$\mathcal{F} = \sum_{t=1}^{T} f_t, \; f_t = \sum_{p=1}^{P+1} \sum_{z \in \mathcal{Z}} |b_t^p(z) - b_*^p(z)|^2$$

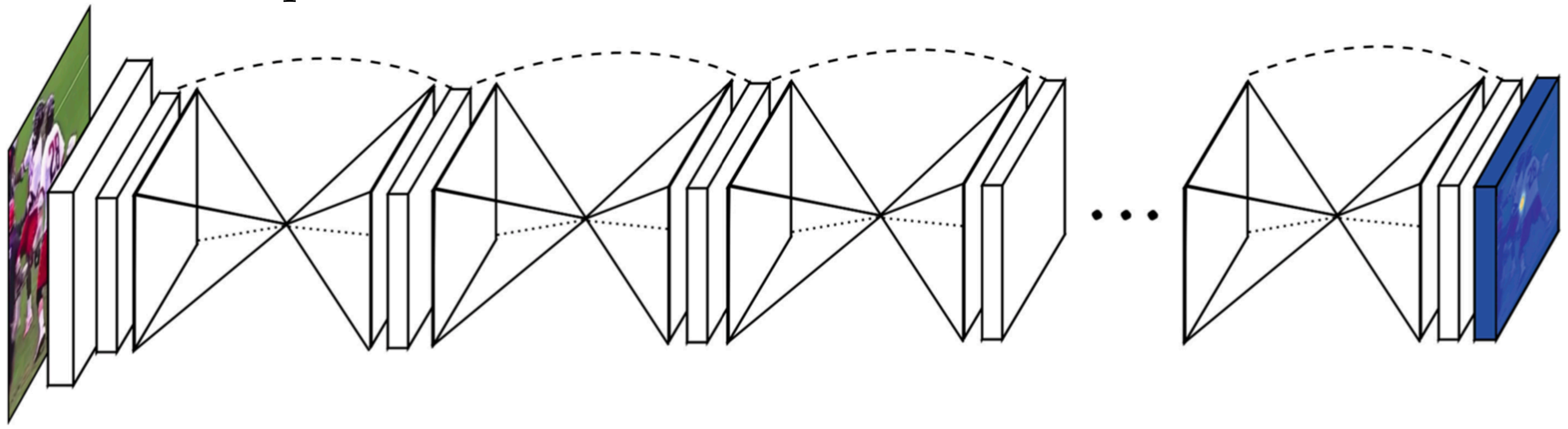
ideal belief map for a part p, $b_*^p(Y_p = z)$: putting Gaussian peaks at ground truth locations of each body part p

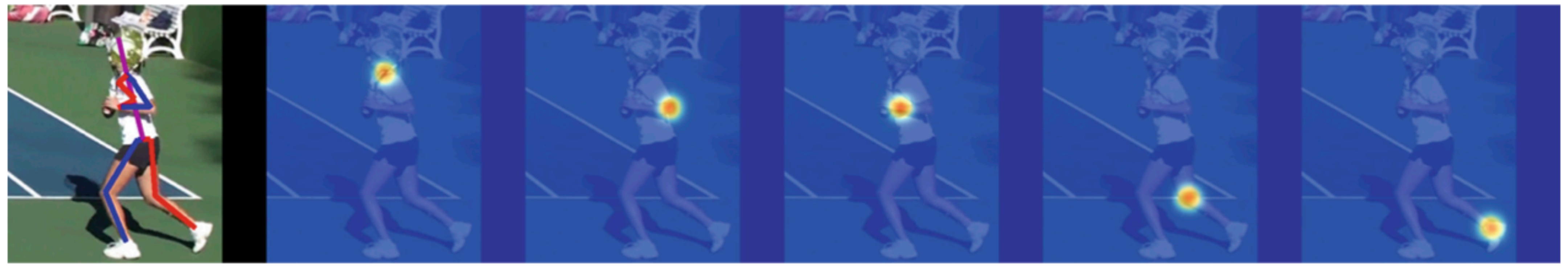
Wei, Shih-En, et al. "Convolutional pose machines." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016.

