Fields of Parts & Friends

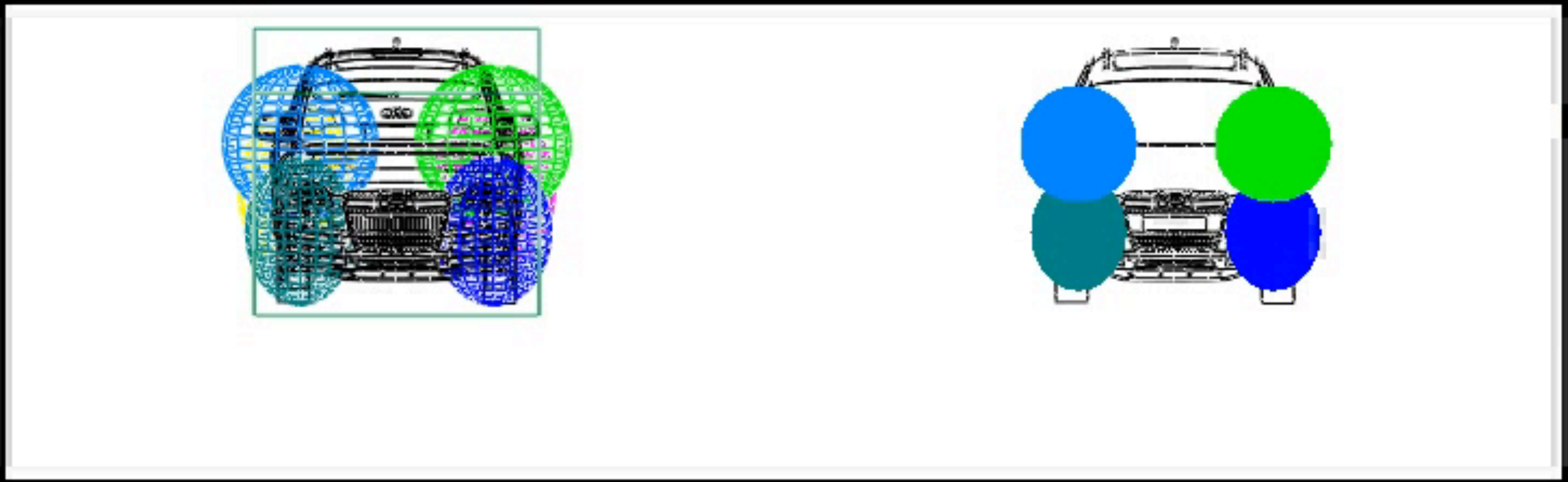
peter.gehler.net

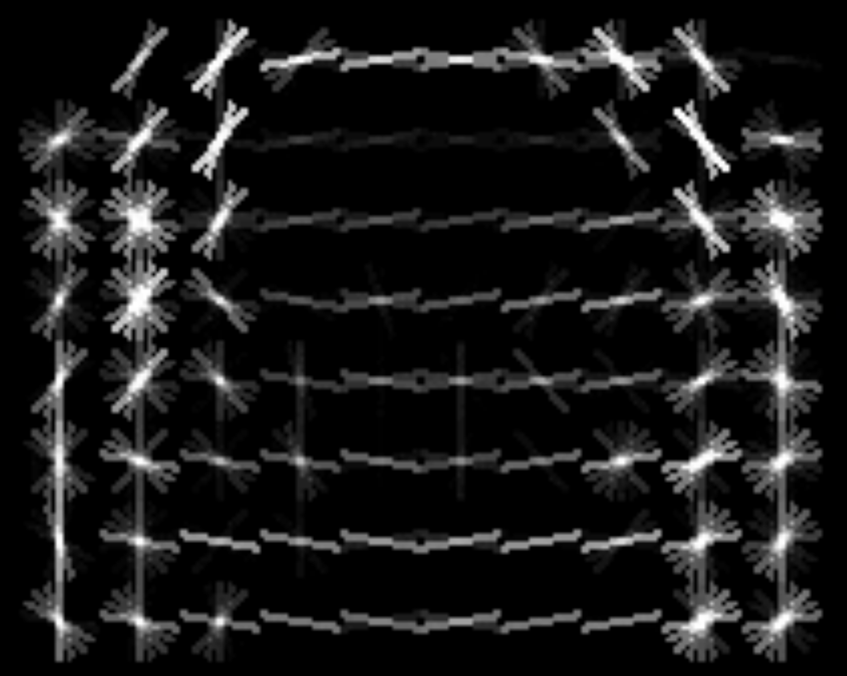
Max Planck Institute for Intelligent Systems
Tübingen

Example visualization of a learned 3D^{2}PM-C full 3D object model

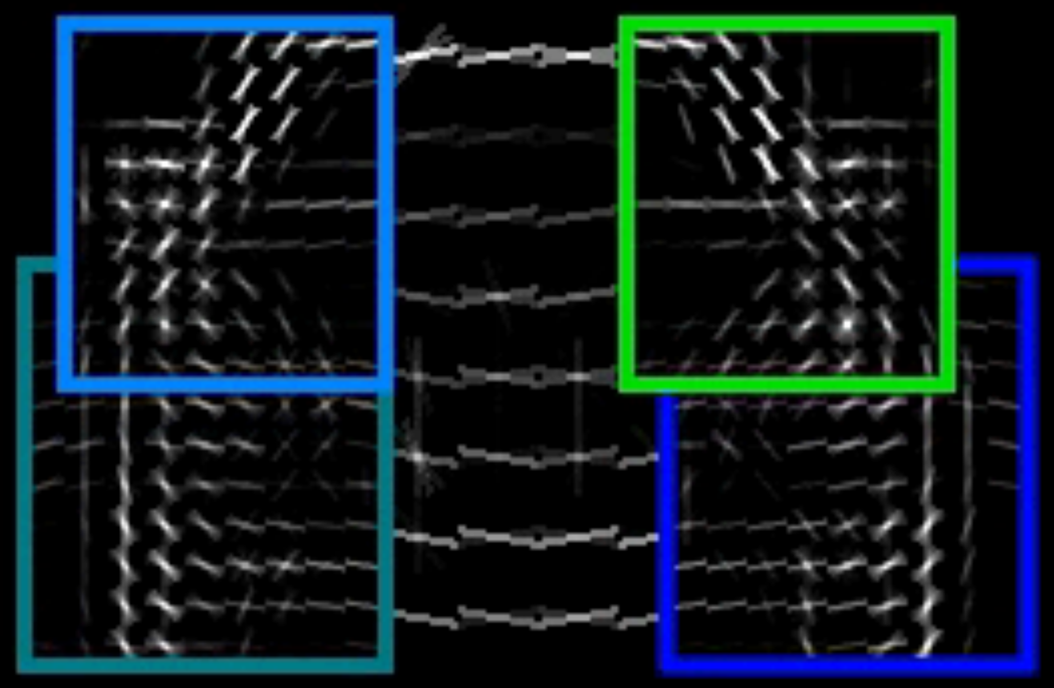
Model of 3D part displacements

2D Projected part displacements

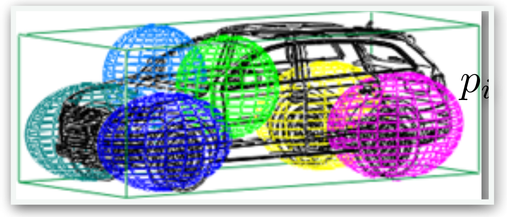




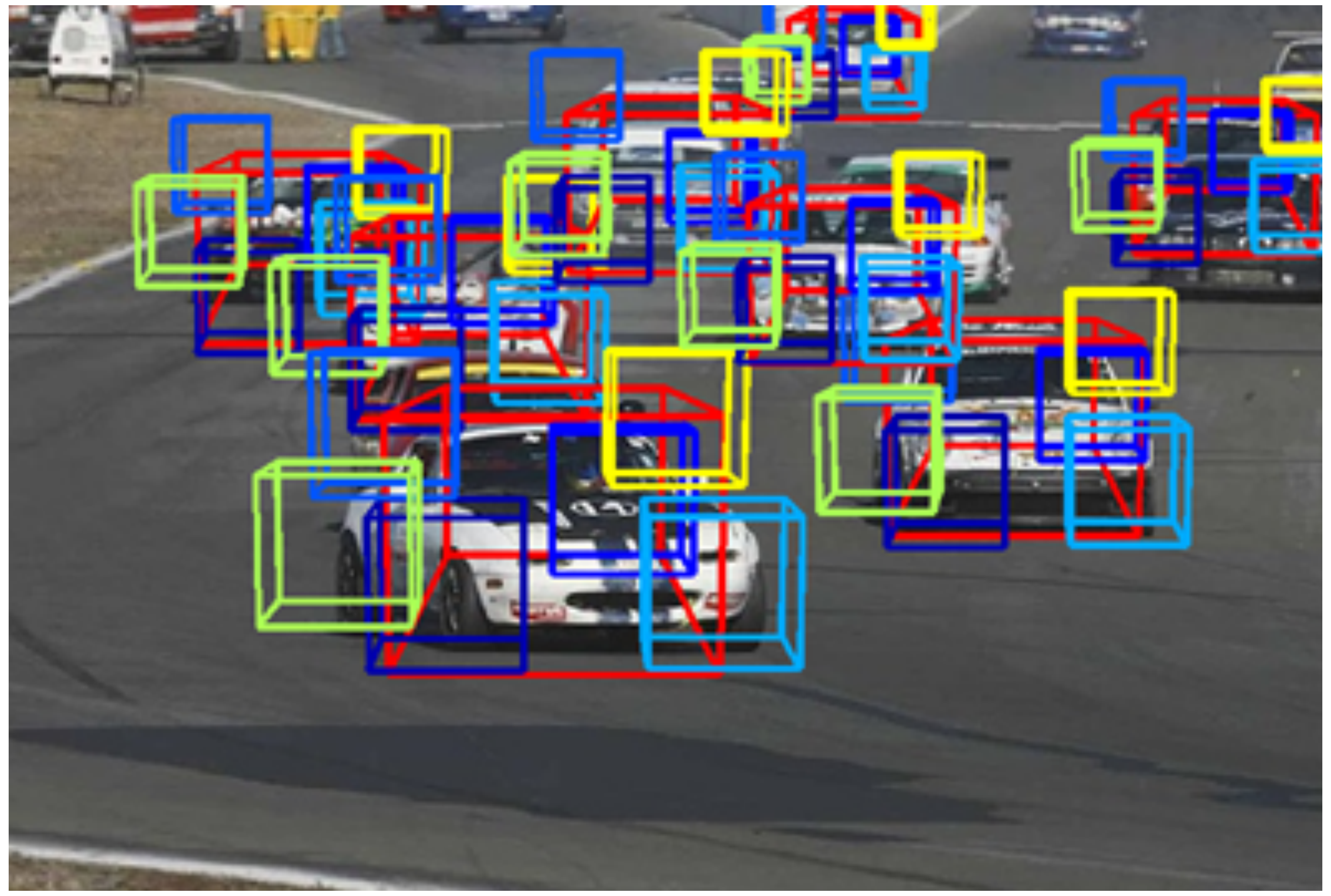
Interpolated appearance model of root part

