Pose Machines: Articulated Pose Estimation via Inference Machines

Varun Ramakrishna, Daniel Munoz*, Martial Hebert,
J. Andrew Bagnell, Yaser Sheikh

Carnegie Mellon University

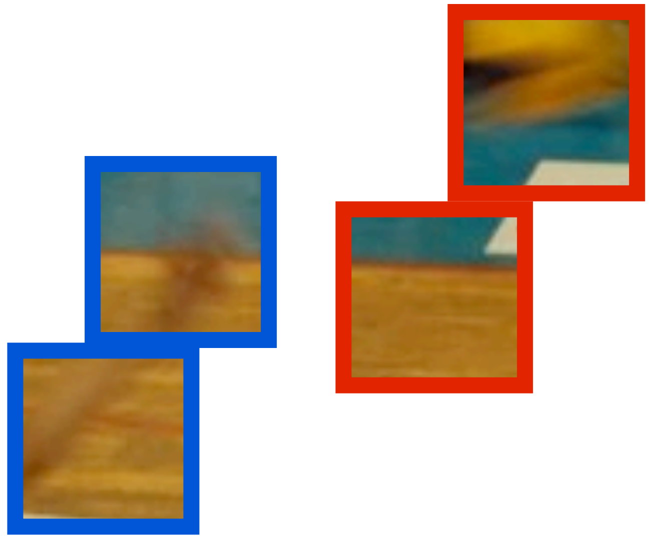
THE ROBOTICS INSTITUTE

*now at Google

Goal: Articulated Pose Estimation

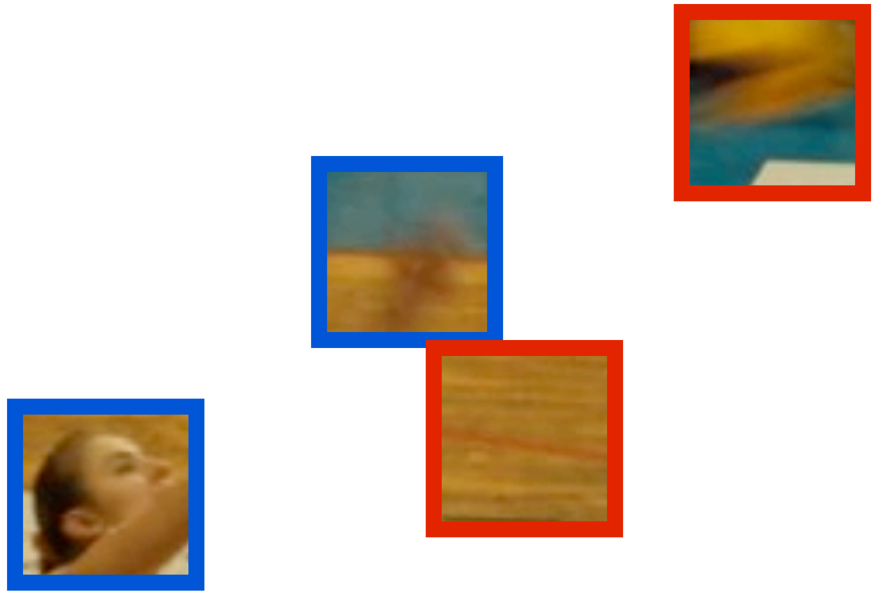
Which patch corresponds to a body part?

Which patch corresponds to a body part?



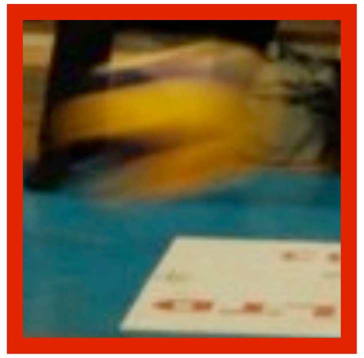
Which patch corresponds to a body part?

Which patch corresponds to a body part?



Which patch corresponds to a body part?