Stacked Hourglass Networks for Human Pose Estimation

[YouTube Video]

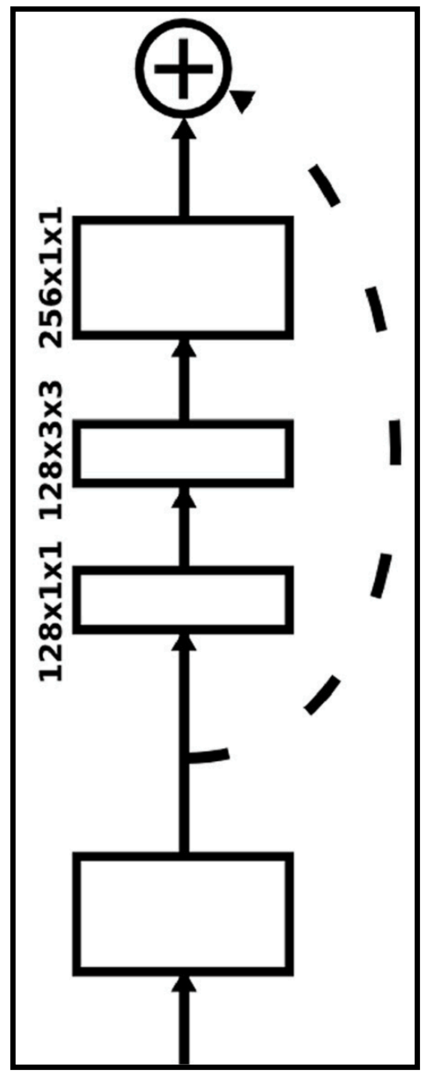
Human-computer interaction and animation



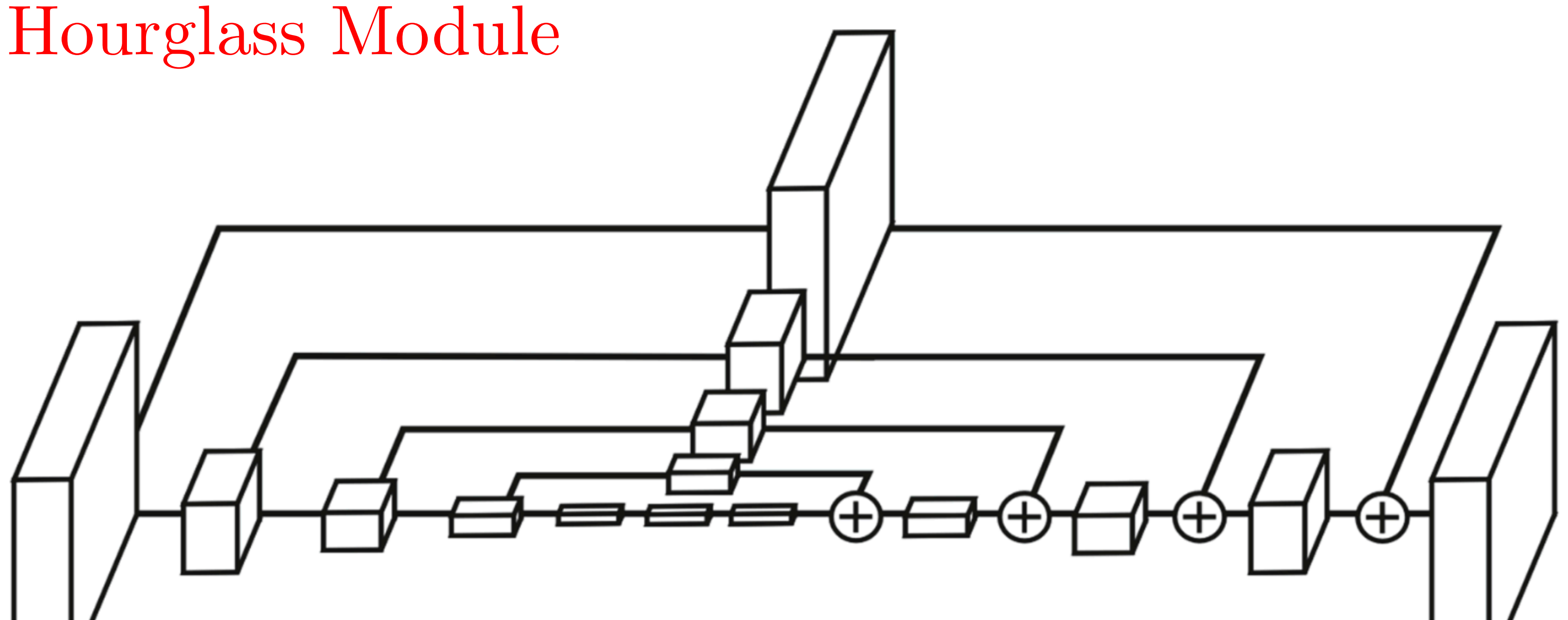


Motivation: Capturing information at every scale!