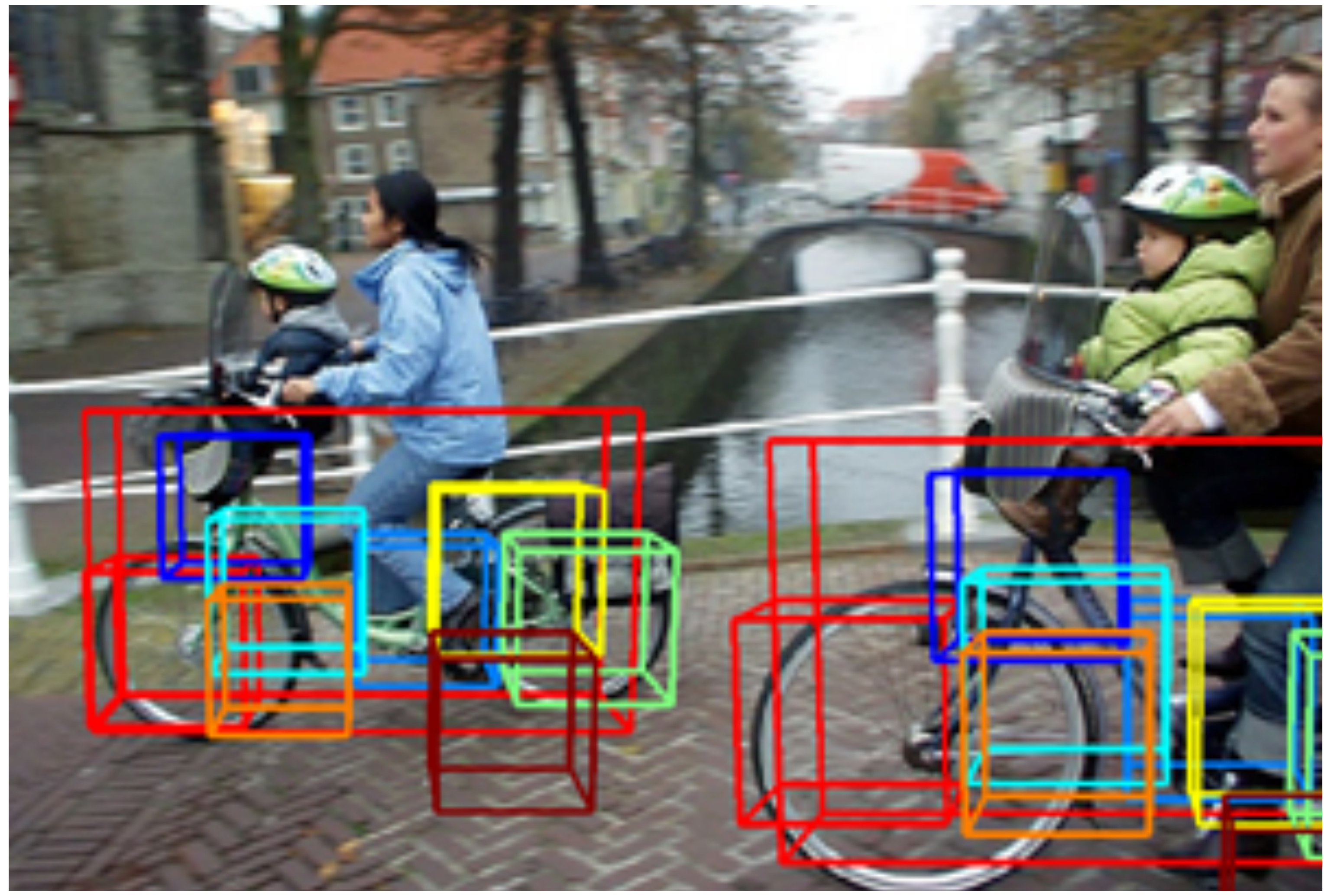


Interpolated appearance model of other parts



Detection + Geometry



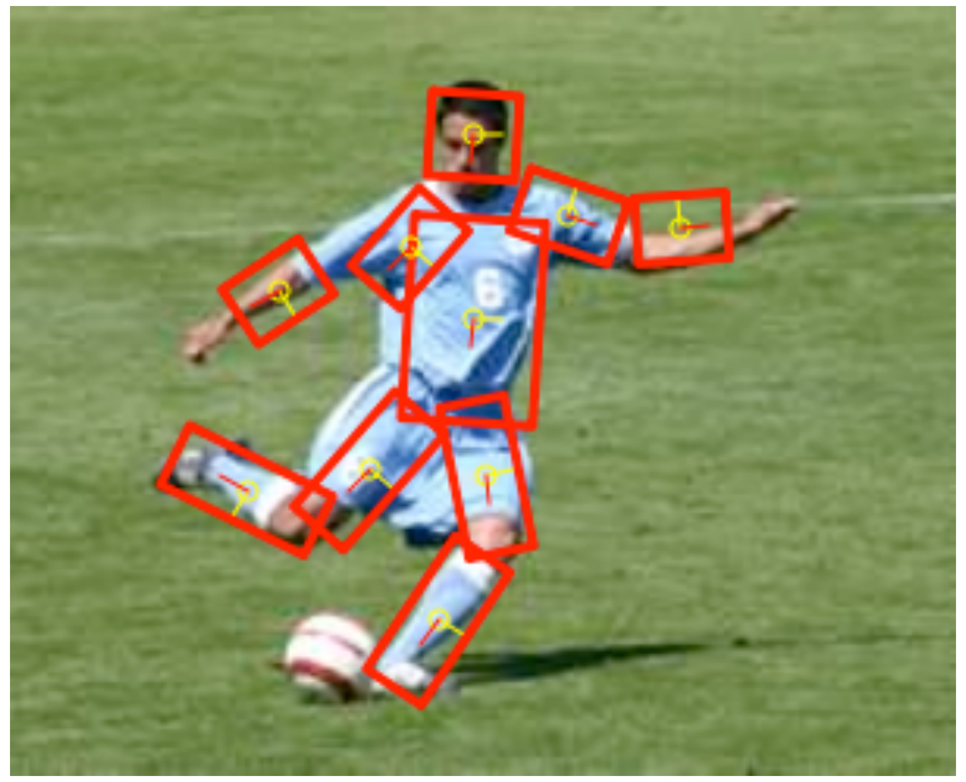


Fundamental matrix estimation experiment

A pair of images of the same instance

Human Pose Estimation

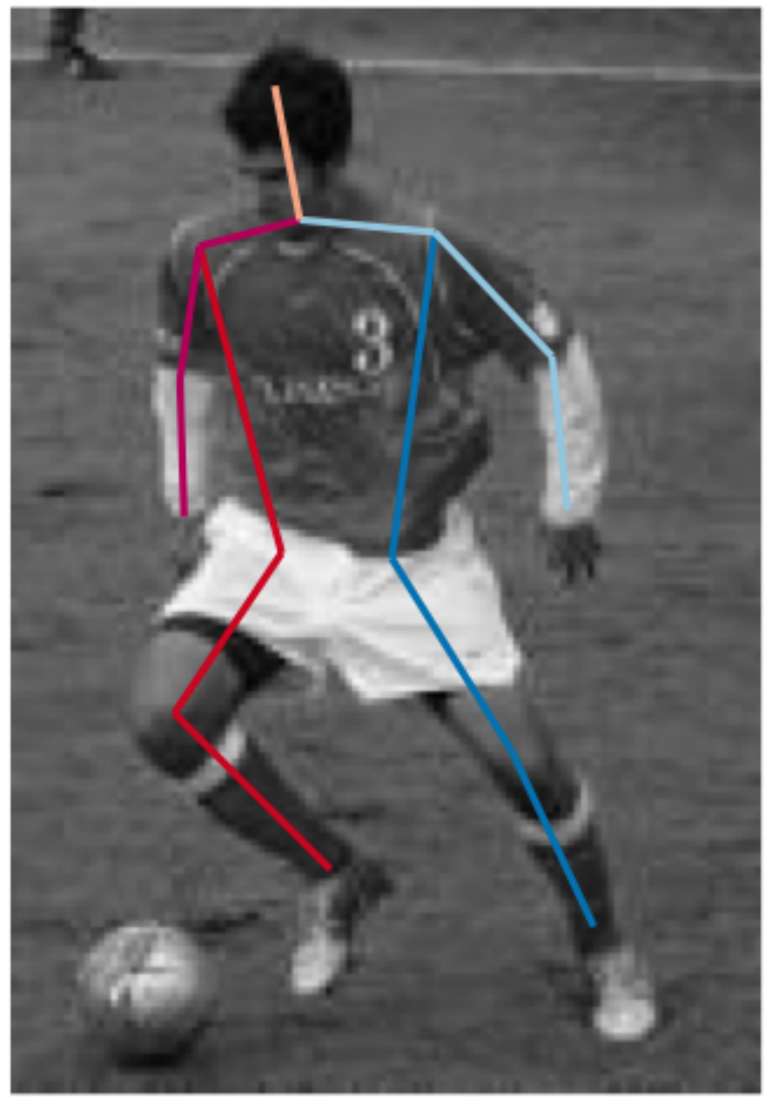
Observation



Predict
Bounding Boxes

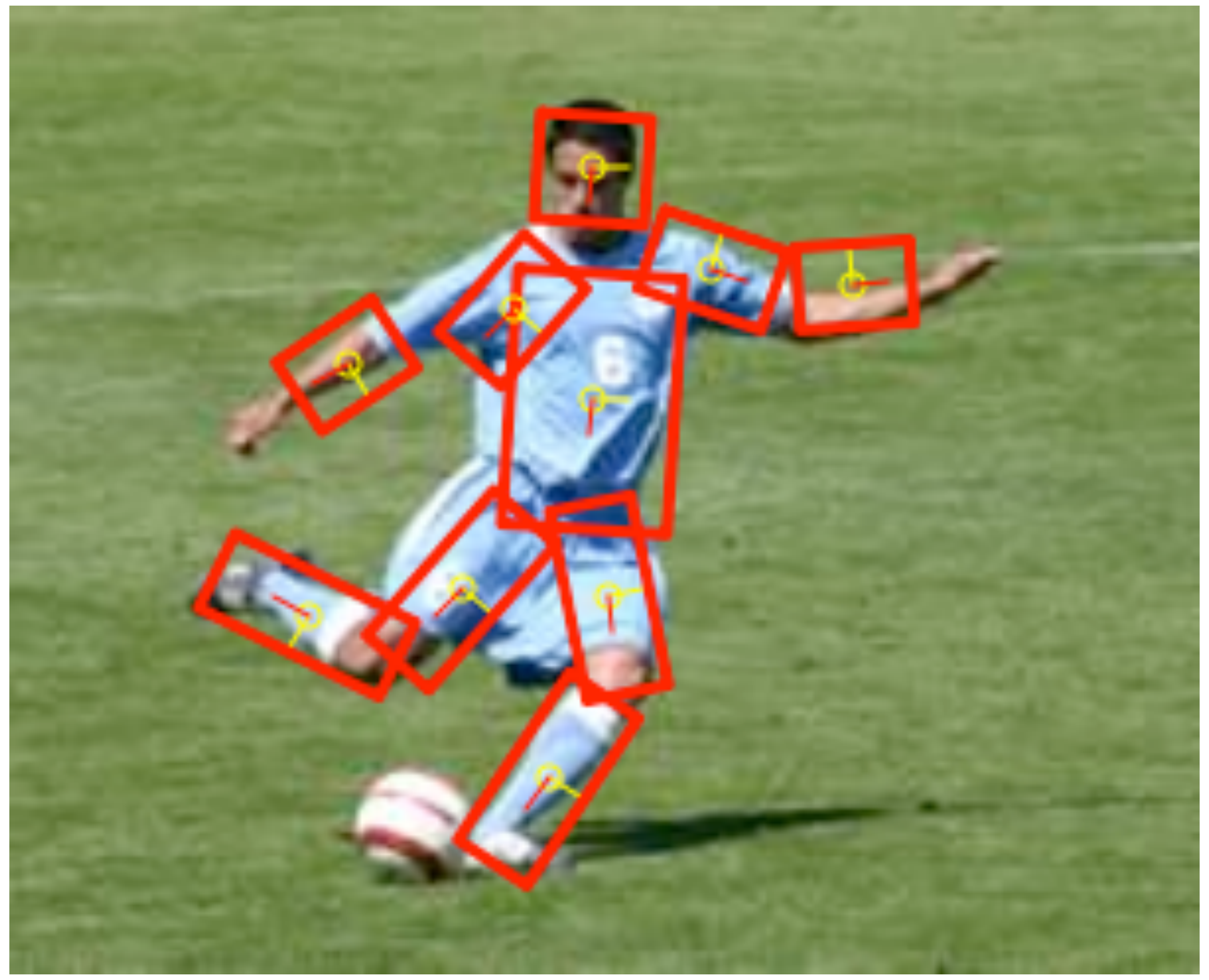
or

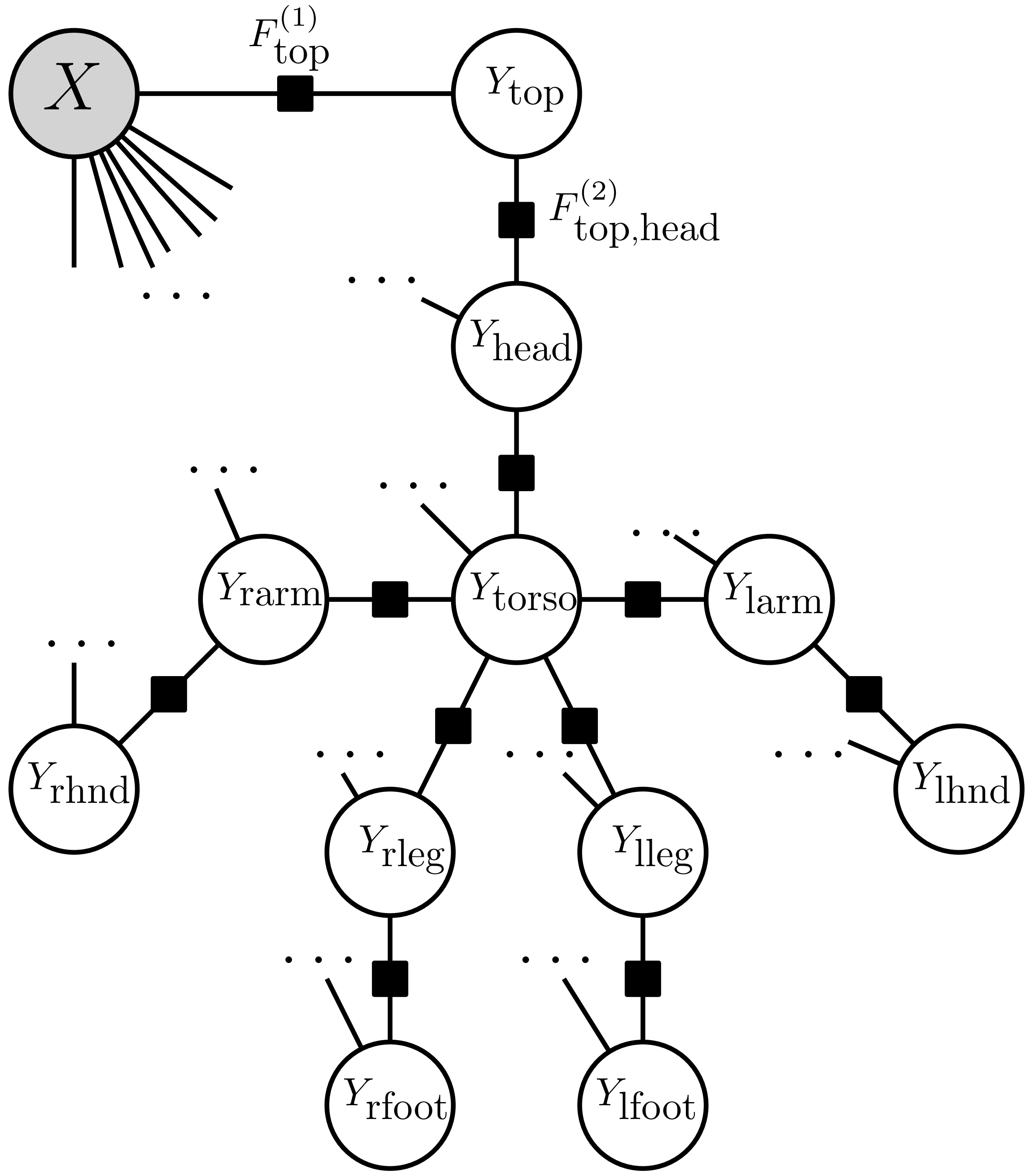
Observation



Predict
Joint Locations

Human Pose Estimation





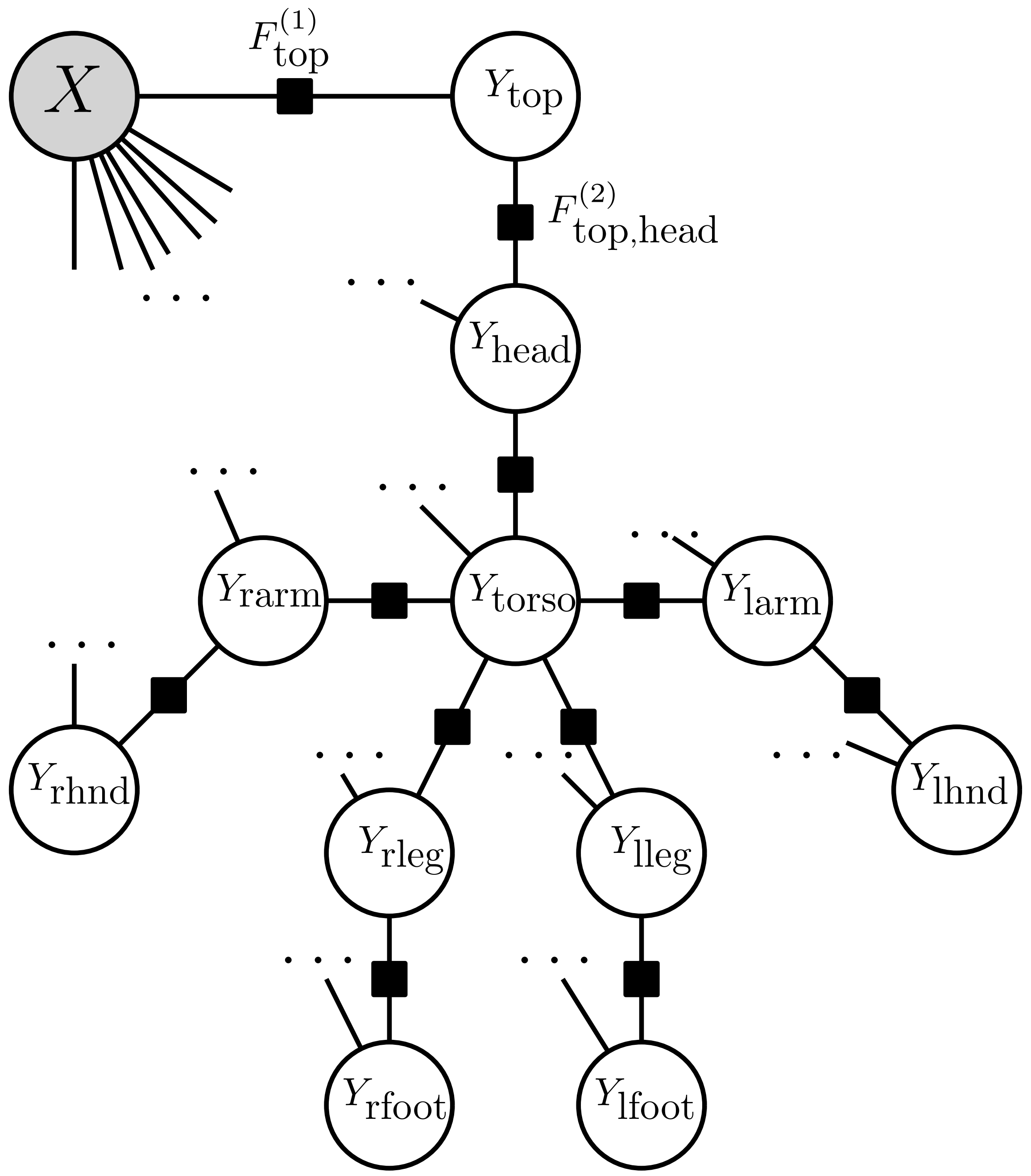
Observation

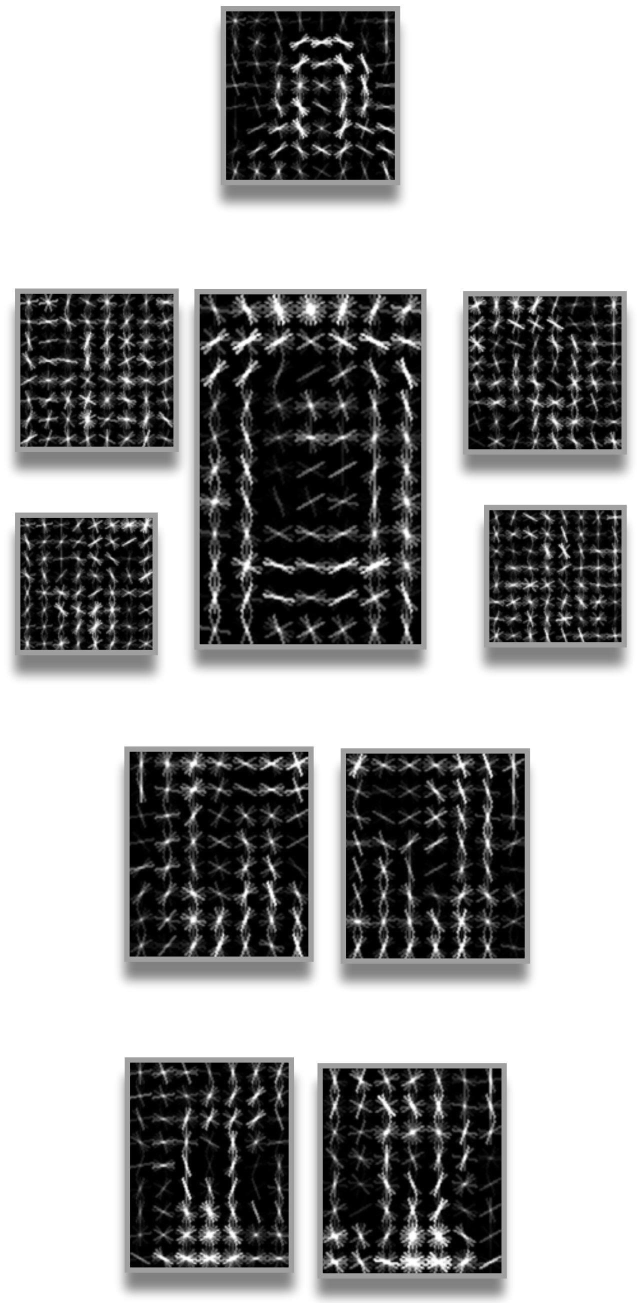
Desired Output

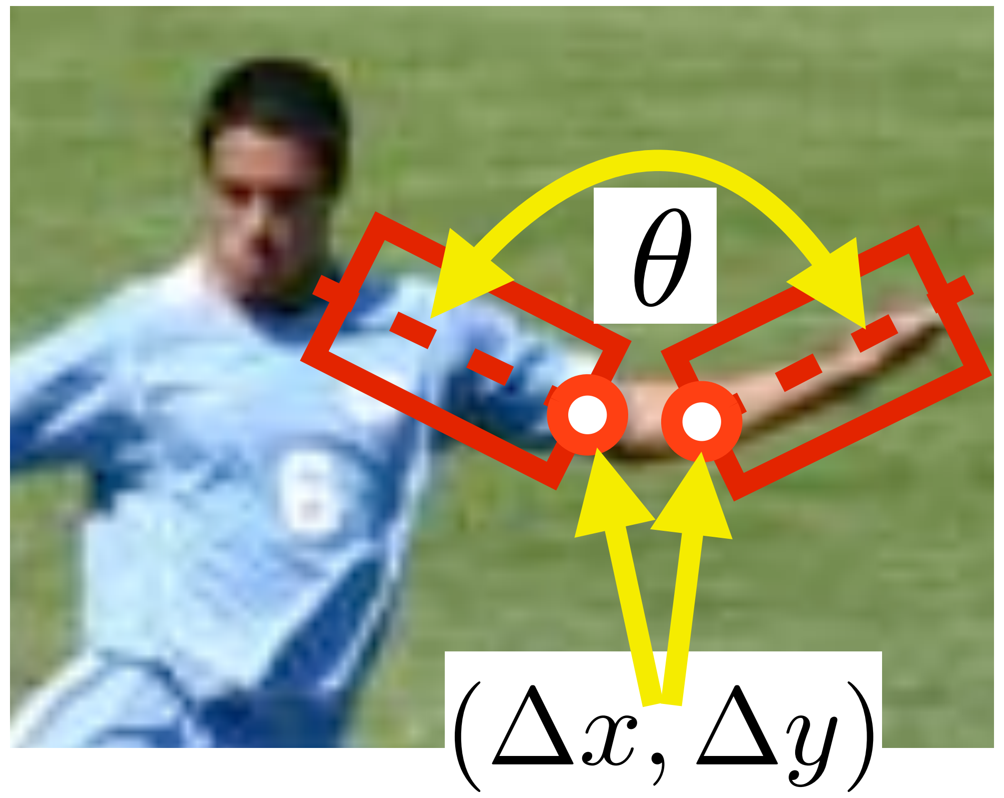
$$p(y|I,w) \propto \sum_{p} \psi(y^p, I; w) + \sum_{p \sim p'} \psi(y^p, y^{p'}; w)$$

P. Felzenszwalb, D. Huttenlocher, **Pictorial Structures for Object Recognition** International Journal of Computer Vision (IJCV), 2005

Pictorial Structures







$$p(y|I,w) \propto \sum_p \psi(y^p; I, w) + \sum_{p\sim p'} \psi(y^p, y^{p'}; I, w)$$

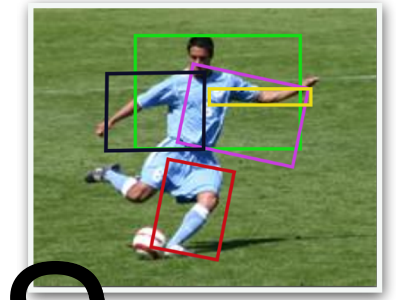
[Johnson&Everingham, BMVC'10], [Yang&Ramanan, CVPR'11],[Eichner&Ferrari, ACCV'12], [Sapp et al., ECCV'10], [Tran&Forsyth, ECCV'10], [Wang et al., CVPR'11], [Agarwal&Triggs, PAMI'02], [Urtasun&Darrell, ICCV'09], [Ionescu et al., ICCV'11]