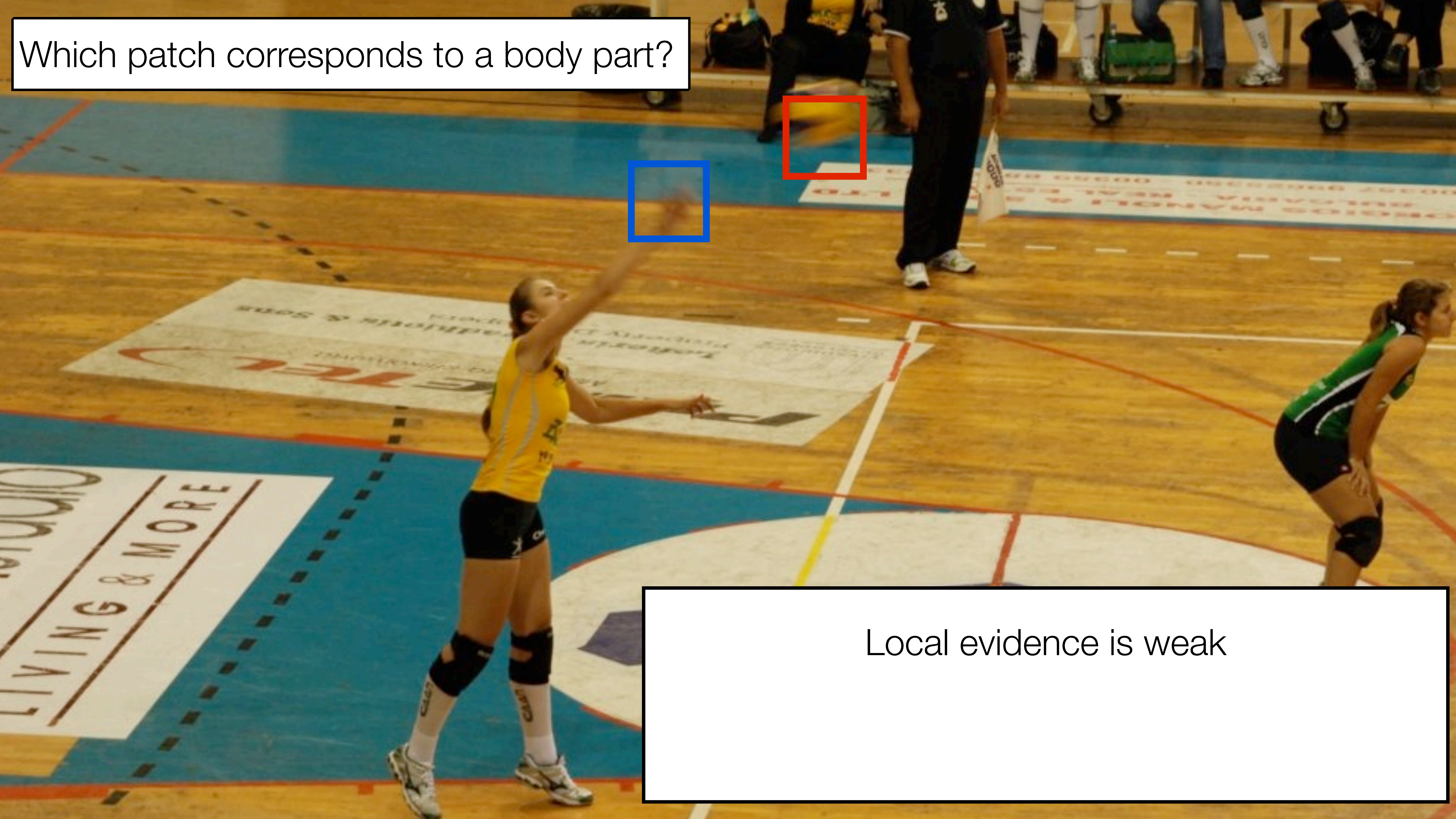
Which patch corresponds to a body part?



Which patch corresponds to a body part?