> “Local evidence is essential for identifying features like faces and hands, while a final pose estimate requires a coherent understanding of the full body.”

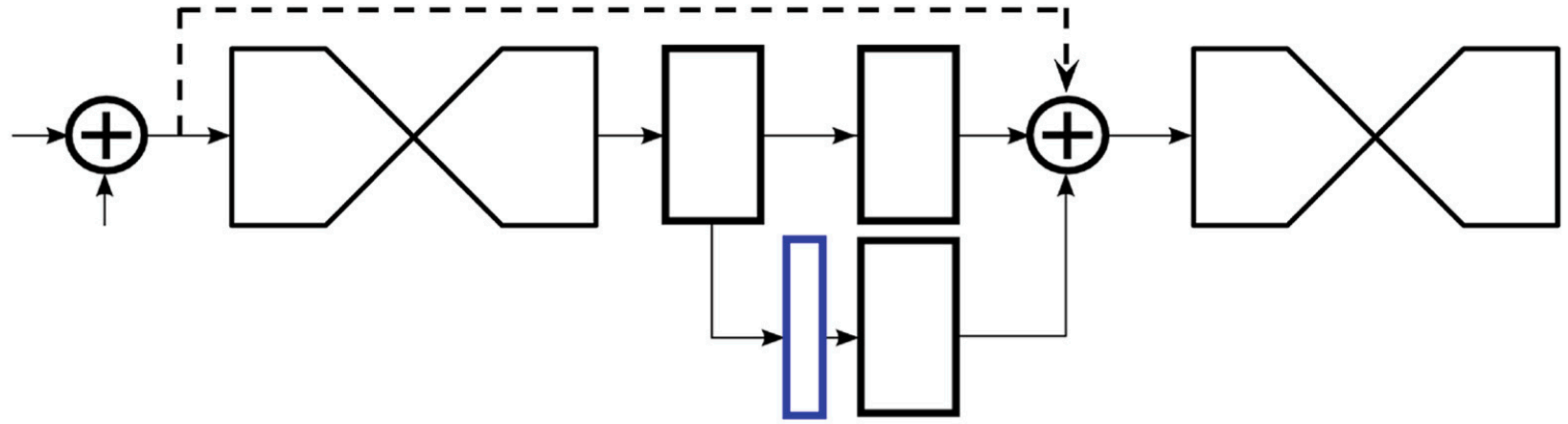


Hourglass Module



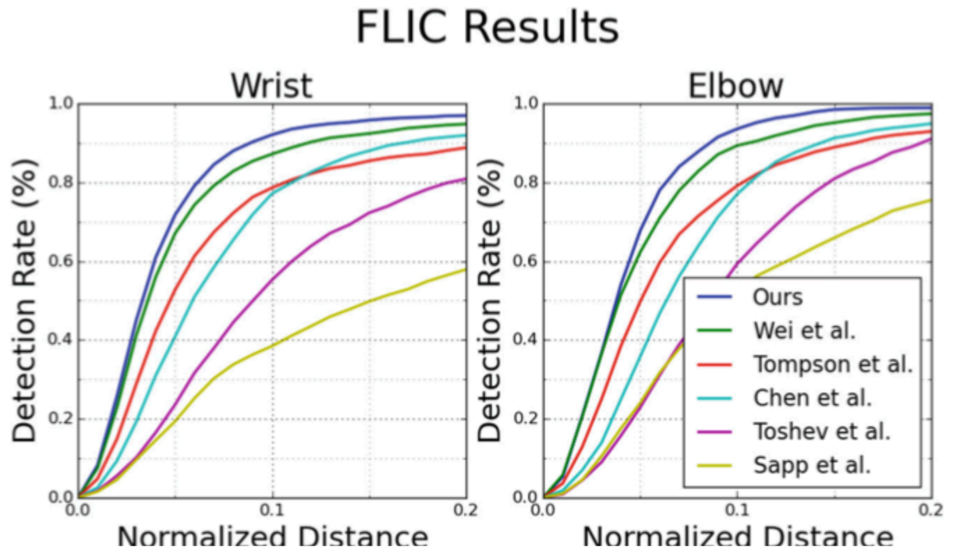
Each box corresponds to a residual module.