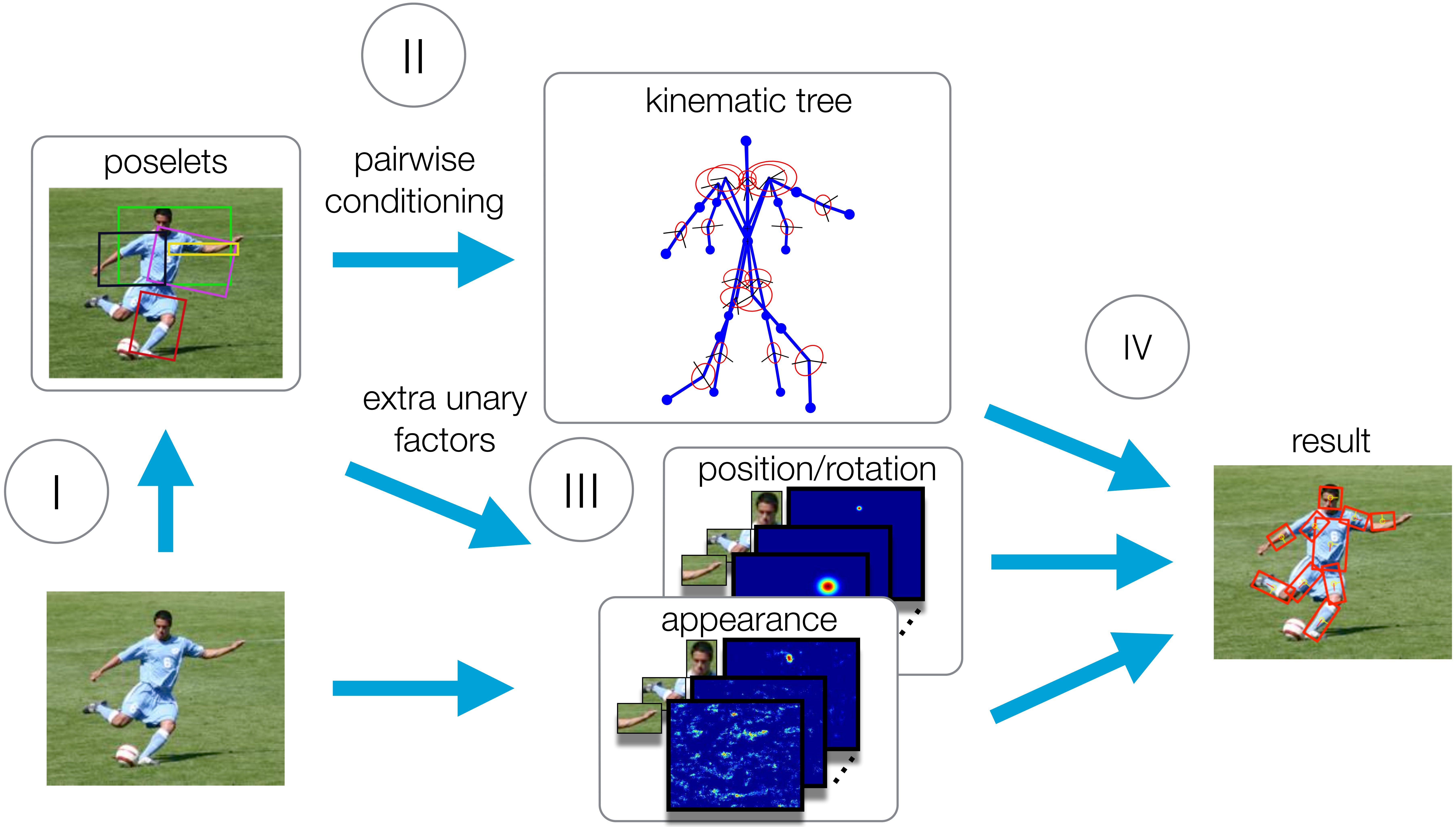
Extensions

- Ever since introduced many extensions are proposed:
- loopy …
- mixture …
- holistic approaches…

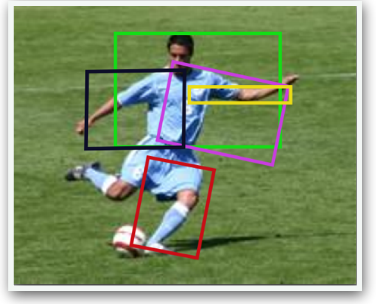
[Johnson&Everingham, BMVC'10]
[Yang&Ramanan, CVPR'11]
[Eichner&Ferrari, ACCV'12]
[Sapp et al., ECCV'10]
[Tran&Forsyth, ECCV'10]
[Wang et al., CVPR'11]
[Agarwal&Triggs, PAMI'02]
[Urtasun&Darrell, ICCV'09]
[Ionescu et al., ICCV'11]
…

Poselet Conditioned Pictorial Structures



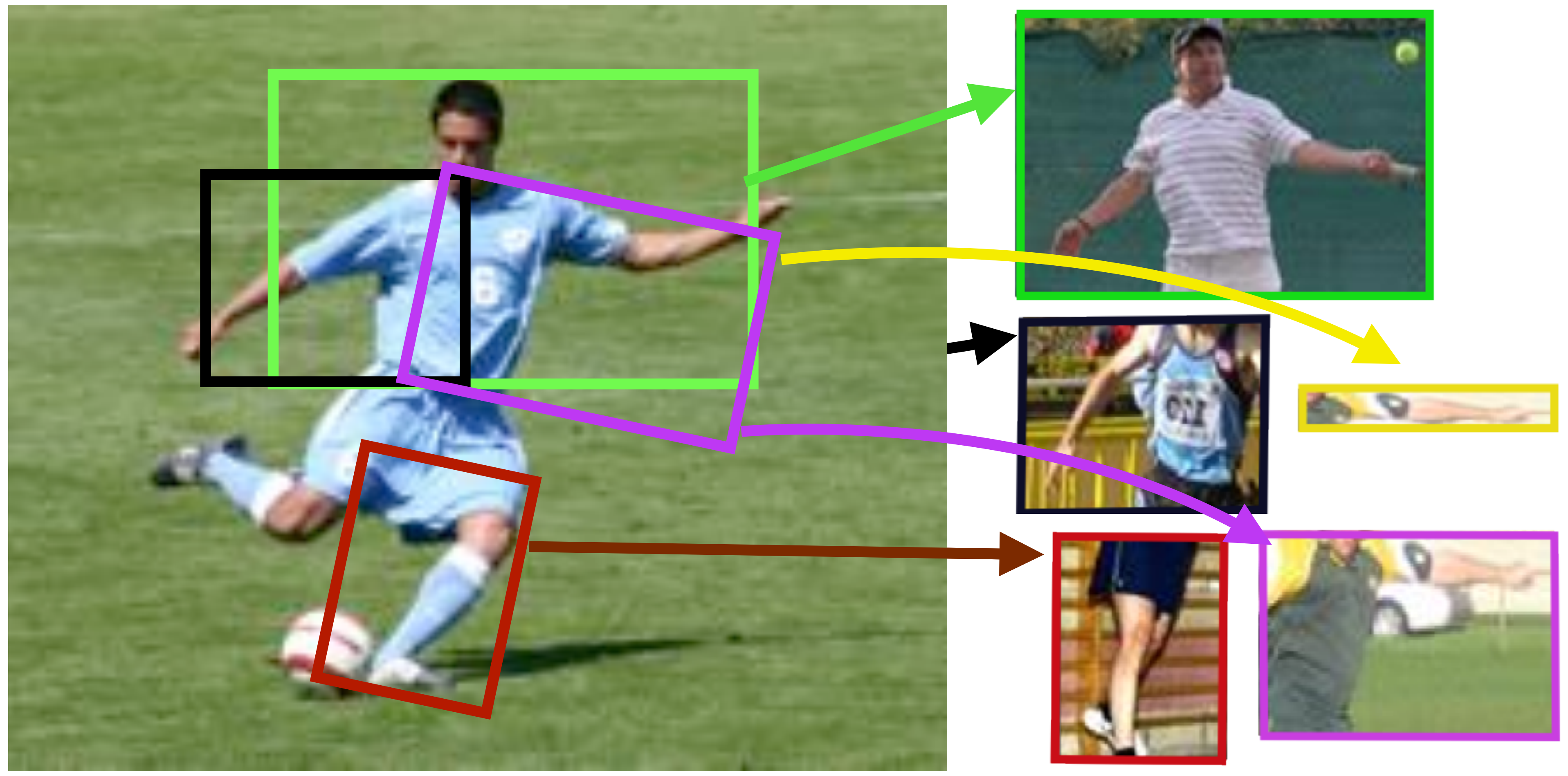


L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele **Poselet Conditioned Pictorial Structures,** CVPR 2013