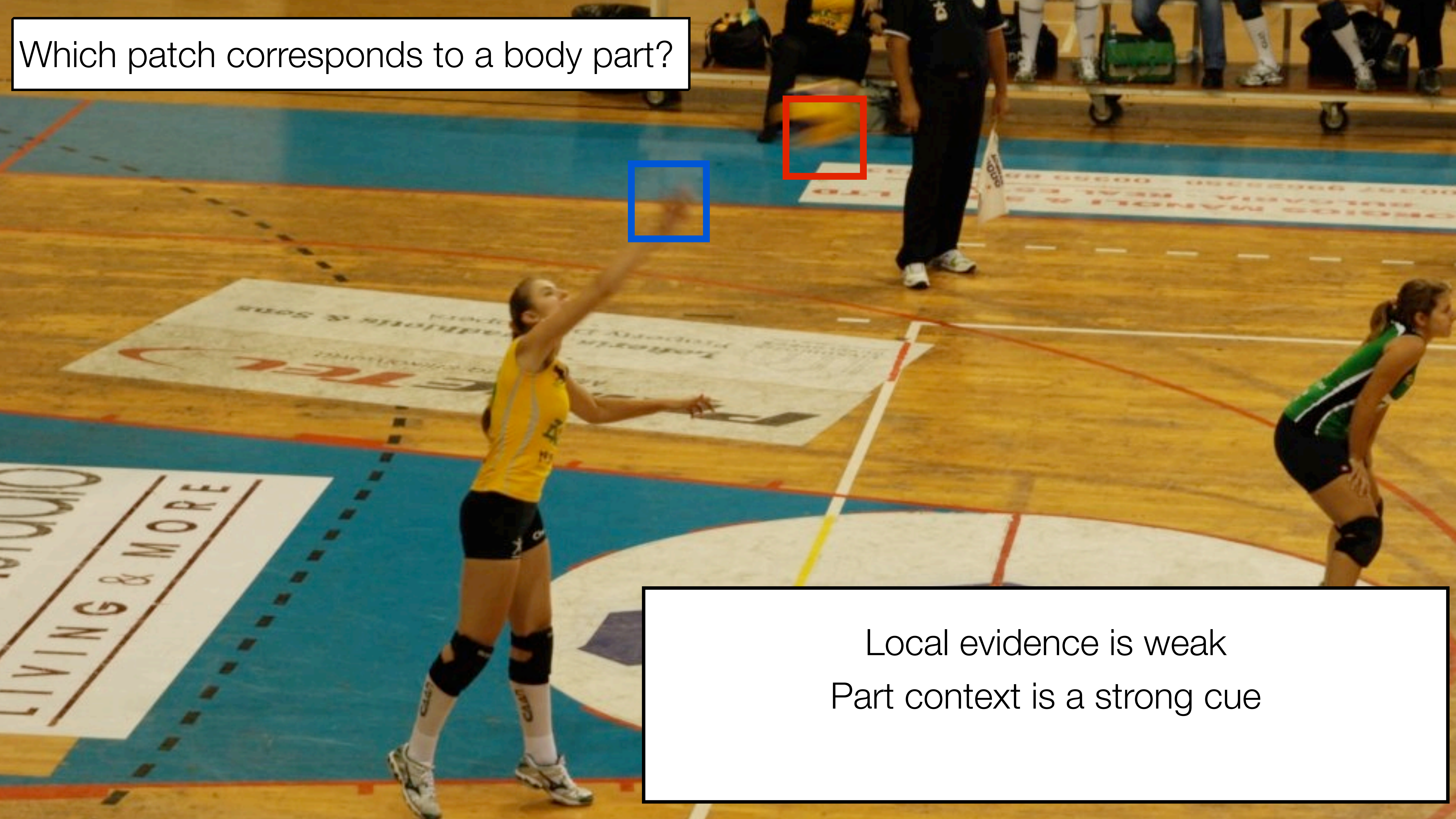
Which patch corresponds to a body part?

Which patch corresponds to a body part?