Intermediate Supervision

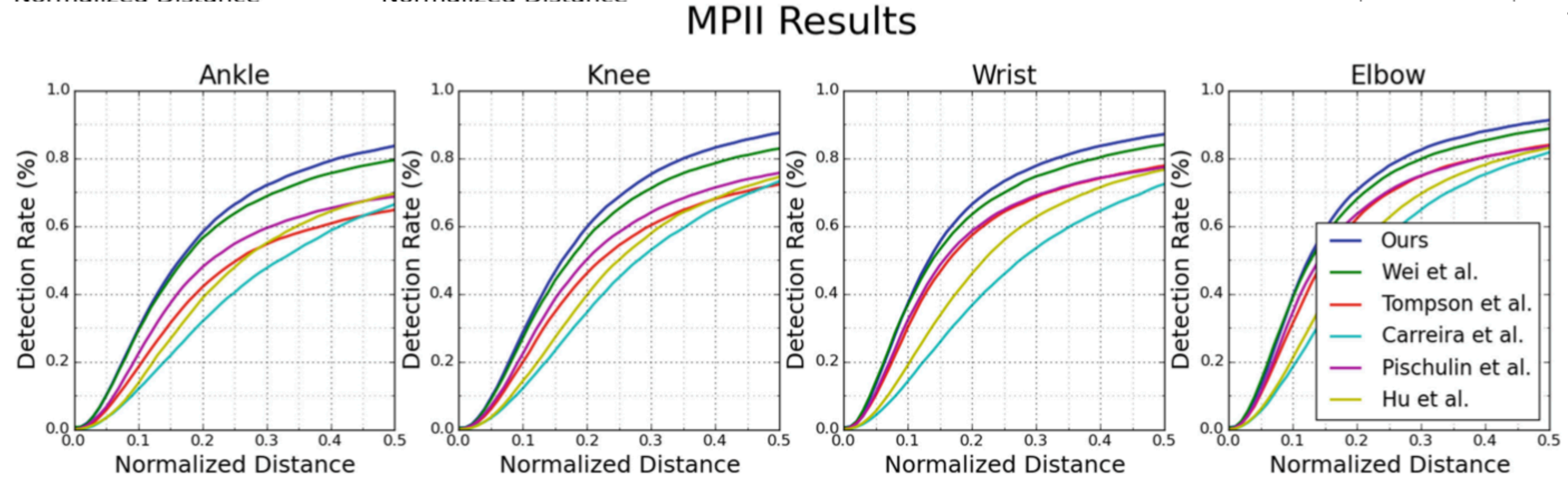


MSE Loss (predicted v.s. $\underbrace{\text{ground truth heatmaps}}_{\text{Gaussian}}$)

Percentage of Correct Keypoints (PCK): Percentage of detections that fall within a normalized distance of the ground truth.



	Elbow	Wrist
Sapp et al. [1]	76.5	59.1
Toshev et al. [24]	92.3	82.0
Tompson et al. [16]	93.1	89.0
Chen et al. [25]	95.3	92.4
Wei et al. [18]	97.6	95.0
Our model	**99.0**	**97.0**



Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *European conference on computer vision*. Springer, Cham, 2016.

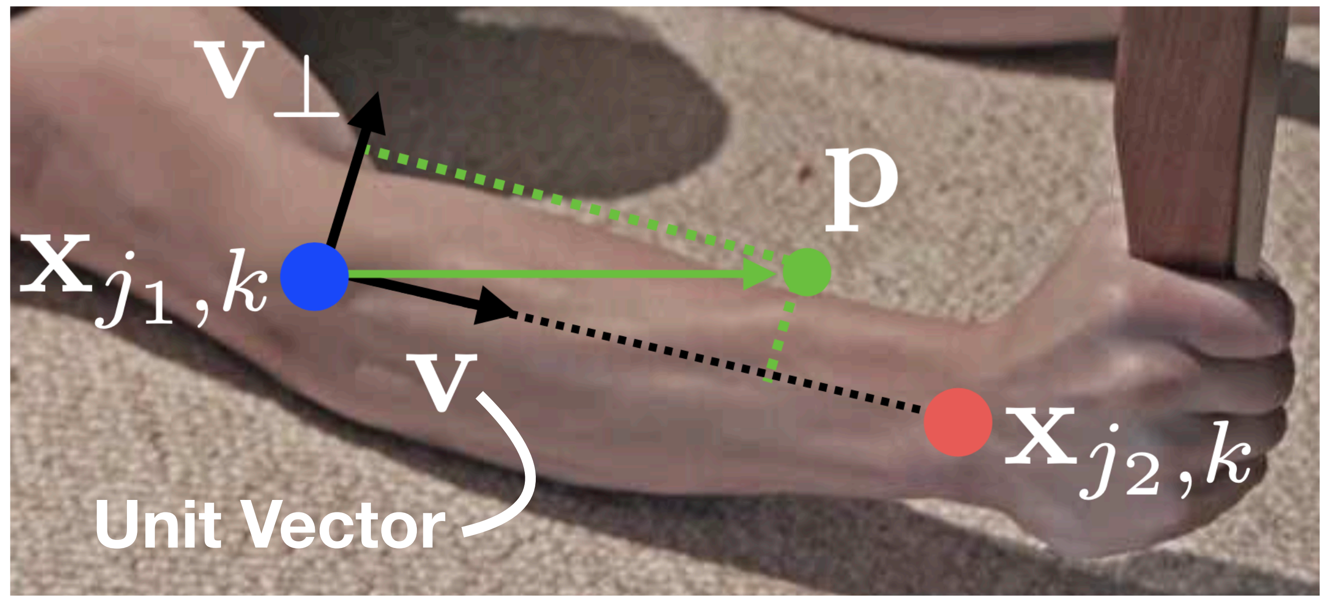
Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