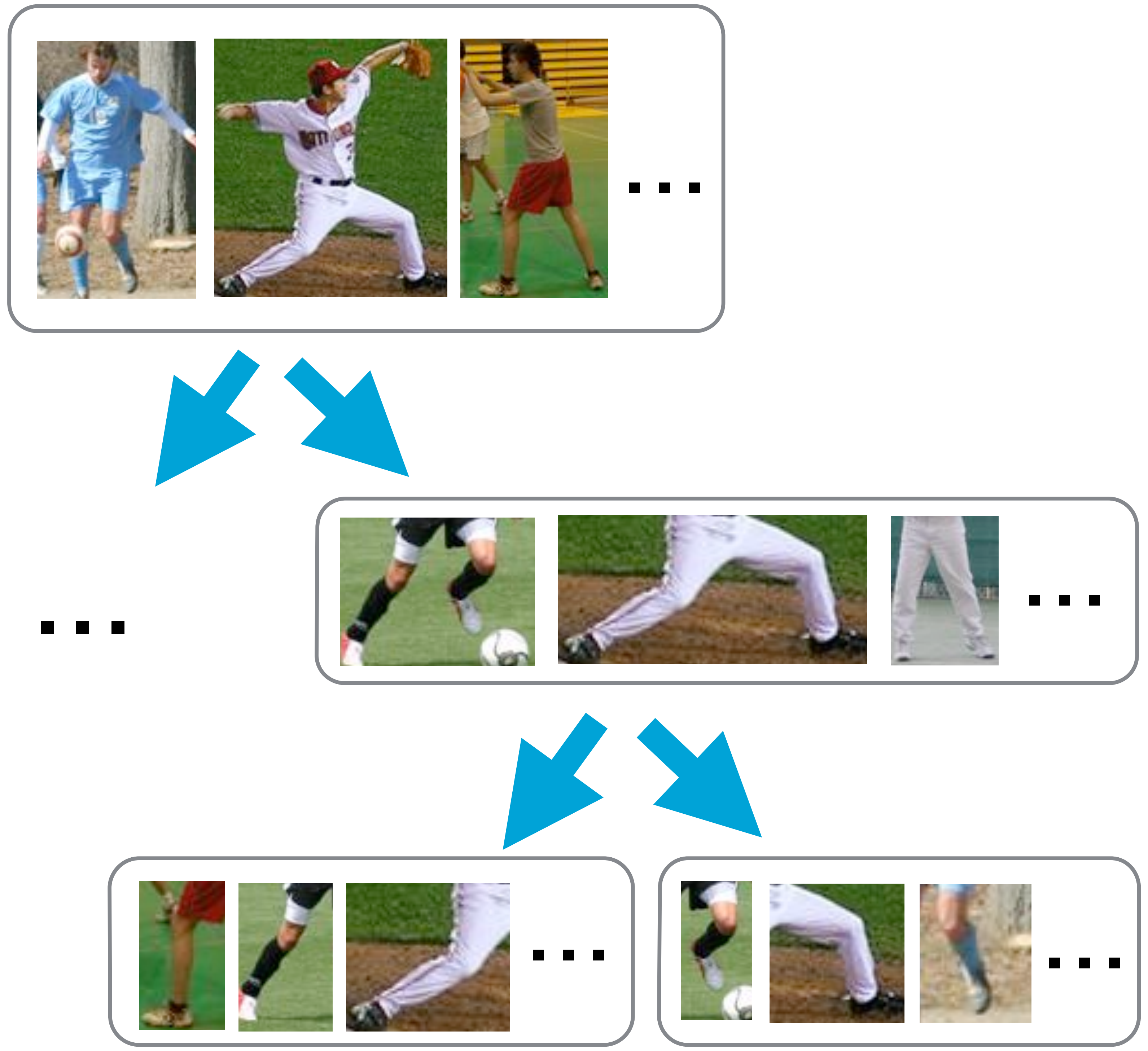


Poselets

- "Clusters" of more parts
- Capture non-adjacent part dependencies

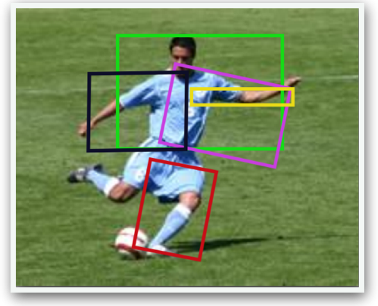


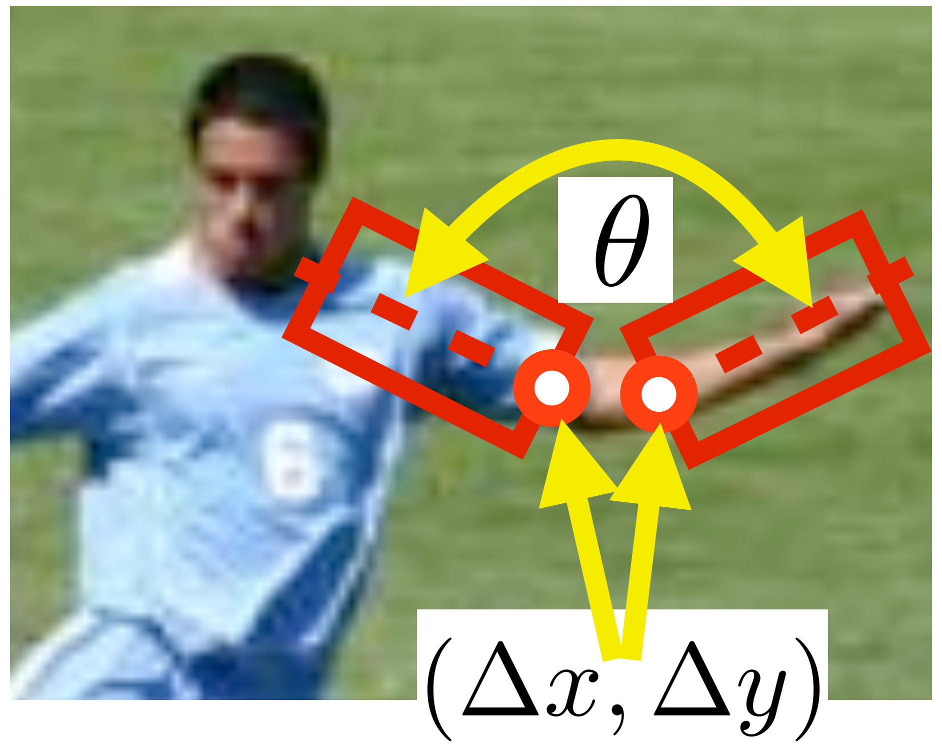
Top detections Poselet cluster medoids

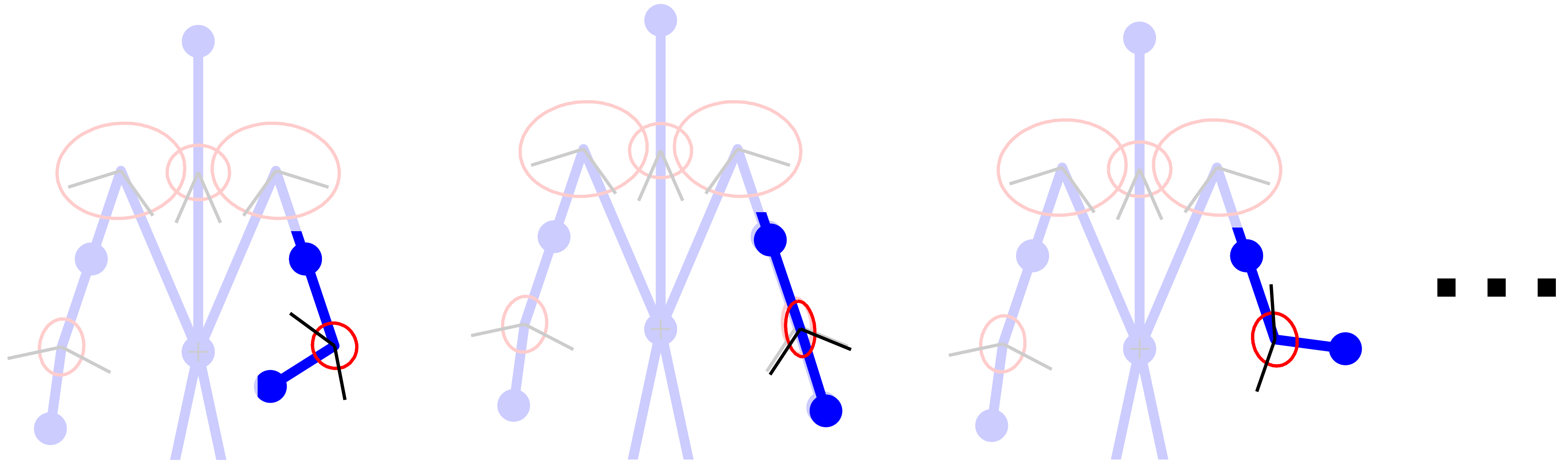


L. Bourdev, J. Malik, **Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations**, ICCV 2009

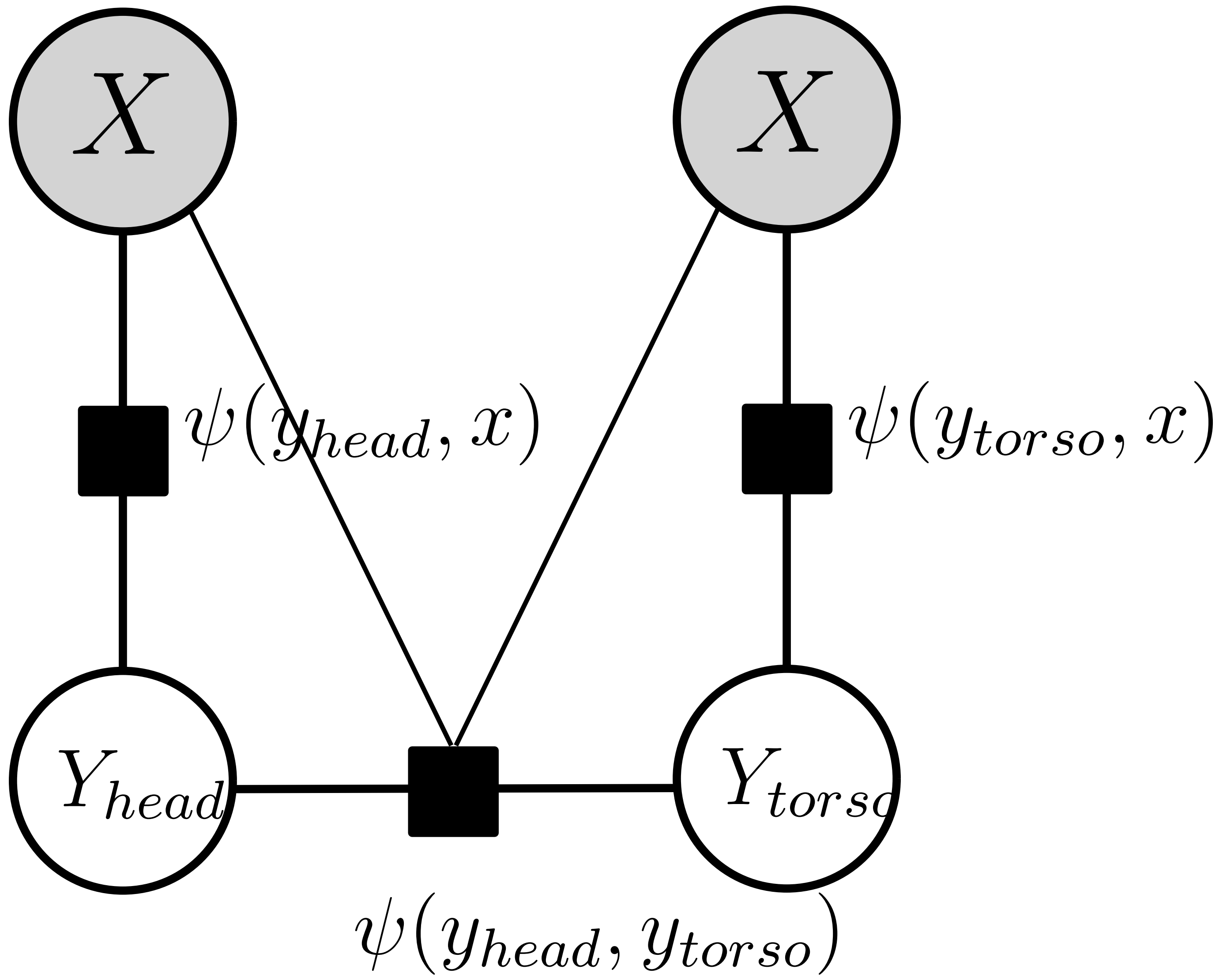
Conditioning Pairwise Terms

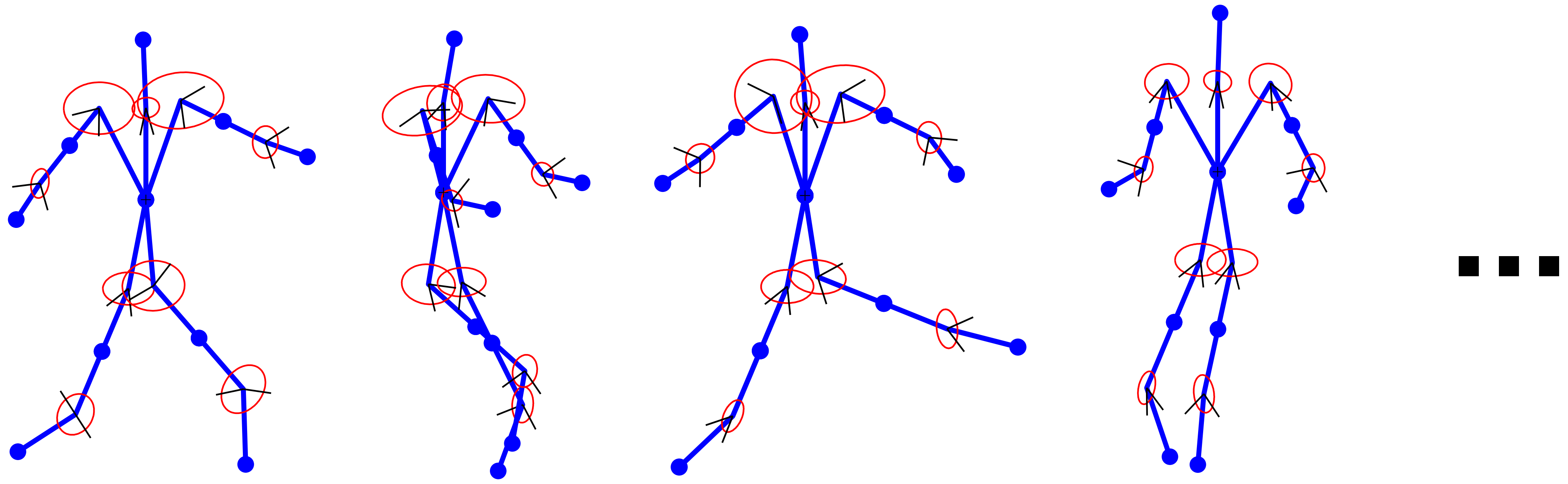




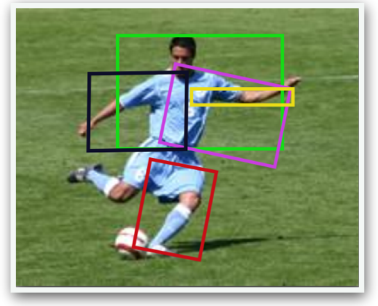


Possible pairwise factors





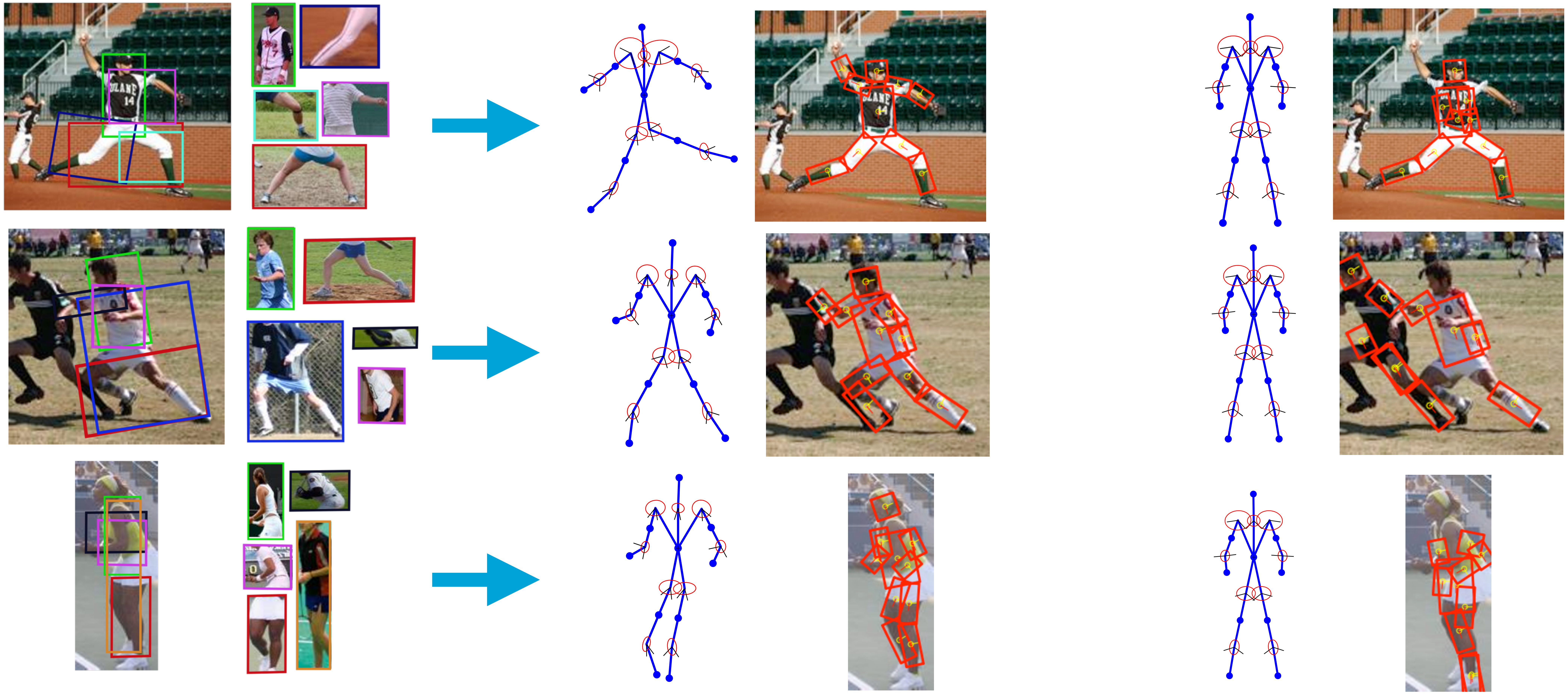
Possible body models



Results

Poselet Conditioned

Baseline PS



Top poselet detections

Cluster medoids

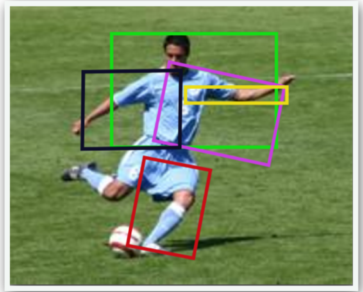
Prediction

Result

Generic Tree

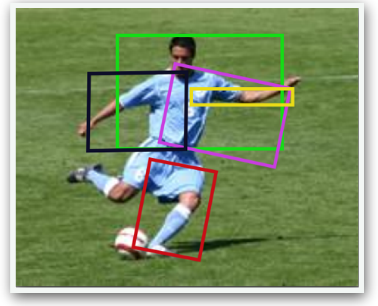
Result

Results on Leeds Sports Poses



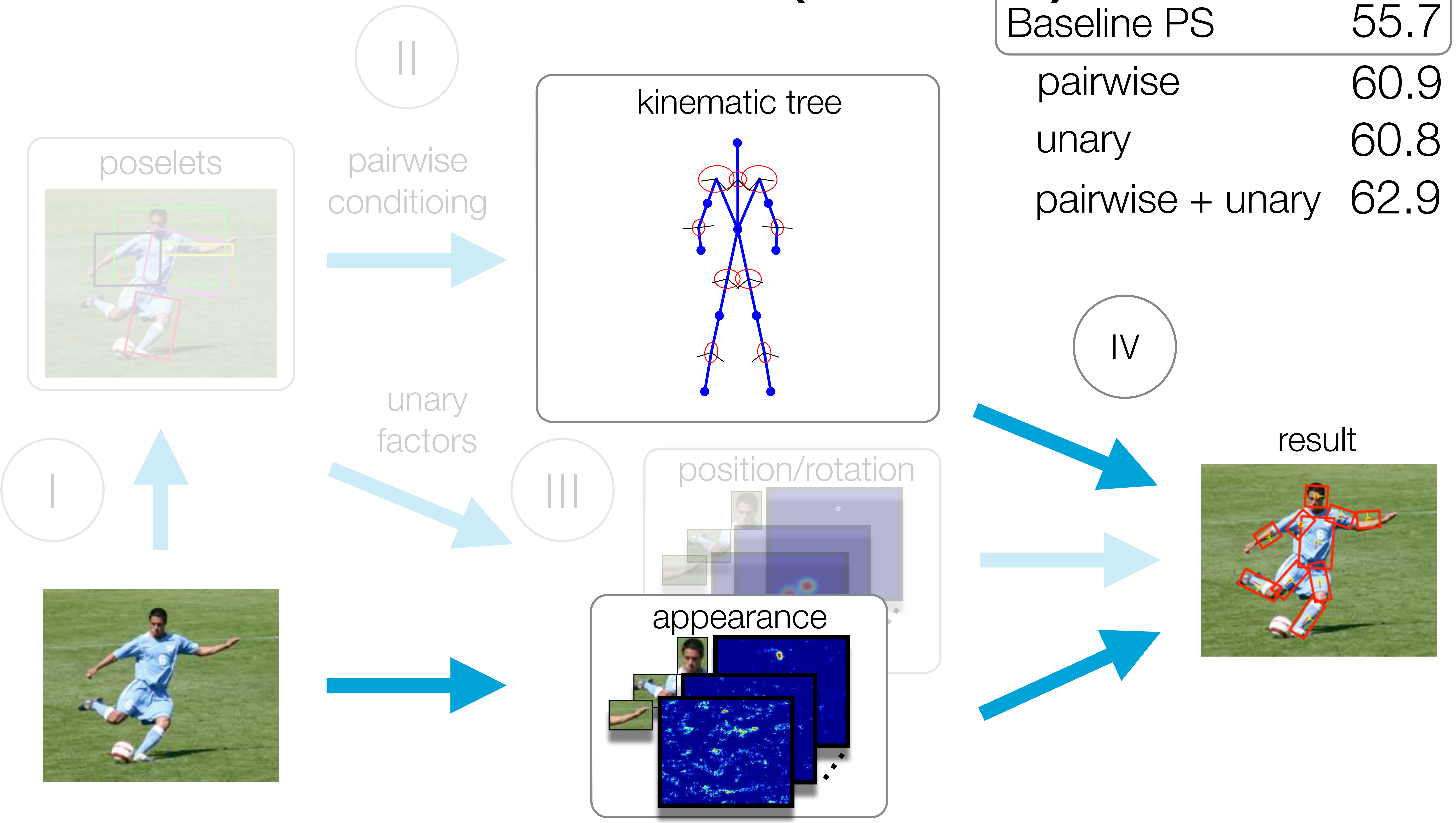
S. Johnson, M. Everingham, **Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation**, BMVC 2010

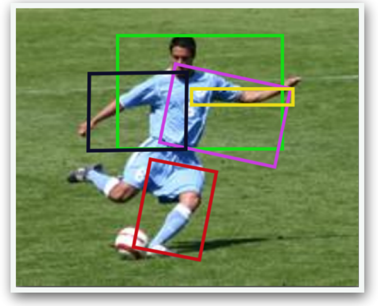
1000 training, 1000 testing images
observer centric annotation [Eichner&Ferrari, ACCV12]
Error: PCP percentage of correct parts



Results (PCP)

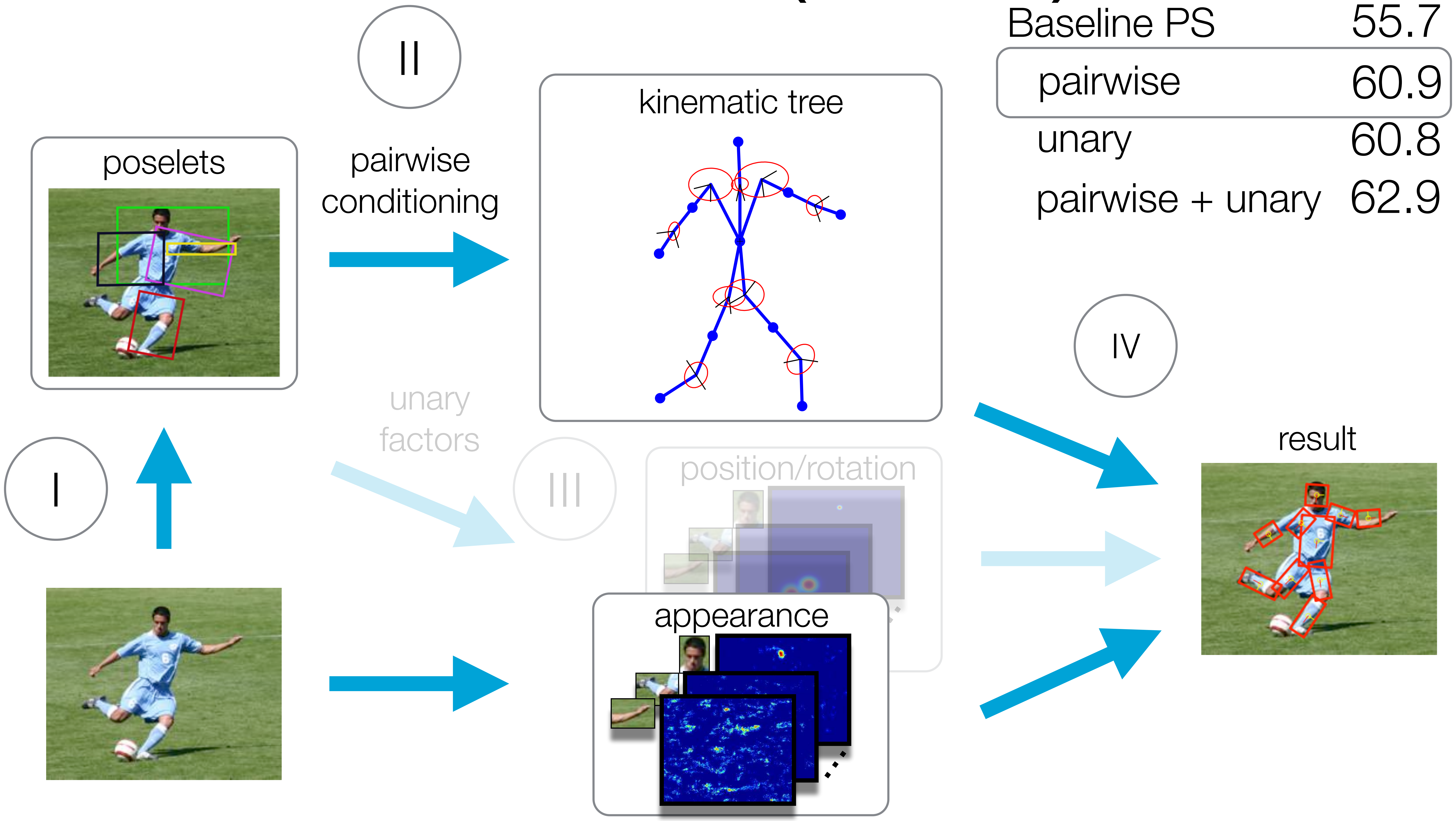
Baseline PS	55.7
pairwise	60.9
unary	60.8
pairwise + unary	62.9