Local evidence is weak

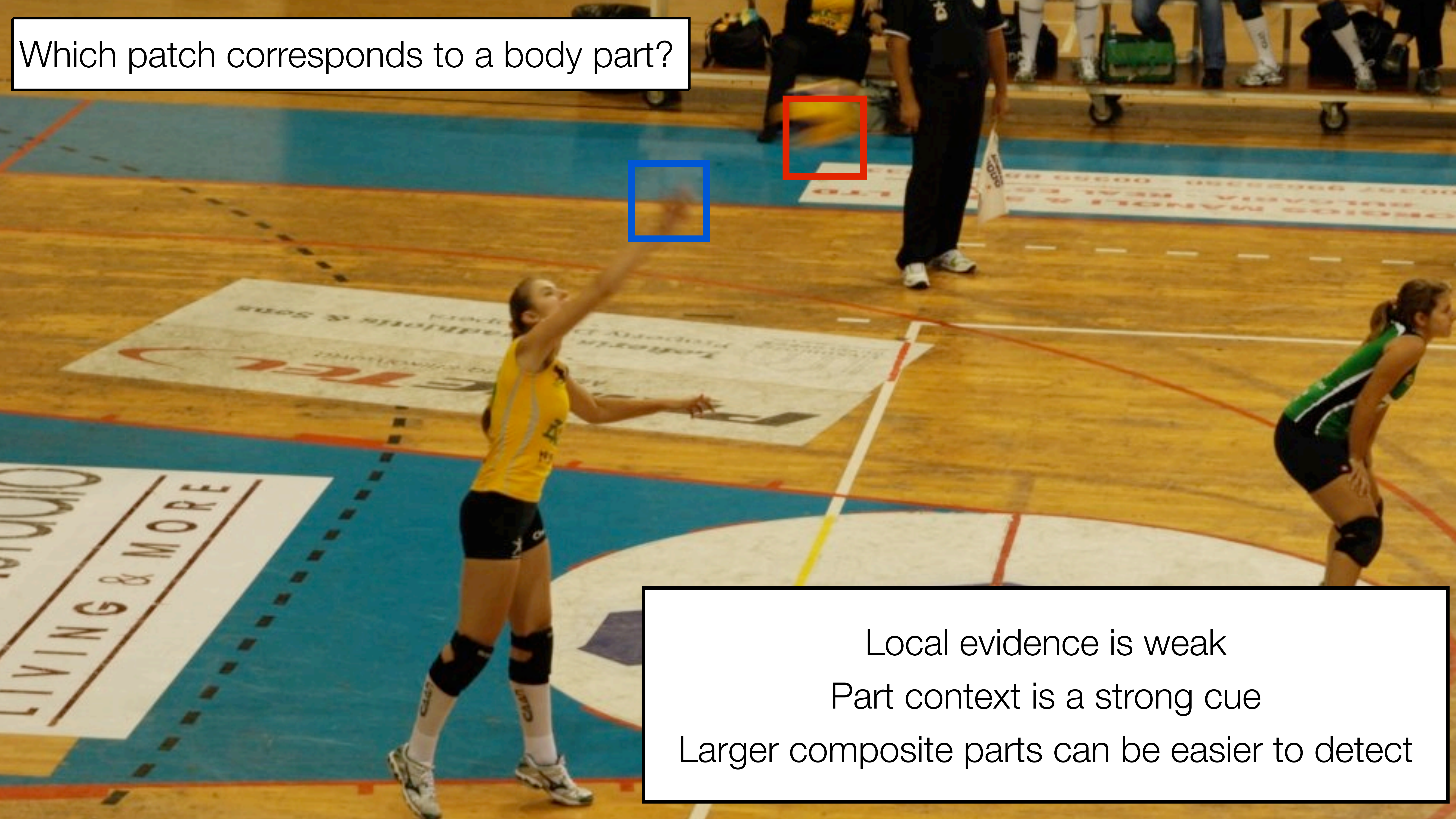
Which patch corresponds to a body part?



Local evidence is weak

Part context is a strong cue

Which patch corresponds to a body part?