Boulder

[YouTube Playlist]

Human 2D pose estimation $\rightarrow$ localizing anatomical keypoints or "parts"

(a) Input Image (b) Part Confidence Maps (c) Part Affinity Fields (d) Bipartite Matching (e) Parsing Results



$L^*_c(p) = \dfrac{1}{n_c(p)} \sum_k L^*_{c,k}(p)$

groundtruth part affinity field

$n_c(p) \rightarrow$ number of non-zero vectors at point p across all k people

$L^*_{c,k}(p) = v$ if p on limb c, k
$= 0$ otherwise

$\boxed{0 \leq \mathbf{v}\cdot(\mathbf{p}-\mathbf{x}_{j_1,k}) \leq l_{c,k} \text{ and } |\mathbf{v}_\perp \cdot (\mathbf{p}-\mathbf{x}_{j_1,k})| \leq \sigma_l}$

$l_{c,k} = \|\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}\|_2$

$S \rightarrow$ set of 2D confidence maps of body part locations

$L \rightarrow$ set of 2D vector fields of part affinities (encode the degree of association between parts)

$S = (S_1, \ldots, S_J)$ has J confidence maps (one per part)

$S_j \in \mathbb{R}^{w\times h}$

$L = (L_1, \ldots, L_C)$ has C vector fields (one per limb)

$L_c \in \mathbb{R}^{w\times h\times 2} \rightarrow$ each image location encodes a 2D vector

$F \leftarrow$ image $\triangleright$ VGG-19 (first 10 layers)

$S^1 = \rho^1(F) \qquad S^t = \rho^t(F, S^{t-1}, L^{t-1})$ for $t \geq 2$

$L^1 = \phi^1(F) \qquad L^t = \phi^t(F, S^{t-1}, L^{t-1})$ for $t \geq 2$

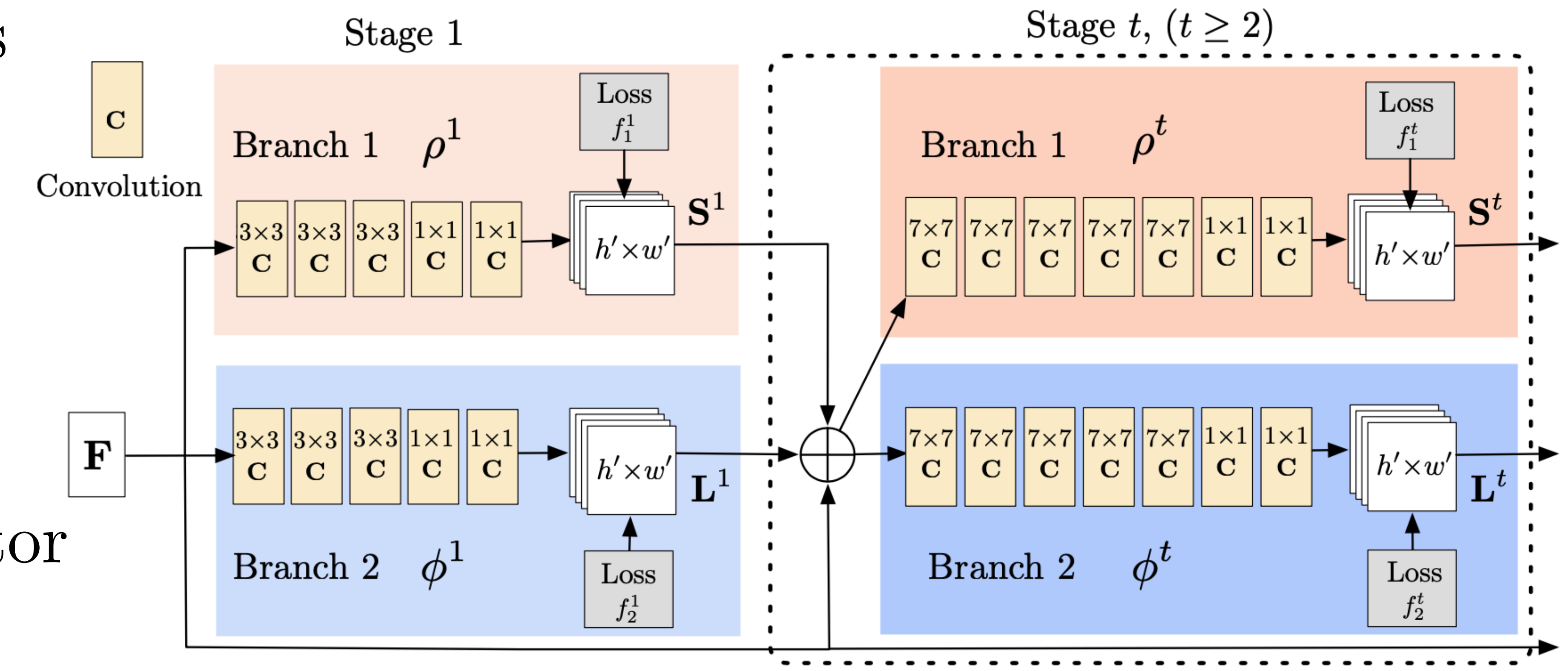
$W(p) = 0$ when the annotation is missing at an image location p