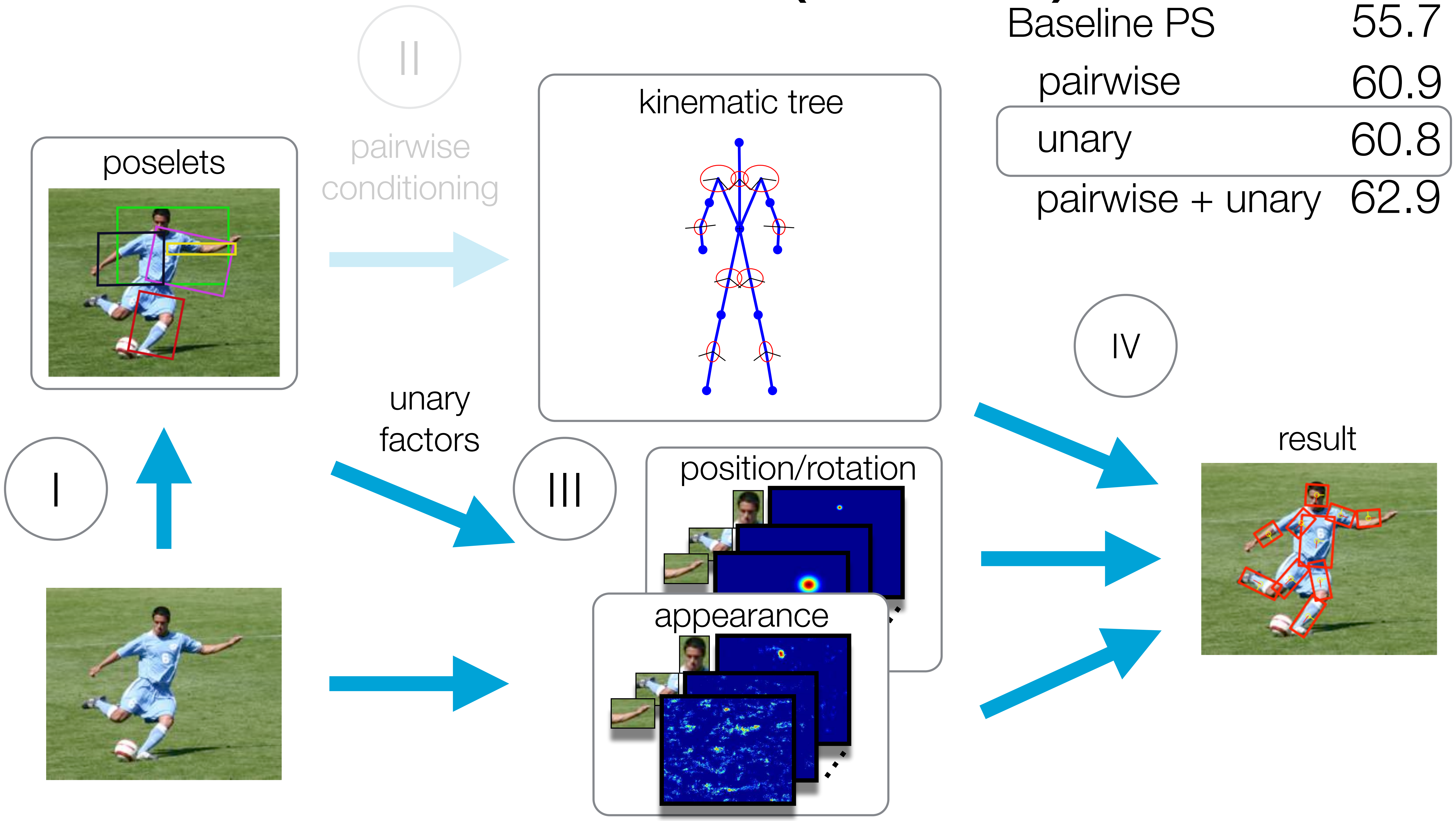
Results (PCP)

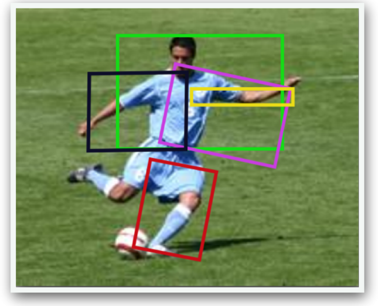
Baseline PS	55.7
pairwise	60.9
unary	60.8
pairwise + unary	62.9



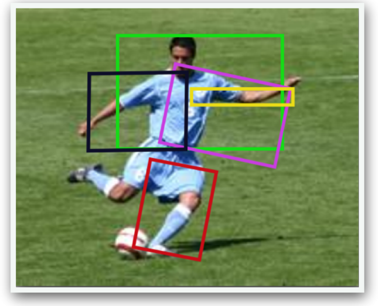
Results (PCP)

Baseline PS	55.7
pairwise	60.9
unary	60.8
pairwise + unary	62.9

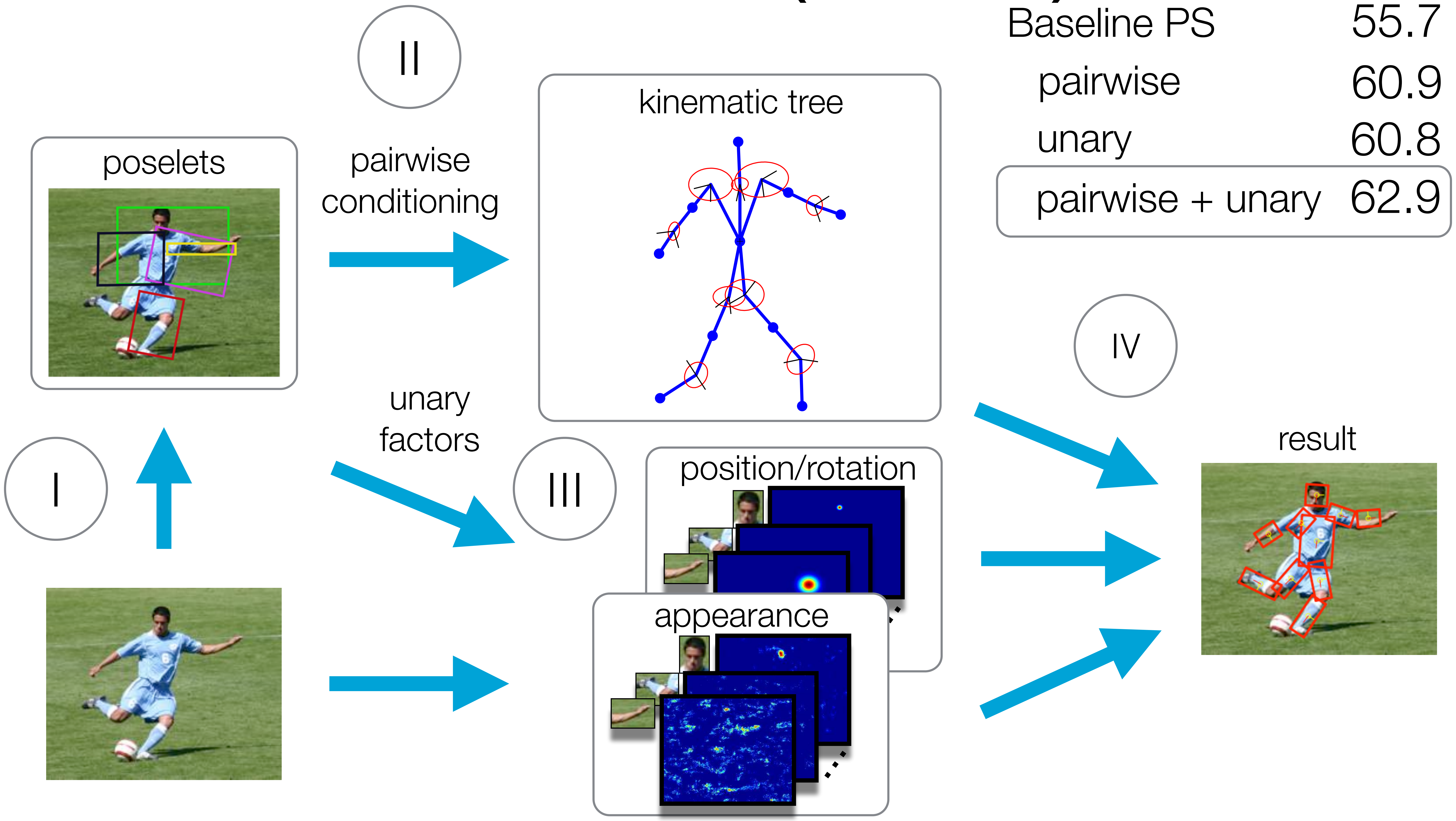


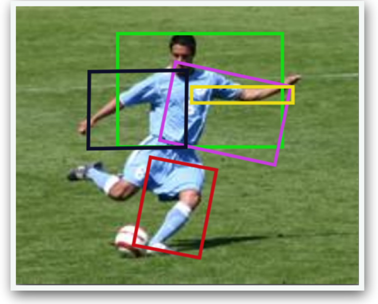


Results (PCP)

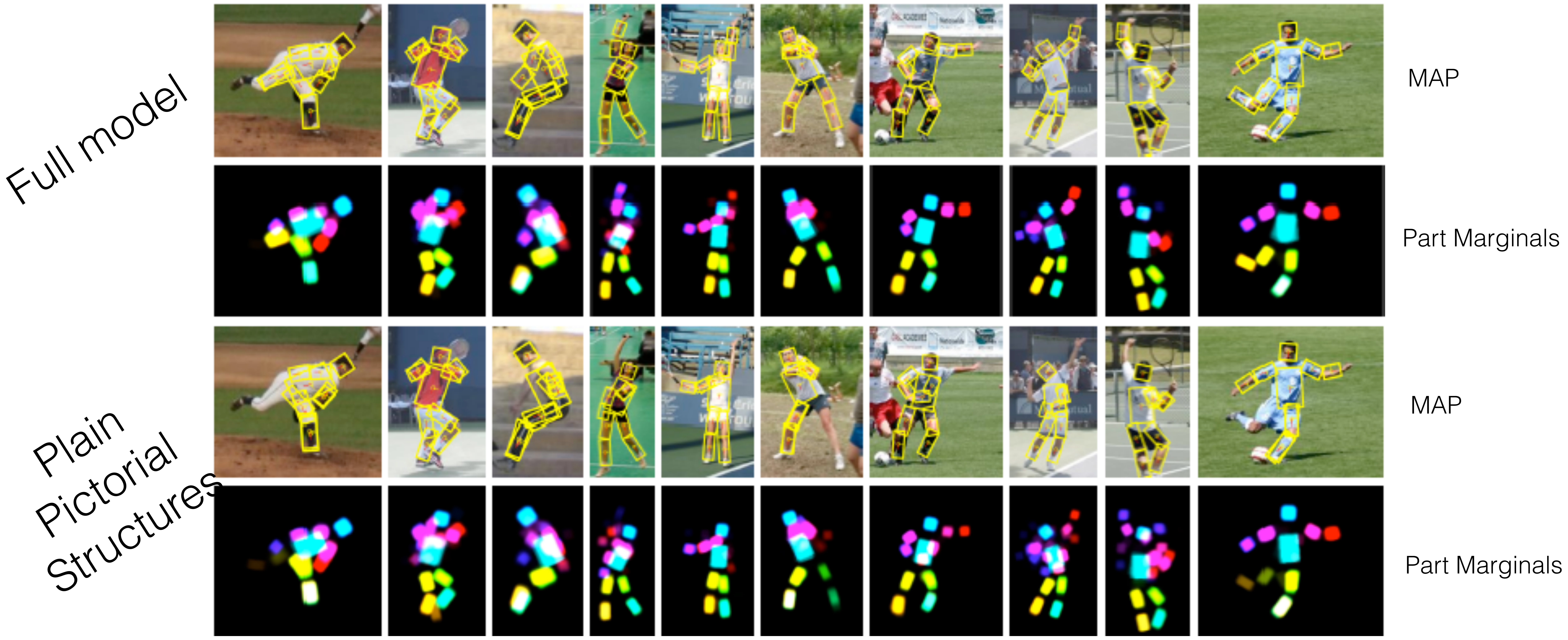


Baseline PS	55.7
pairwise	60.9
unary	60.8
pairwise + unary	62.9

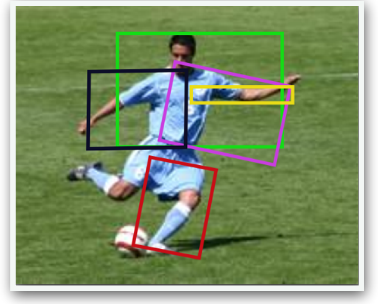




Results



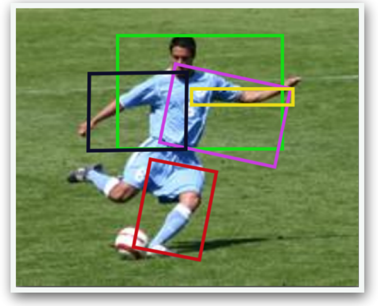
Full model
Plain Pictorial Structures
MAP
Part Marginals
MAP
Part Marginals



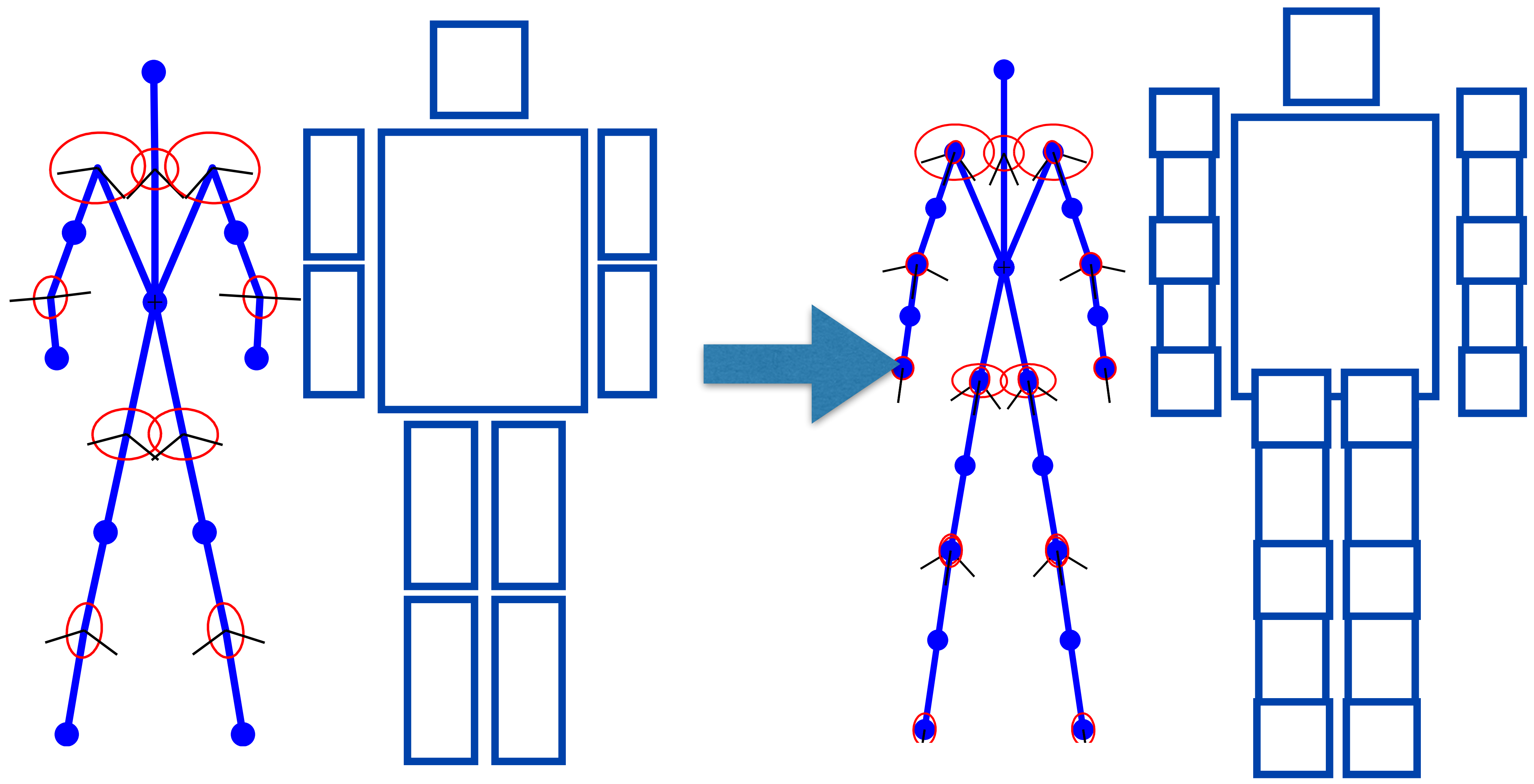
Only 62.9% ??? Why not 100%?

What are we missing?

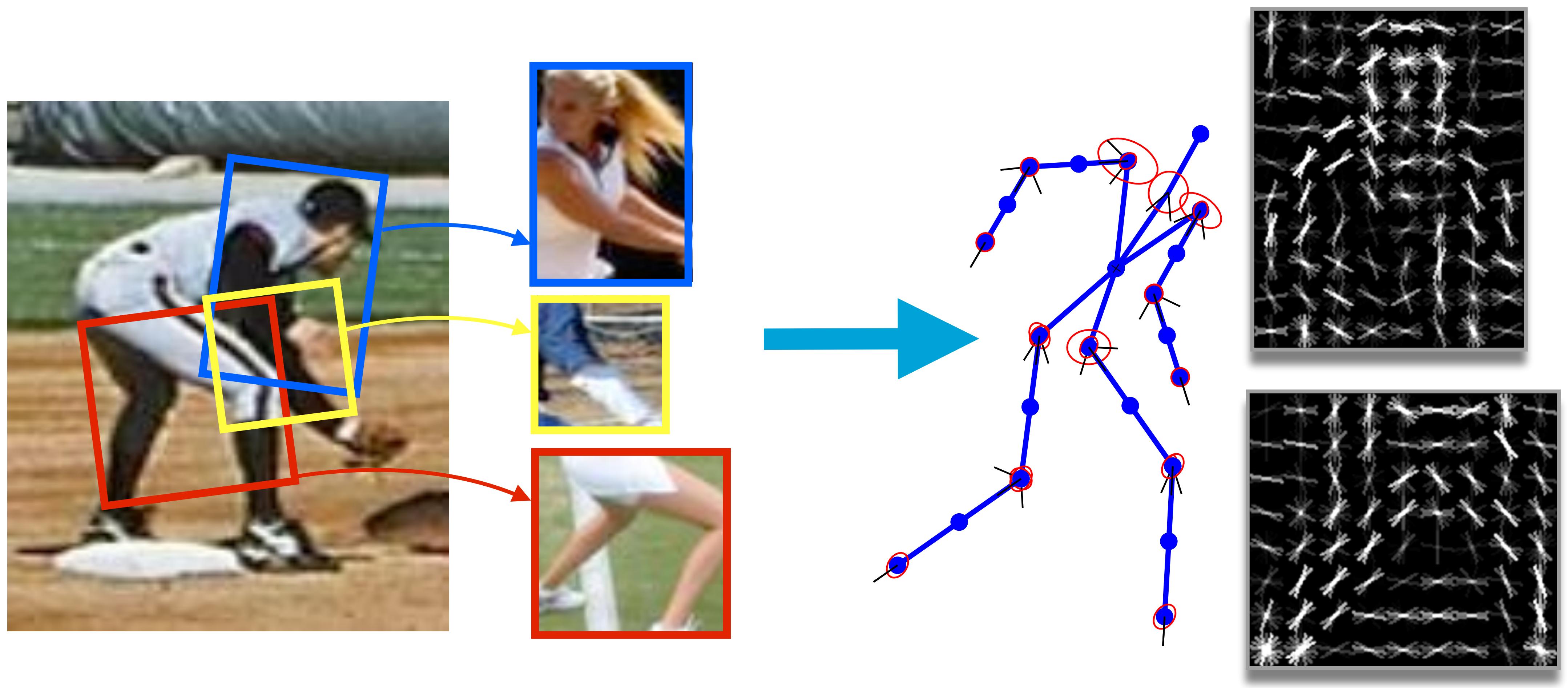
Expressive Spatial Models…



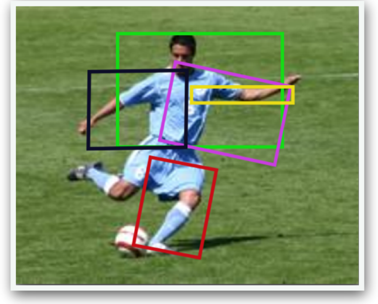
Joint model for body parts and body joints