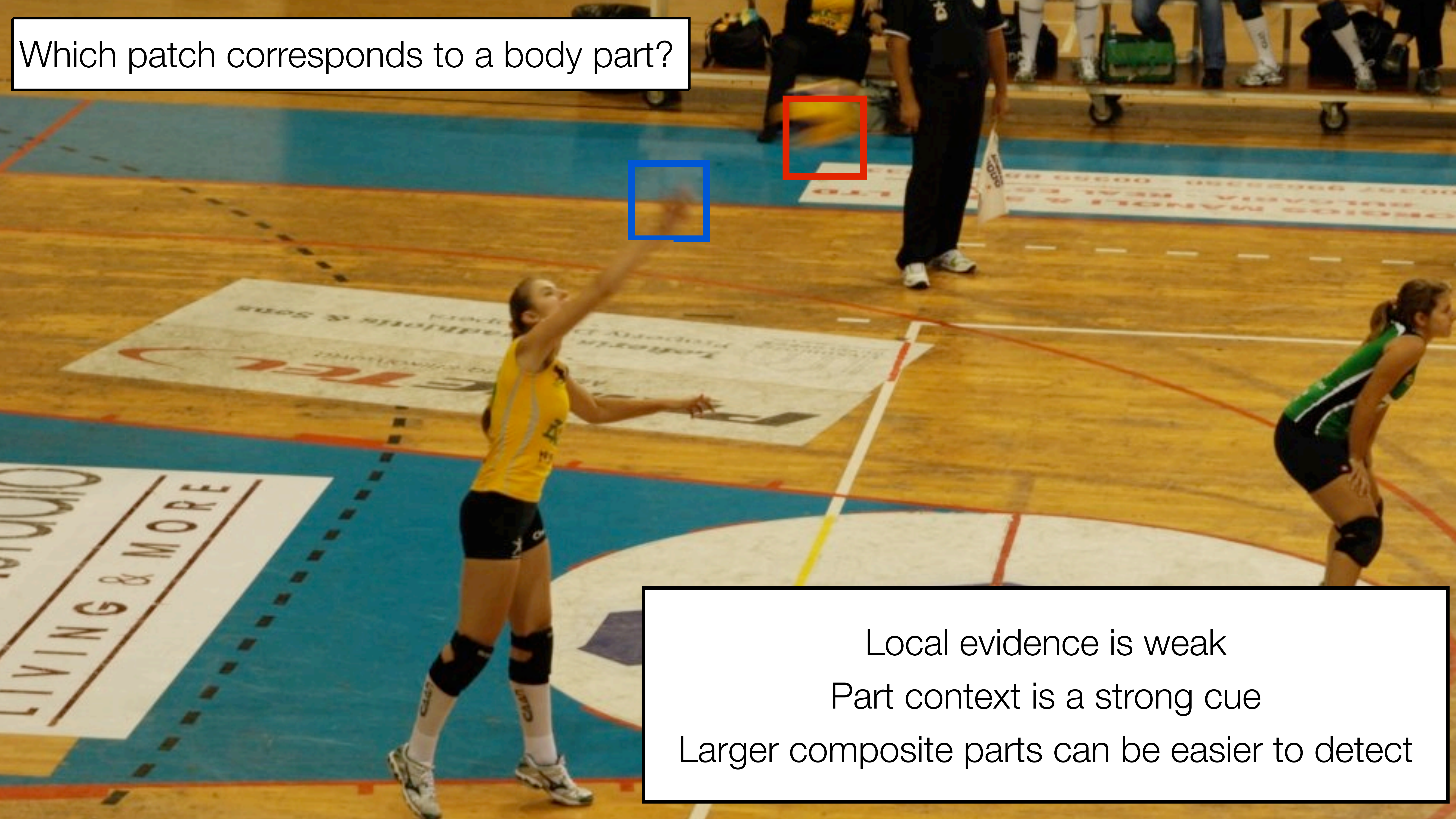


Local evidence is weak

Part context is a strong cue

Larger composite parts can be easier to detect

Which patch corresponds to a body part?