$$f^t_S = \sum_{j=1}^{J} \sum_p W(p) \|S^t_j(p) - S^*_j(p)\|^2_2$$

loss

$$f^t_L = \sum_{c=1}^{C} \sum_p W(p) \|L^t_c(p) - L^*_c(p)\|^2_2$$

$$f = \sum_{t=1}^{T} (f^t_S + f^t_L)$$



$S^*_j(p) = \max_k S^*_{j,k}(p) \rightarrow$ groundtruth confidence map

$S^*_{j,k}(p) \rightarrow$ confidence map for person k

$$S^*_{j,k}(p) = \exp\left(-\frac{\|p - x_{j,k}\|^2_2}{\sigma^2}\right)$$

$x_{j,k} \in \mathbb{R}^2 \rightarrow$ groundtruth position of body part j for person k in the image

Testing

$d_{j_1}, d_{j_2} \rightarrow$ two candidate part locations

$p(u) = (1-u)d_{j_1} + u d_{j_2}$

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot \frac{d_{j_2} - d_{j_1}}{\|d_{j_2} - d_{j_1}\|_2} du$$

association confidence

Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Deep High-Resolution Representation Learning for Human Pose Estimation

Boulder

YouTube Video

Human pose estimation (a.k.a keypoint estimation)

$K \rightarrow$ number of keypoints or parts

$I \in \mathbb{R}^{W \times H \times 3} \rightarrow$ image

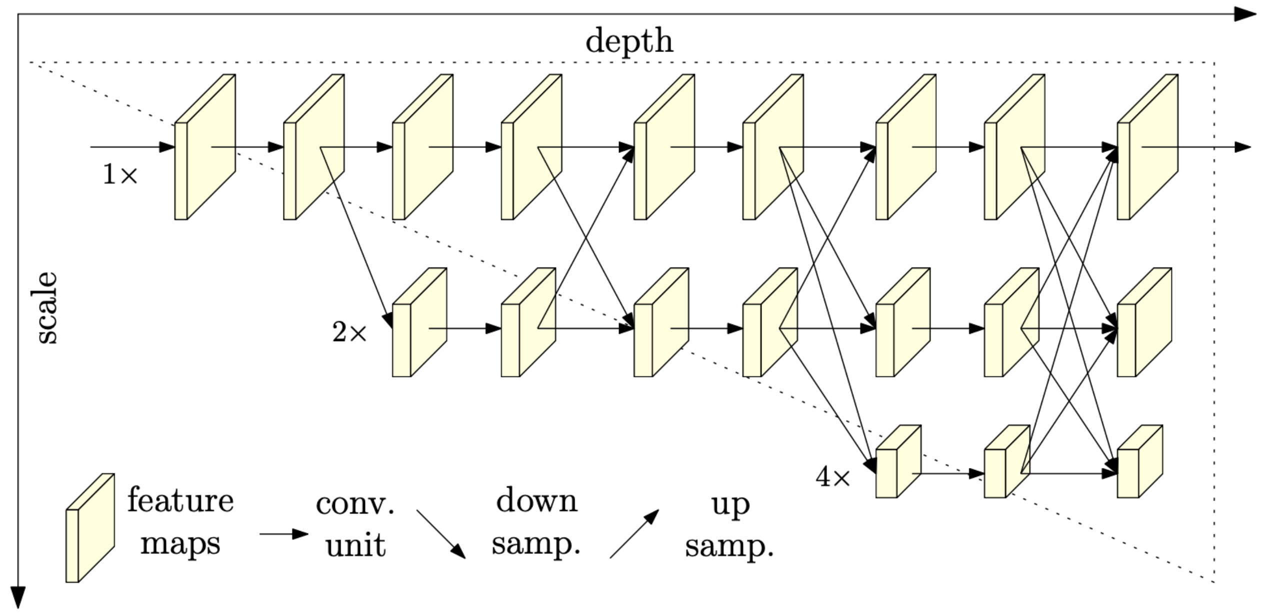
$\{H_1, H_2, \ldots, H_K\} \rightarrow$ K heatmaps