Mid-Level representation

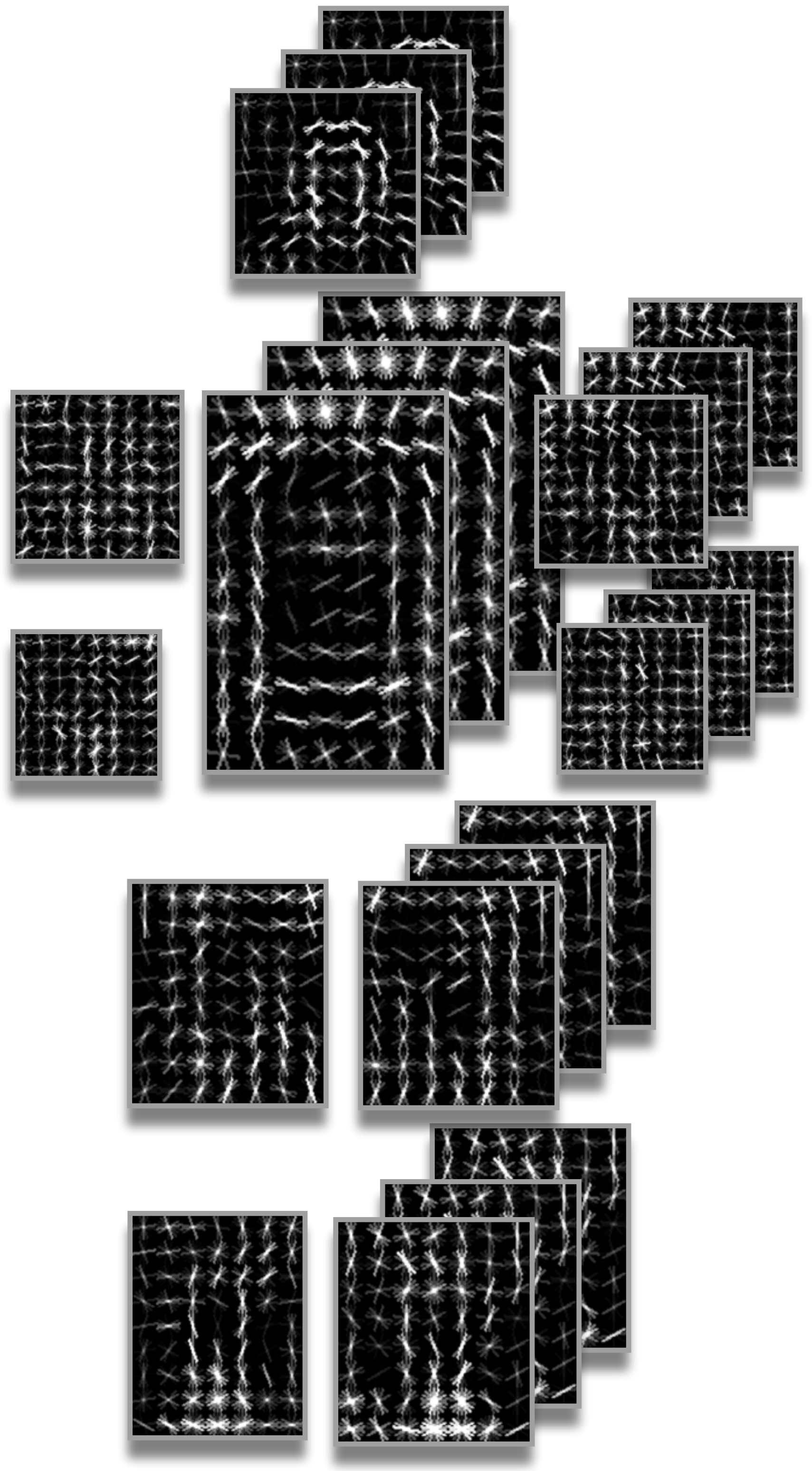


L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, **Strong Appearance and Expressive Spatial Models for Human Pose Estimation**, ICCV 2013

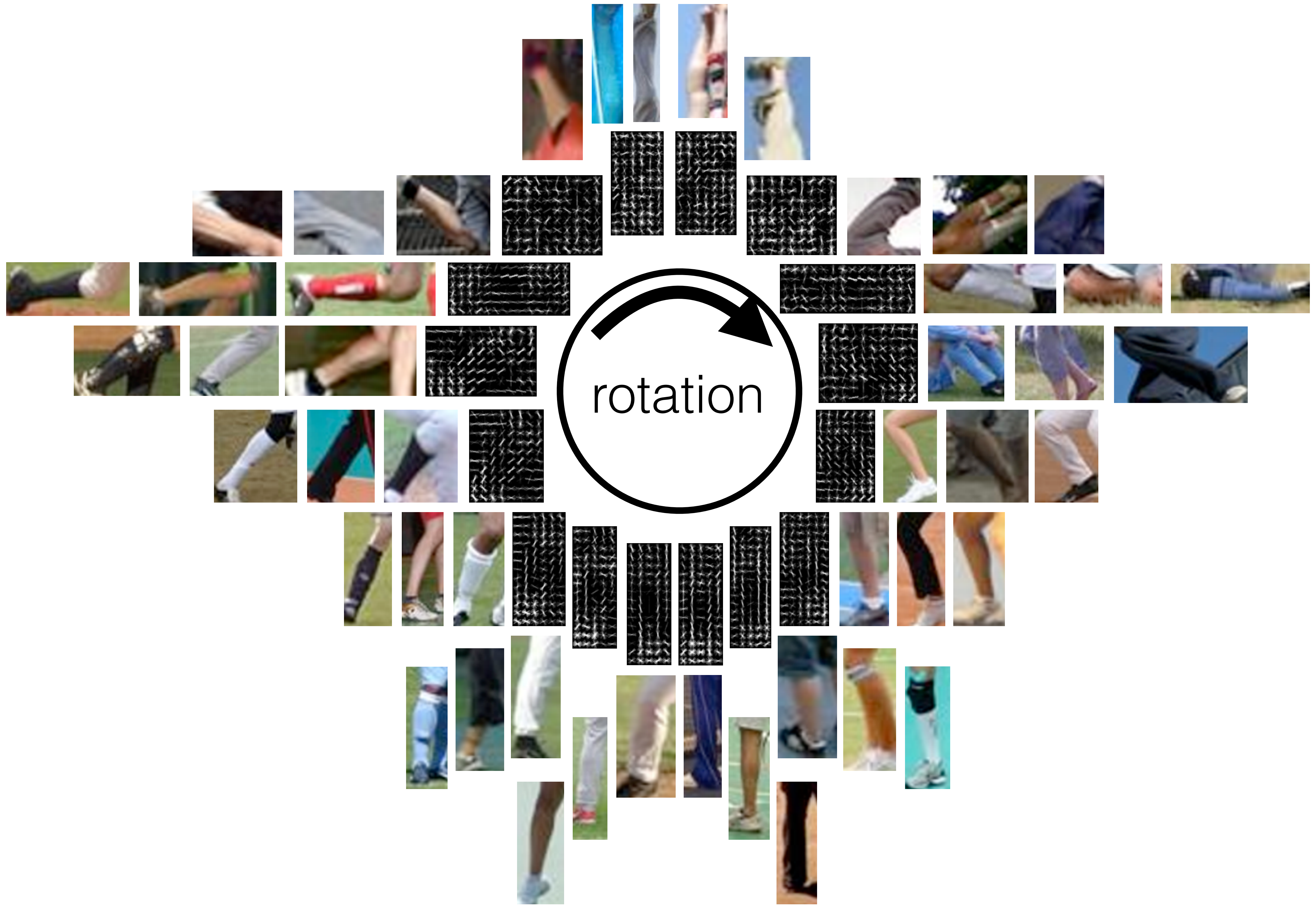


… and Strong Appearance

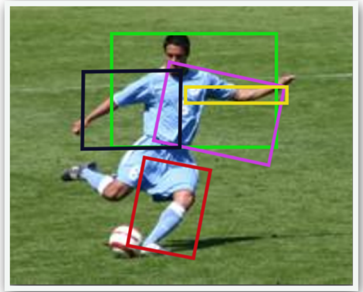
Mixtures of DPM for local Appearance



Rotation Dependent Part Detectors

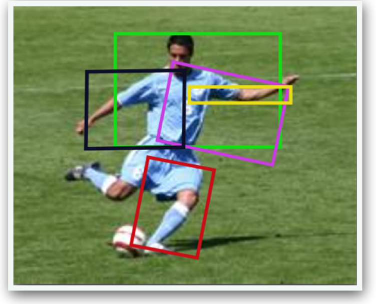


L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, **Strong Appearance and Expressive Spatial Models for Human Pose Estimation**, ICCV 2013



Empirical Results

Setting	PCP [%]	
model so far	62.9	
Andriluka et al. CVPR 09	55.7	
+ flexible body model	56.9	
+ local mixtures	65.2	
+ Poselet conditioned unaries	68.5	
+ Poselet conditioned pairwise	**69.0**	
Yang & Ramanan, CVPR 11	60.8	
Eichner & Ferrari, ACCV 12	64.3	
Ramakrishna et al. ECCV 14	67.6	(Pose Inference Machines)
Chen & Yuille arXiv 14	**76.6**	(CNNs)

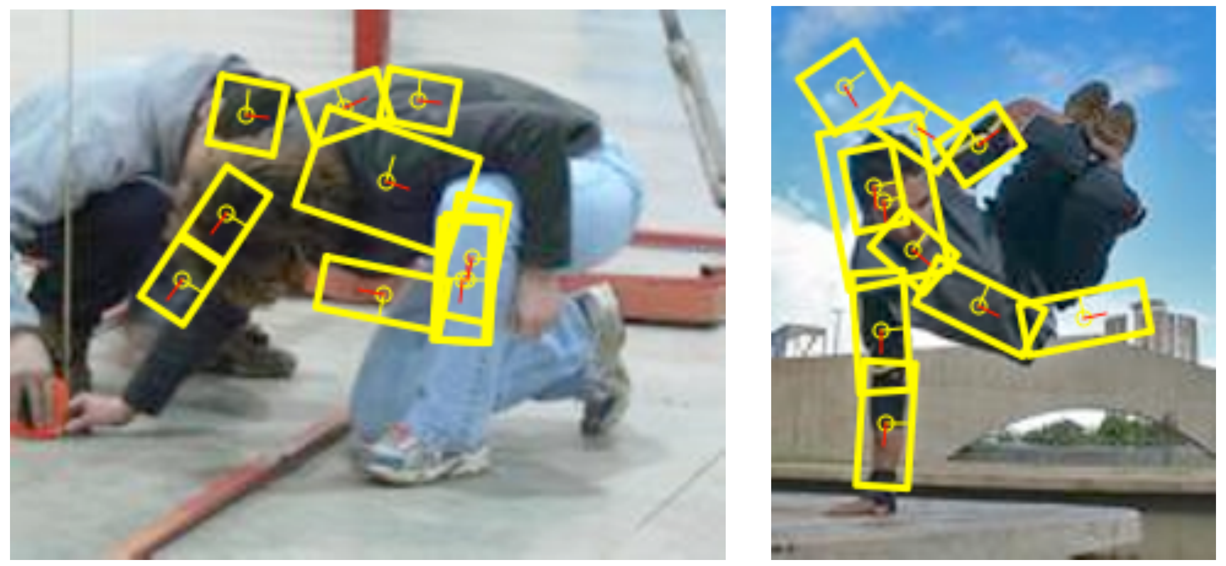


Still not perfect … ?

- All remaining failure cases are of these types

Self-occlusion

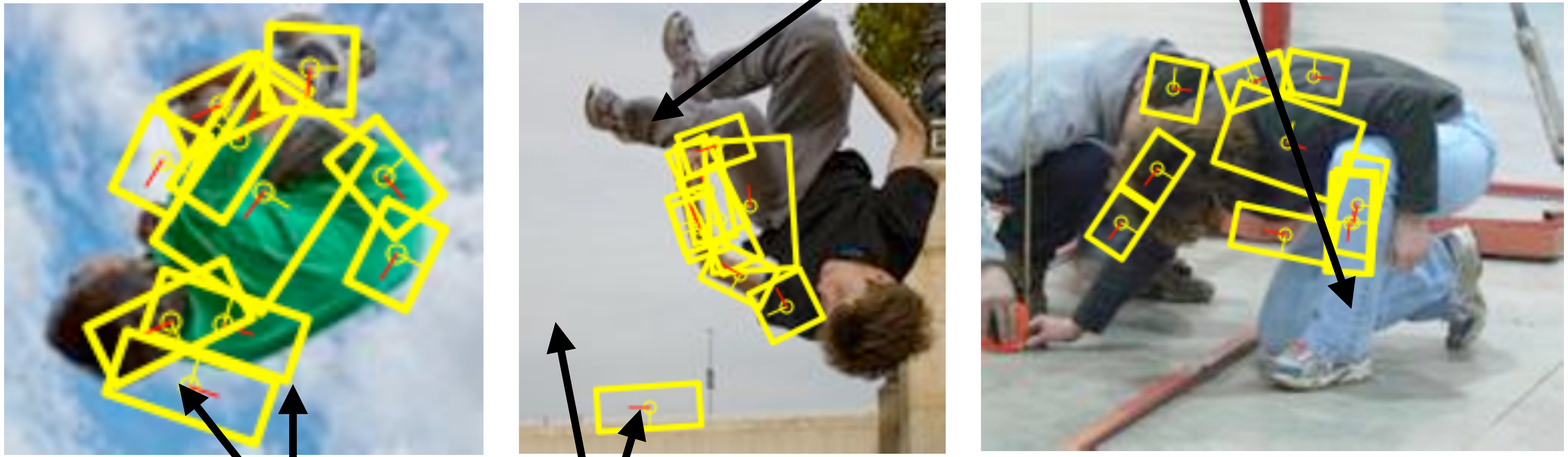
Rare poses



Strong foreshortening

Only detection!

Explain this then!

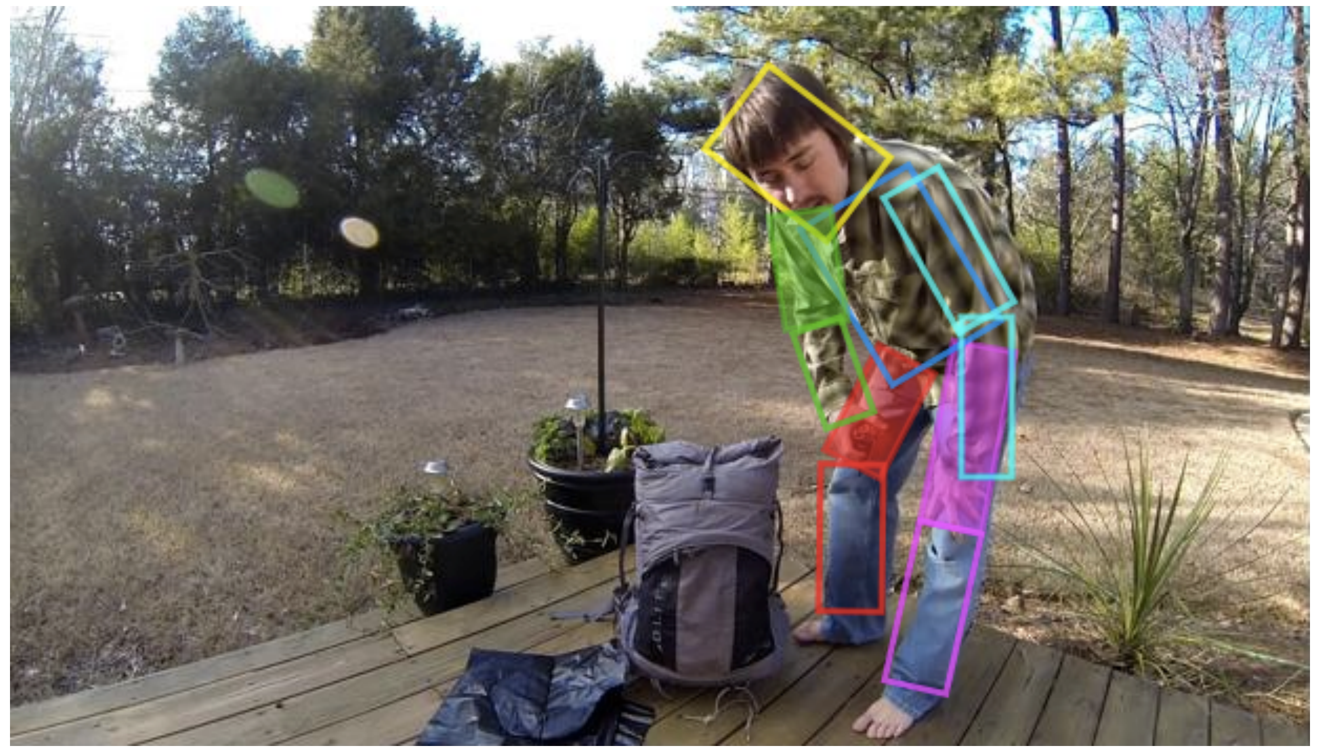


Same color!