Local evidence is weak

Part context is a strong cue

Larger composite parts can be easier to detect

Part Detection using Local Image Evidence

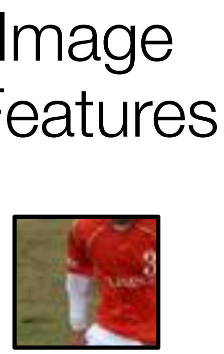
Multi-class classification of each patch into one of *P* part-types + background

Image Location z

Input Image

Part Detection using Local Image Evidence

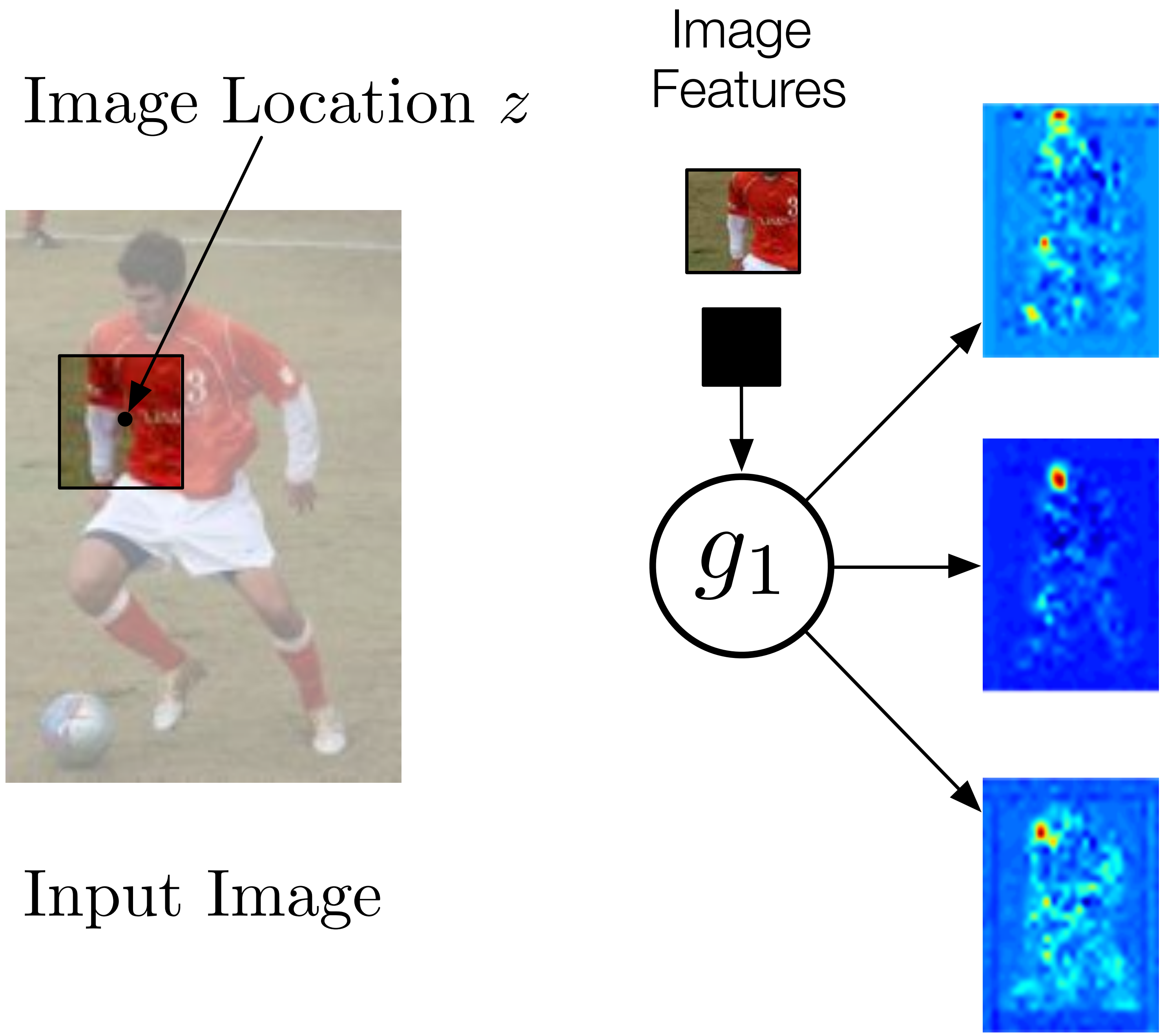
Multi-class classification of each patch into one of *P* part-types + background