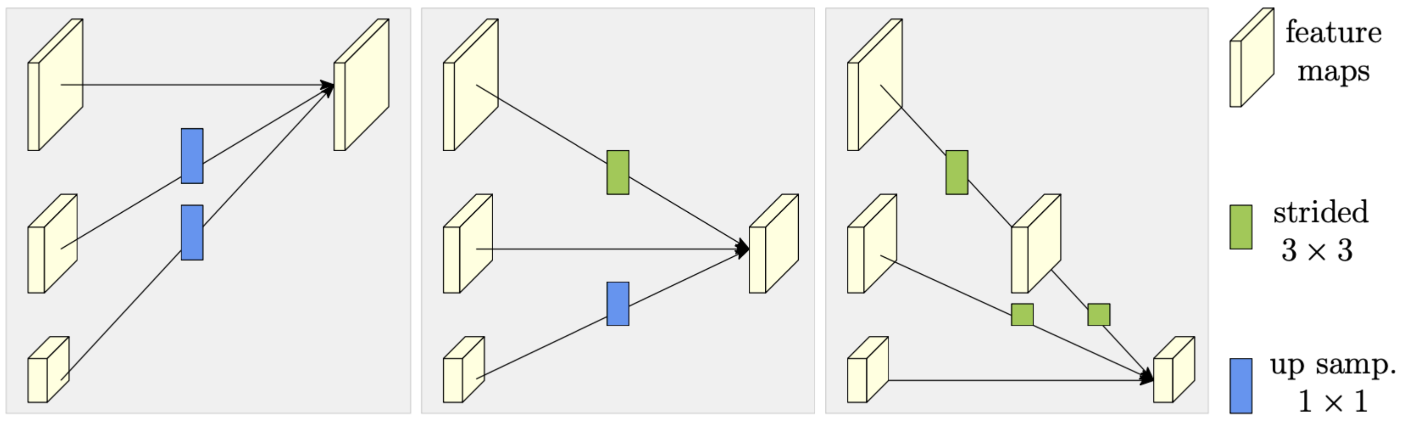
$H_k \in \mathbb{R}^{W' \times H'} \rightarrow$ represents location confidence of the k-th keypoint

– stem: two strided convolutions decreasing the resolution

– body: outputting feature maps with the same resolution as its input feature maps

– regressor: estimating the heatmaps where the keypoint positions are estimated and then transformed to the full resolution