Challenging Pose Dataset

- 400 activities
- 40000 examples
- multiple people
- video

joint positions and occlusions



part occlusions

3D torso and head orientation

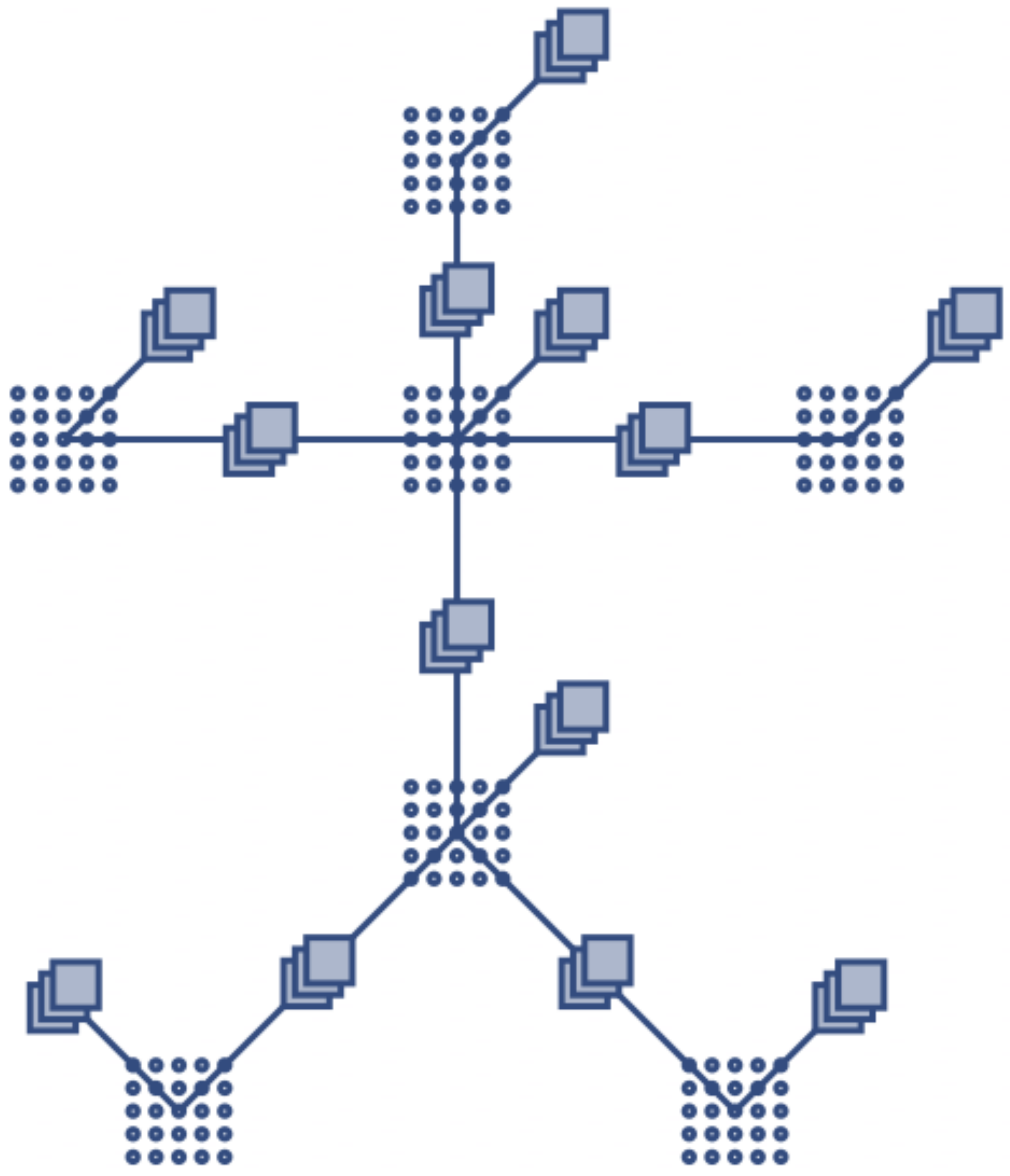
activity labels

M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, **Human Pose Estimation: A new Benchmark and State of the Art Analysis**, CVPR 2014

Fields of Parts — Parametrization

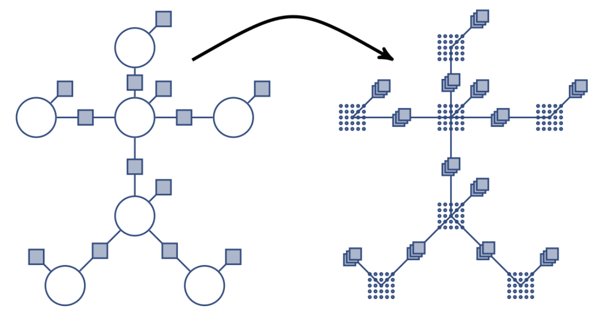
- for every body part… $p = 1, \ldots, P$
- …and every possible state $|\mathcal{Y}^p|$
- … a binary random variable

$$x_i^p \in \{0, 1\}, i = 1, \ldots, |\mathcal{Y}^p|$$



Kiefel & Gehler, **Human Pose Estimation with a Fields of Parts**, ECCV 2014

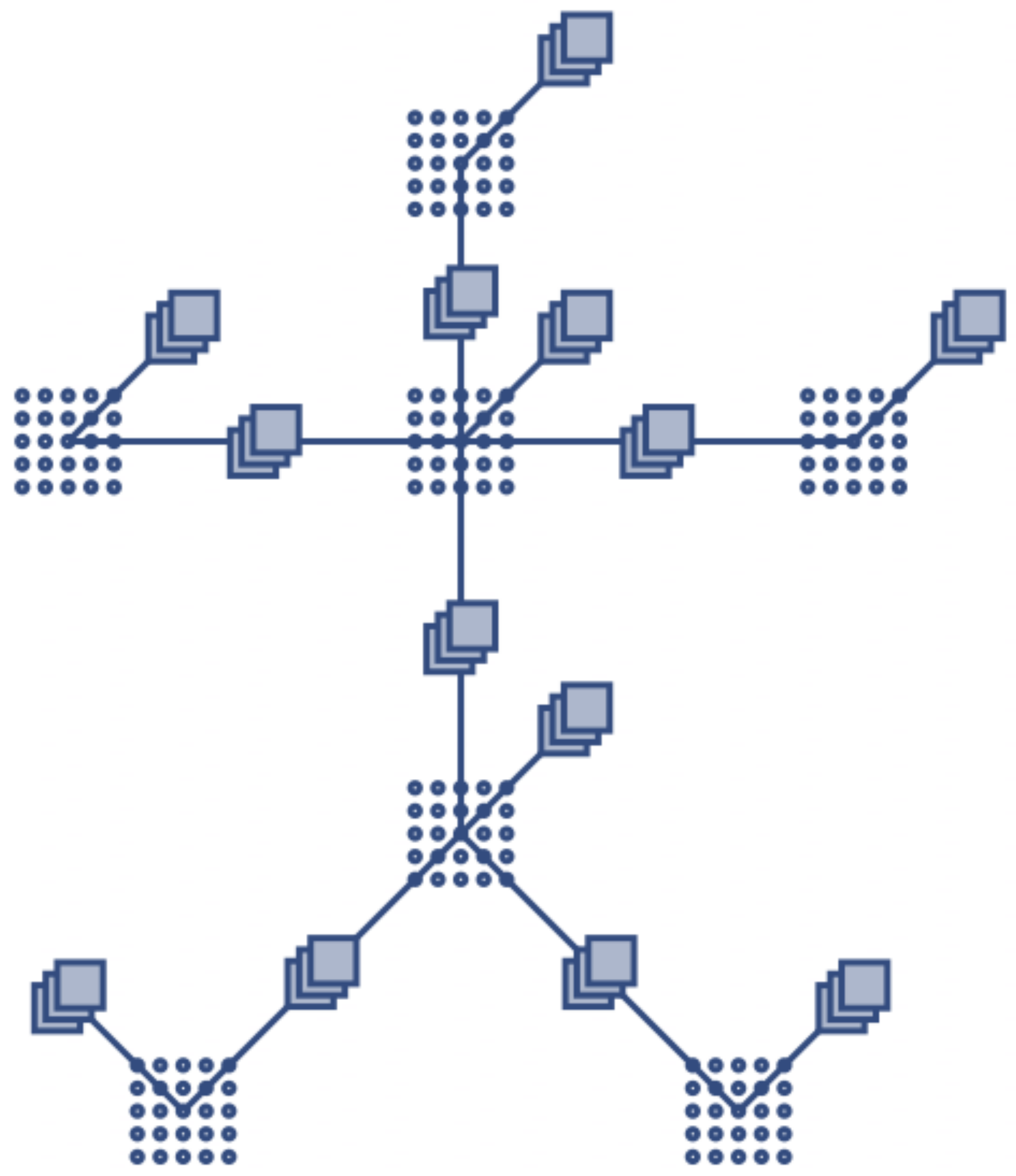
Fields of Parts — Energy



- Pairwise binary CRF (looooooopy)

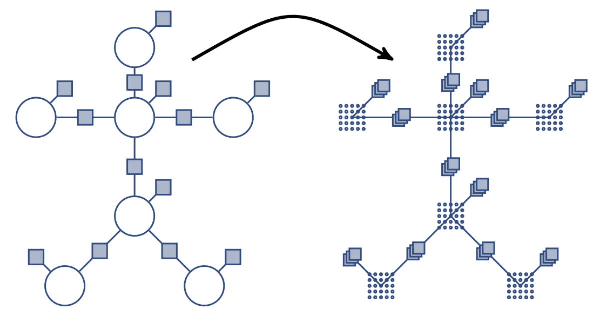
$$E(x|I,\theta) = \sum_{p=1}^{P} \sum_{i=1}^{|\mathcal{Y}^p|} \Psi_{\text{unary}}(x_i^p|I,\theta)$$

$$+ \sum_{p \sim p'} \sum_{i=1}^{|\mathcal{Y}^p|} \sum_{j=1}^{|\mathcal{Y}^{p'}|} \Psi_{\text{pairwise}}(x_i^p, x_j^{p'}|I,\theta)$$

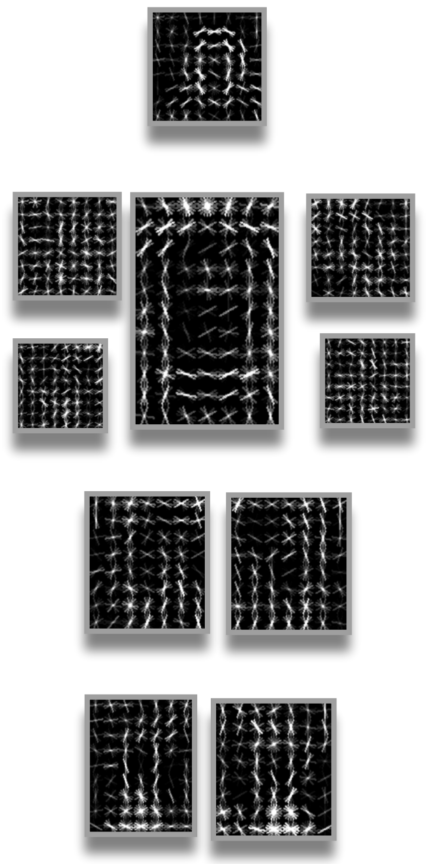


Kiefel & Gehler, **Human Pose Estimation with a Fields of Parts**, ECCV 2014

Fields of Parts — Factors

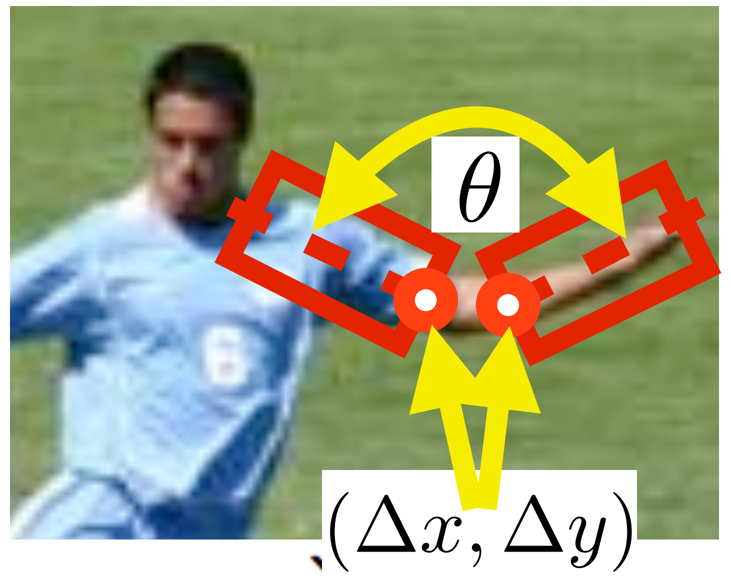


- Unary Factors — your usual HOG filter



$$\Psi_{\text{unary}}(x_i^p | I, \theta) = \langle \theta^p_{\text{unary}}, \psi_i(I) \rangle$$

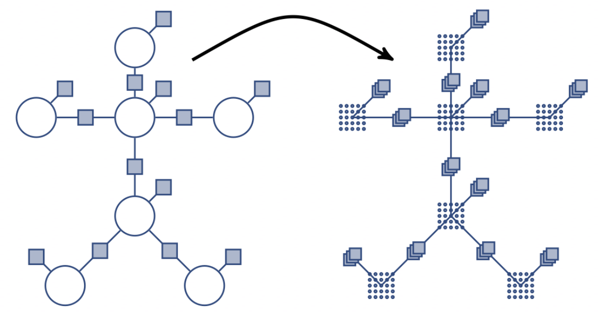
- Pairwise Factors — your usual displacement factor (and more)



$$\Psi_{\text{pairwise}}(x_i^p, x_j^{p'} | I, \theta) = \sum_m L_m(x_i^p, x_j^{p'}) \; k_m^{p,p'} \left(f(i, p; I, \theta), f(j, p'; I, \theta); \theta \right)$$

$$k_m^{p,p'}(f_i^p, f_j^{p'}; \theta) = \exp\left(-\frac{1}{2} (f_i^p - f_j^{p'} - \mu_m^{p,p'})^T (\Sigma_m^{p,p'})^{-1} (f_i^p - f_j^{p'} - \mu_m^{p,p'}) \right)$$

Comparison to PS

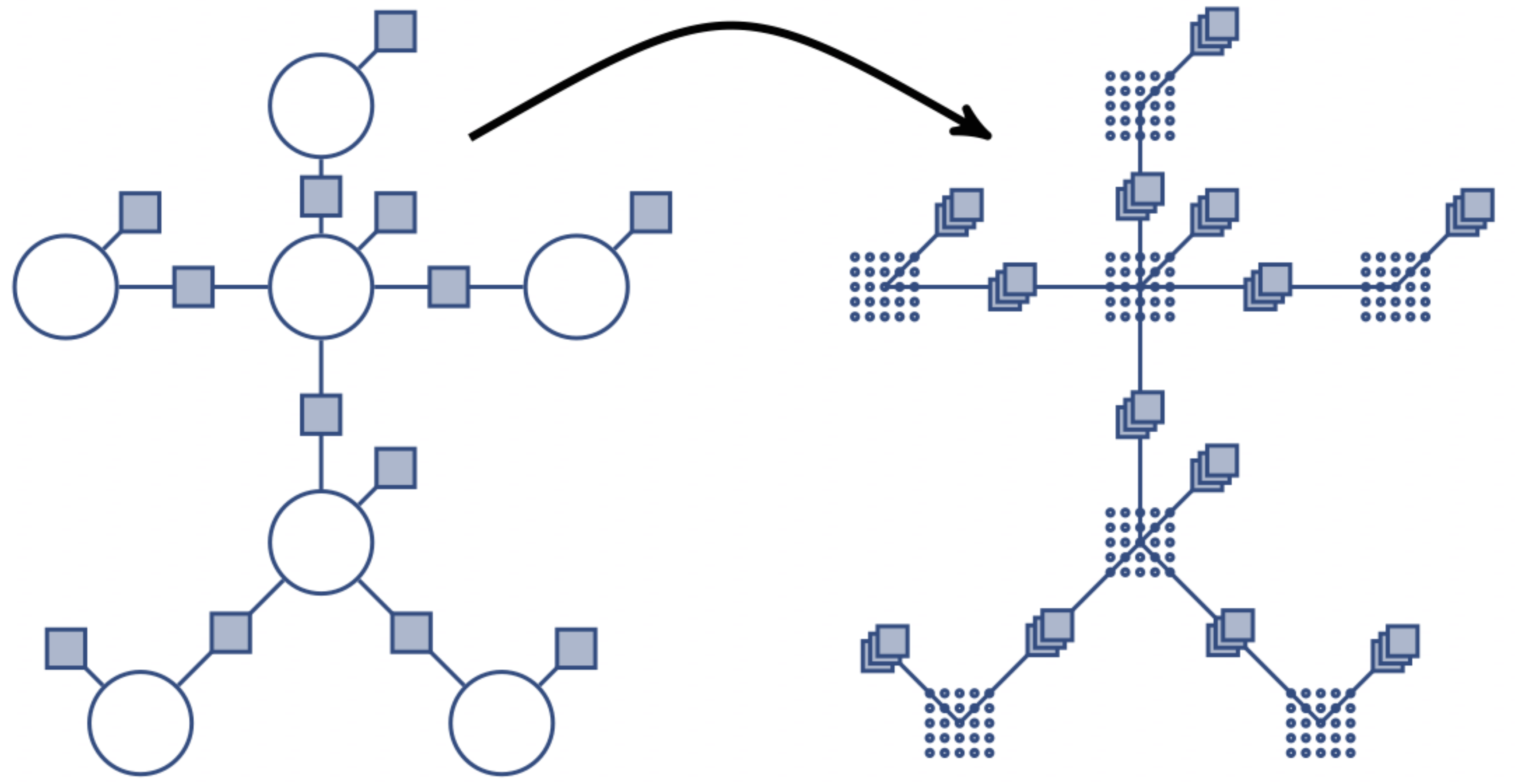


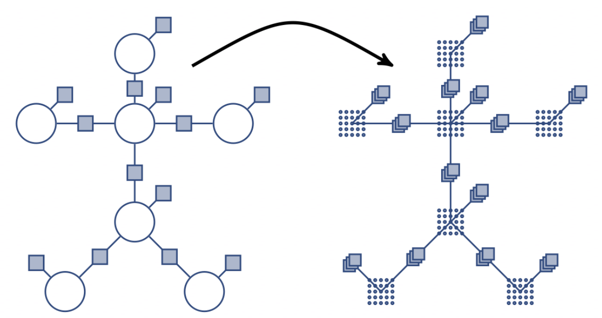
- Number of (body) parts $p = 1, \ldots, P$
- Pictorial Structures — few parts, huge state space

$$y^p \in \{1, \ldots, M\} \times \{1, \ldots, N\} = \mathcal{Y}^p$$

- Fields of Parts — many parts, small state space

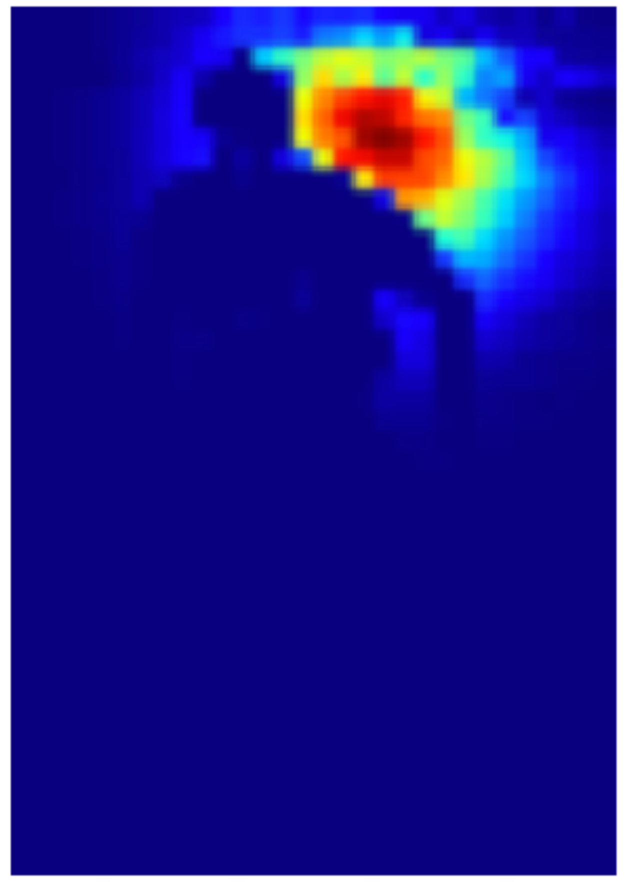
$$x_i^p \in \{0, 1\}, i = 1, \ldots, |\mathcal{Y}^p|$$





Gain: Bilateral

- Locally image conditioned pairwise factors (bilateral, segmentation)
- Not possible in distance transform for pictorial structures