Input Image

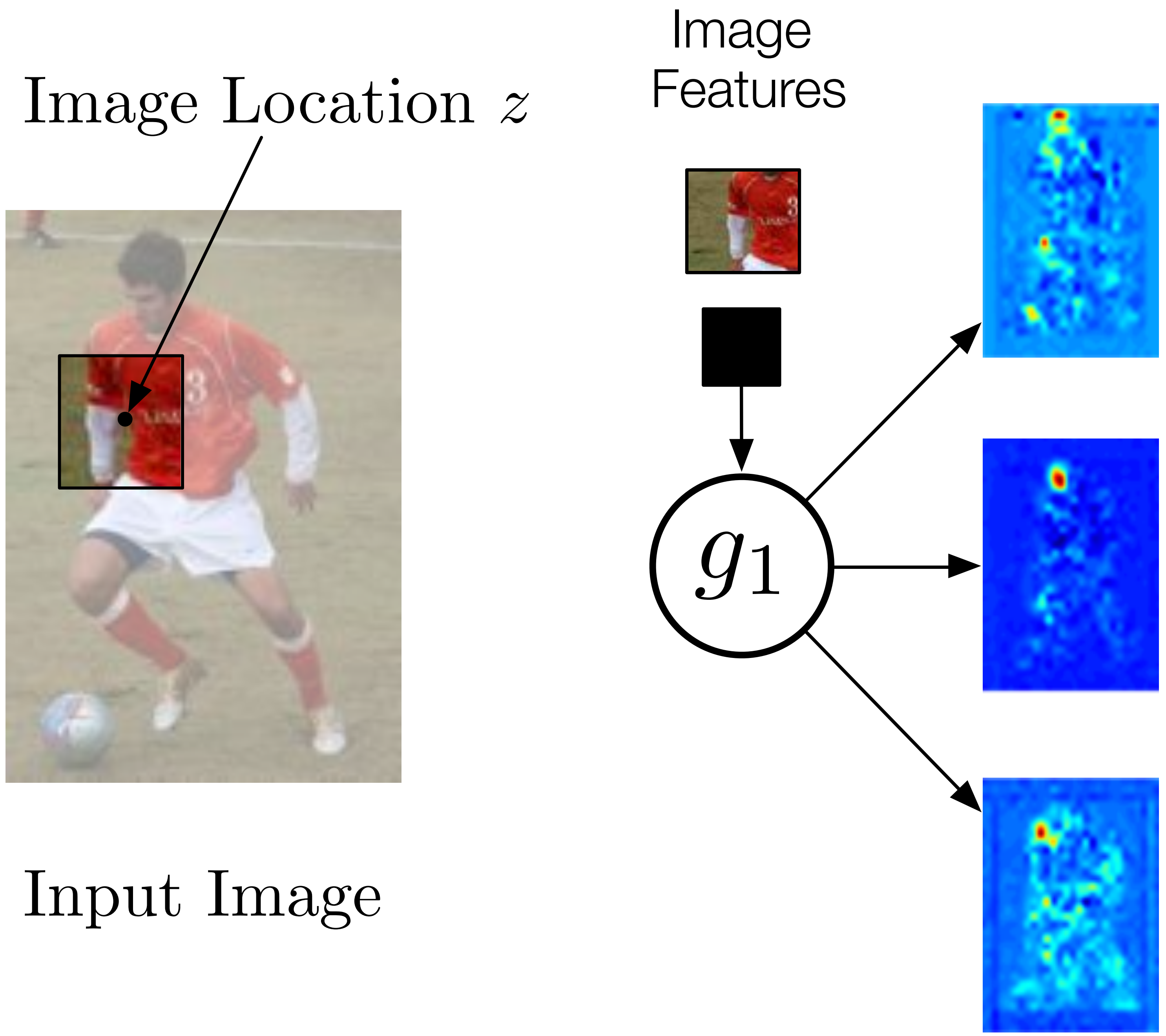
Part Detection using Local Image Evidence