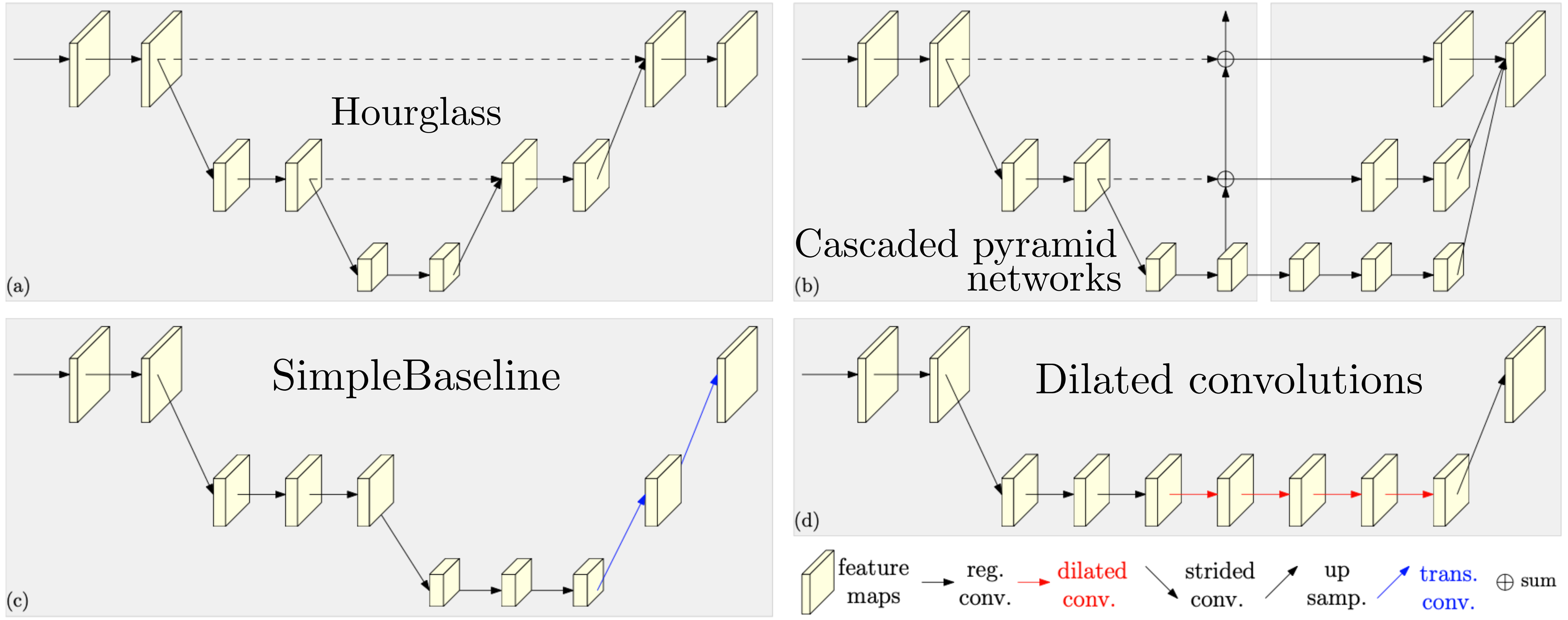


$AP^{50} \rightarrow$ average precision at OKS $= 50$

OKS: Object Keypoint Similarity (see the paper!)

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up: keypoint detection and grouping										
OpenPose [6]	–	–	–	–	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [38]	–	–	–	–	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [45]	–	–	–	–	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [32]	–	–	–	–	69.6	86.3	76.6	65.0	76.3	73.5
Top-down: human detection and single-person keypoint detection										
Mask-RCNN [21]	ResNet-50-FPN	–	–	–	63.1	87.3	68.7	57.8	71.4	–
G-RMI [46]	ResNet-101	353 × 257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [58]	ResNet-101	256 × 256	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	–
G-RMI + extra data [46]	ResNet-101	353 × 257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3
CPN [11]	ResNet-Inception	384 × 288	–	–	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [17]	PyraNet [74]	320 × 256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	–
CFN [24]	–	–	–	–	72.6	86.1	69.7	78.3	64.1	–
CPN (ensemble) [11]	ResNet-Inception	384 × 288	–	–	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [70]	ResNet-152	384 × 288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W32	HRNet-W32	384 × 288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet-W48	HRNet-W48	384 × 288	63.6M	32.9	**75.5**	**92.5**	**83.3**	**71.9**	**81.5**	**80.5**
HRNet-W48 + extra data	HRNet-W48	384 × 288	63.6M	32.9	**77.0**	**92.7**	**84.5**	**73.4**	**83.1**	**82.0**

Sequential multi-resolution subnetworks



Parallel multi-resolution subnetworks & Repeated multi-scale fusion

$\mathbf{Y}_k = \sum_{i=1}^{s} a(\mathbf{X}_i, k) \rightarrow$ exchange unit

$\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_s\} \rightarrow$ input response maps

$\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_s\} \rightarrow$ output response maps

$a(\mathbf{X}_i, k) \rightarrow$ upsampling or downsampling from resolution i to resolution k

$a(\mathbf{X}_i, k) = \mathbf{X}_i \rightarrow$ if $i = k$

$\mathbf{Y}_{s+1} = a(\mathbf{Y}_s, s+1) \rightarrow$ across stages

upsampling: nearest-neighbor followed by 1×1 conv & downsampling: strided 3×3 conv (stride 2)

Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Boulder

Questions?

YouTube Playlist