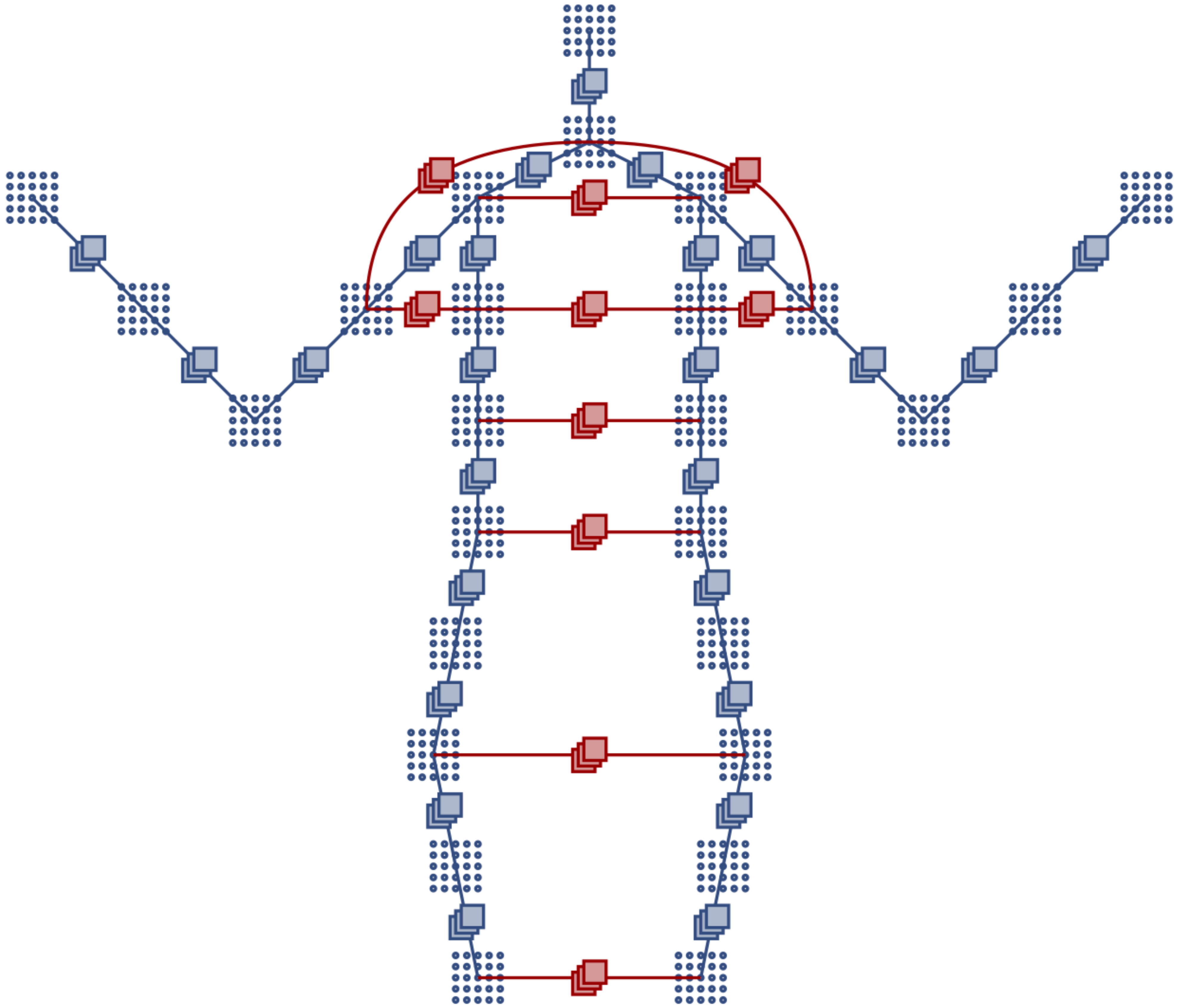


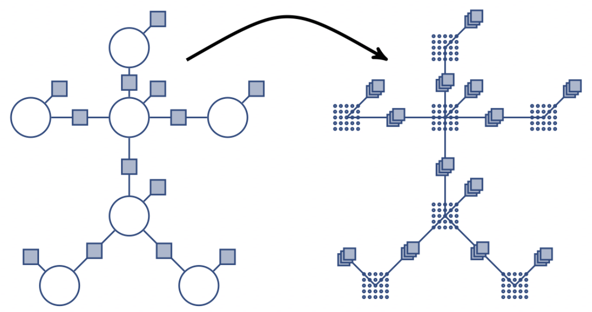
$$\Psi_{\text{pairwise}}(x_i^p, x_j^{p'} | I, \theta) = \sum_m L_m(x_i^p, x_j^{p'})\ k_m^{p,p'}\left(f(i,p;I,\theta), f(j,p';I,\theta);\theta\right)$$

$$k_m^{p,p'}(f_i^p, f_j^{p'};\theta) = \exp\left(-\frac{1}{2}(f_i^p - f_j^{p'} - \mu_m^{p,p'})^T (\Sigma_m^{p,p'})^{-1} (f_i^p - f_j^{p'} - \mu_m^{p,p'})\right)$$

More connections

- Block-dense connections already
- New connections scale linearly





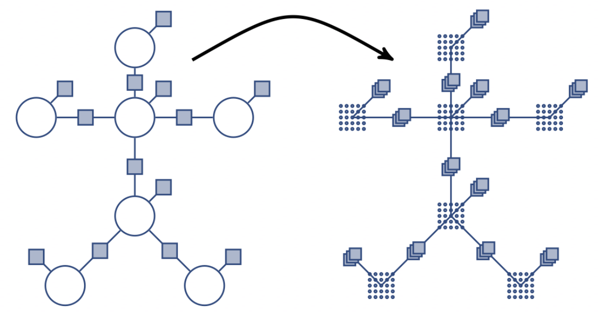
Inference

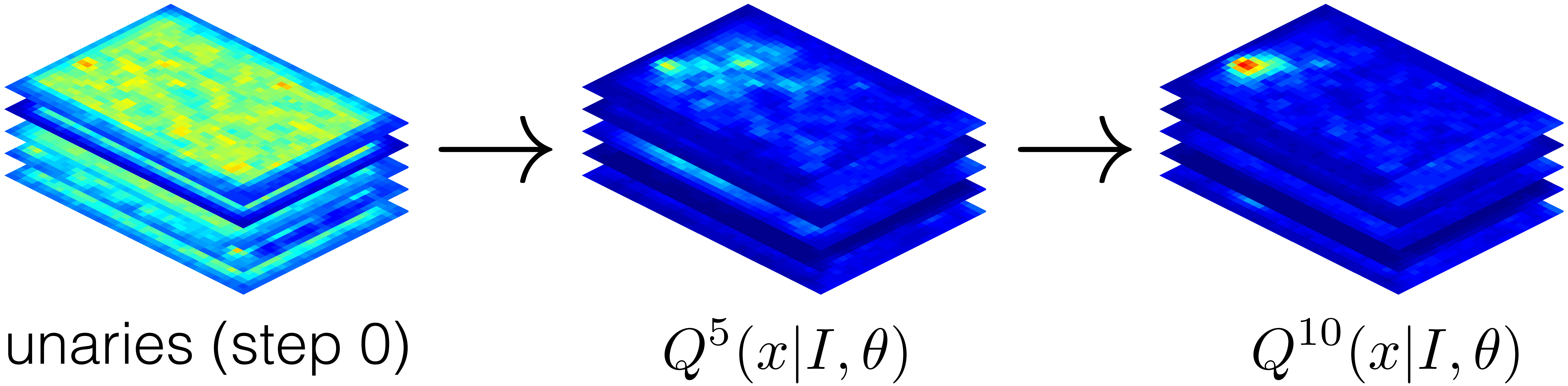
- Intractable Inference
- Mean Field Approximation $Q(x|I,\theta) = \prod_i Q(x_i^p|I,\theta)$
- Update Equation — Bilateral Filtering Operation (linear complexity)

$$Q(x_i^p|I,\theta) \propto \exp\left(-\sum_{p=1}\sum_{i=1}^{|\mathcal{Y}^p|}\Psi_{\text{unary}}(x_i^p|I,\theta) - \sum_{p\sim p'}\sum_{l'\in\{0,1\}}\sum_m L_m(x_i^p,l')\right.$$
$$\left.\sum_{j=1}^{|\mathcal{Y}^{p'}|} k_m^{p,p'}(f(i,p;I,\theta),f(j,p';I,\theta);\theta)\ Q(x_j^{p'}=l'|I,\theta)\right).$$

Krähenbühl & Koltun, **Efficient inference in fully connected CRFs with Gaussian edge potentials**, NIPS 2011

Fields of Parts — Inference





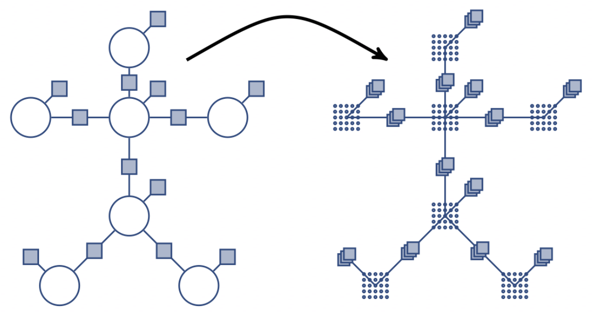
- Mean Field updates (here 10)

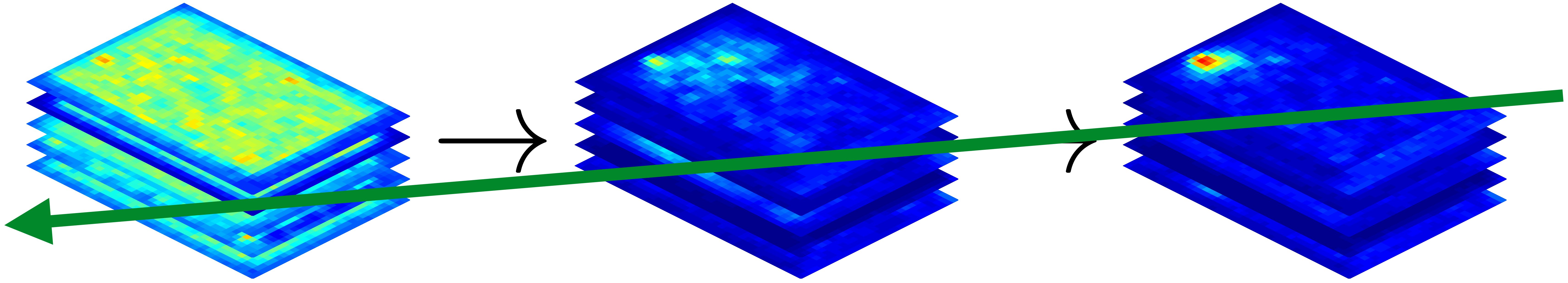
$$Q^0(x|I,\theta) \rightarrow Q^1(x|I,\theta) \rightarrow \cdots \rightarrow Q^{10}(x|I,\theta)$$

- Predict the maximum marginal state

$$\hat{i}^p = \operatorname*{argmax}_{i \in \mathcal{Y}^p} Q^{10}(x_i^p = 1|I)$$

Fields of Parts — Objective





unaries (step 0) $\quad Q^5(x|I,\theta) \quad Q^{10}(x|I,\theta)$

- Objective: Max-Margin Max-Marginal (structured SVM)

$$\text{minimize}_{\theta,\xi^p \geq 0} \sum_p \ell(\xi^p) + C(\theta)$$

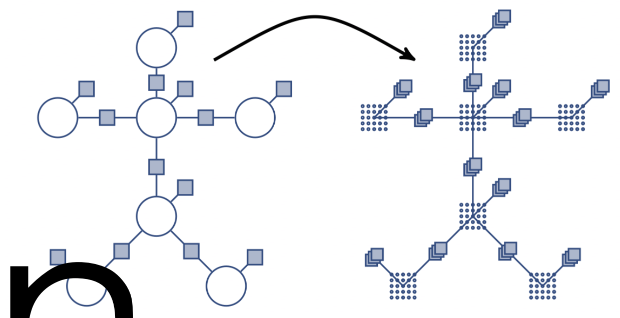
$$\text{sb.t. } s^p_{i_*} - s^p_i \geq \Delta^p(i_*, i) - \xi^p \;\forall p, \forall i \in \mathcal{Y}^p$$

- Backpropagation Mean Field — autodiff through bilateral filtering

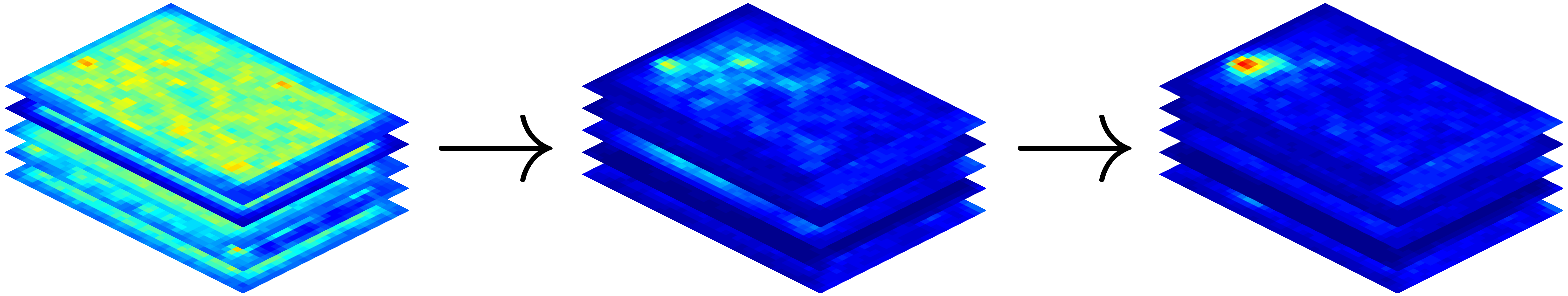
$$Q^0(x|I,\theta) \to Q^1(x|I,\theta) \to \cdots \to Q^{10}(x|I,\theta)$$

J. Domke, **Learning Graphical Model Parameters with Approximate Marginal Inference**, PAMI 2013

P. Krähenbühl & V. Koltun, **Parameter Learning and Convergent Inference for Dense Random Fields**, ICML 2013



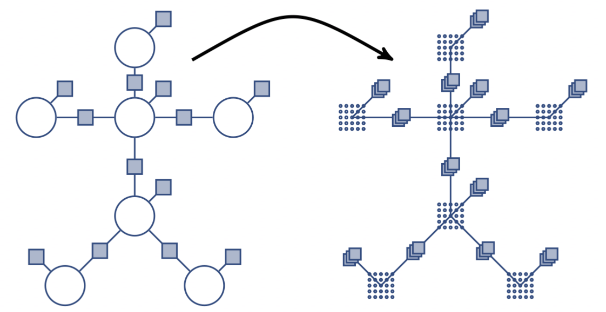
Neural Network Interpretation



unaries (step 0) $Q^5(x|I,\theta)$ $Q^{10}(x|I,\theta)$

- Non-linear convolutional Filter defined by dense graphical model and mean field inference

$$Q^{i+1}(x|I,\theta) = \mathcal{F}(Q^i(x|I,\theta))$$



Results — APK

Model	Setting	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	avg
Yang&Ramanan [12]		80.0	75.2	48.2	28.9	70.4	60.5	53.2	59.5
Yang&Ramanan [12]	(single det.)	79.5	74.9	47.6	28.4	69.9	59.0	51.6	58.7
Fields of Parts		83.1	76.5	55.2	29.0	74.8	70.3	63.7	64.7

Table 1. Comparison of pose estimation results on the LSP dataset. Shown are the APK results (observer-centric annotations [13]).

- On equal ground: same features, same “pairwise” terms
- Pairwise conditionals improve

Disclaimer: Not state-of-the-art

Model	Setting	Torso	Upper leg	Lower leg	Upper arm	Fore-arm	Head	Total
Fields of Parts		82.2	71.8	66.5	52.0	27.7	76.8	59.5
Fields of Parts	Bilateral	83.4	72.8	67.0	52.2	28.0	77.0	60.0
Fields of Parts	Segmentation	84.4	74.4	67.1	53.3	27.4	78.4	60.7
Fields of Parts	Loopy	81.8	73.7	66.9	52.0	26.8	77.3	59.8
Yang&Ramanan	[12]	81.0	67.4	63.9	51.0	31.8	77.3	58.6
Andriluka et al.,	[28]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
Pishchulin et al.,	[9]	87.5	75.7	68.0	54.2	33.9	78.1	62.9
Pishchulin et al.,	[10]	**88.7**	**78.8**	**73.4**	**61.5**	**44.9**	**85.6**	**69.2**
Eichner&Ferrari	[13]	86.2	74.3	69.3	56.5	37.4	80.1	64.3

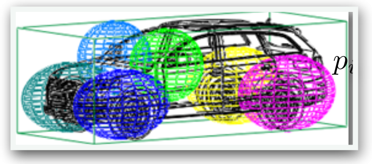
Table 2. Pose estimation results using the PCP criterion on the LSP dataset. We compare our method against the current top performing methods in the literature.

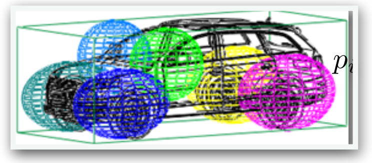
- PCP error measure

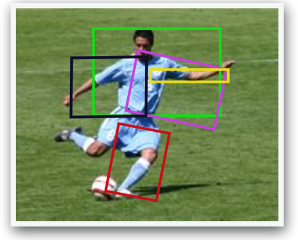
Conclusion & Future Work

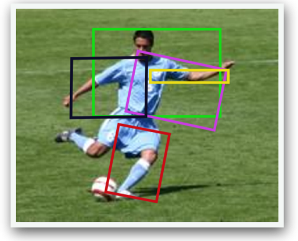
- Parts are important for better models/understanding, not necessarily for performance
  - Richer image interpretation: joint pose estimation & image segmentation
  - More output: 3D pose, clothing, body measurements, etc
- Robustness and speed
  - Will see more models that put tractable inference first

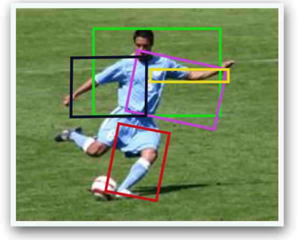
Reference List

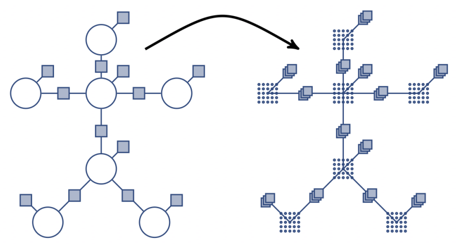
 Teaching Geometry to Deformable Part Models, CVPR12

 3D2DPM — 3D Deformable Part Models, ECCV12

 Poselet Conditioned Pictorial Structures, CVPR13

 Strong Appearance and Expressive Spatial Models for Human Pose Estimation, ICCV13

 Human Pose Estimation: A new Benchmark and State of the Art Analysis, CVPR14

 Human Pose Estimation with a Fields of Parts, ECCV14

Martin Kiefel

Leonid Pishchulin

Micha Andriluka

Bernt Schiele

Thank You! Feedback Welcome!