Multi-class classification of each patch into one of *P* part-types + background



Part Detection using Local Image Evidence

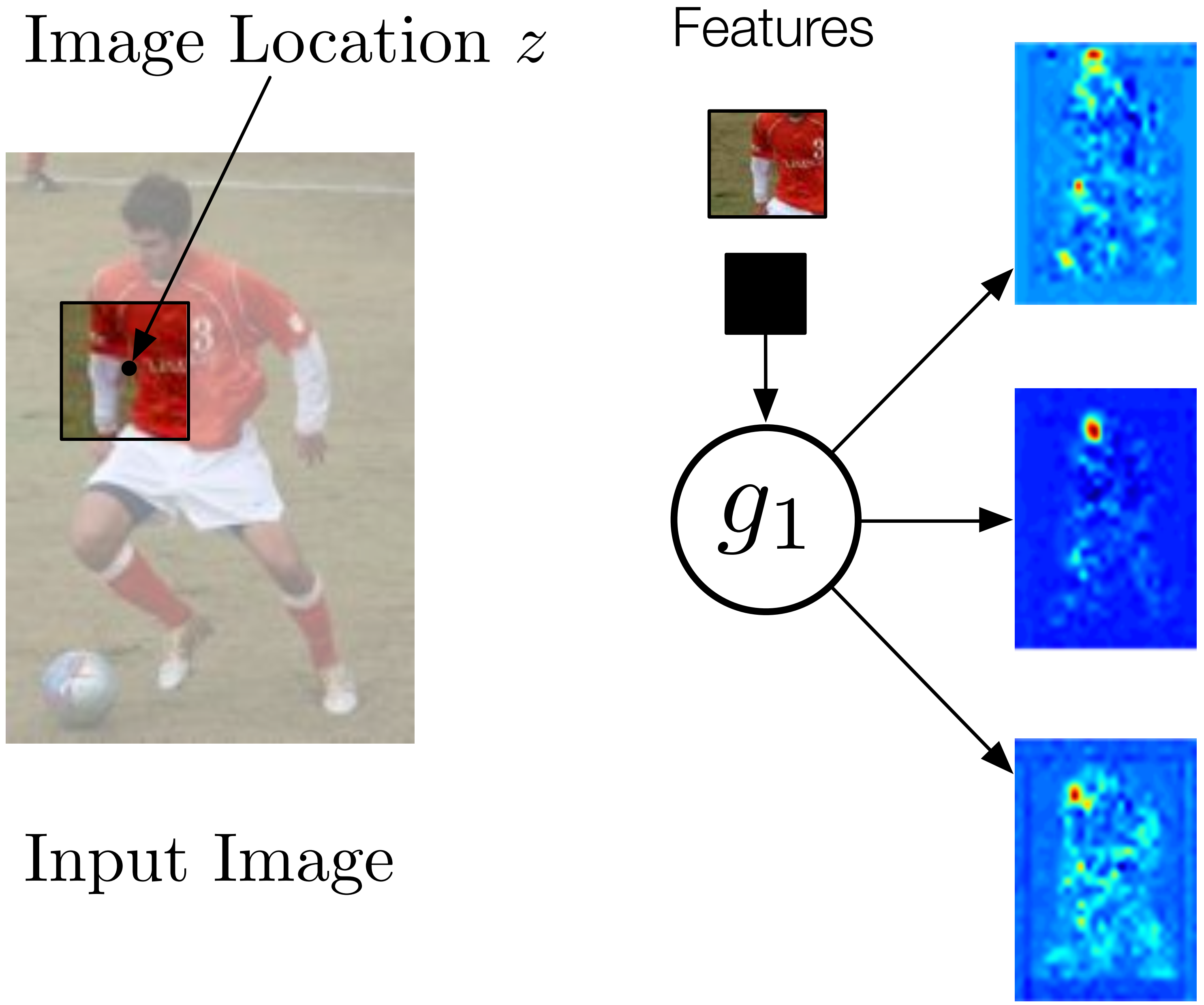
Multi-class classification of each patch into one of P part-types + background



Parts have highly multi-modal appearance variation

Part Detection using Local Image Evidence

Multi-class classification of each patch into one of P part-types + background

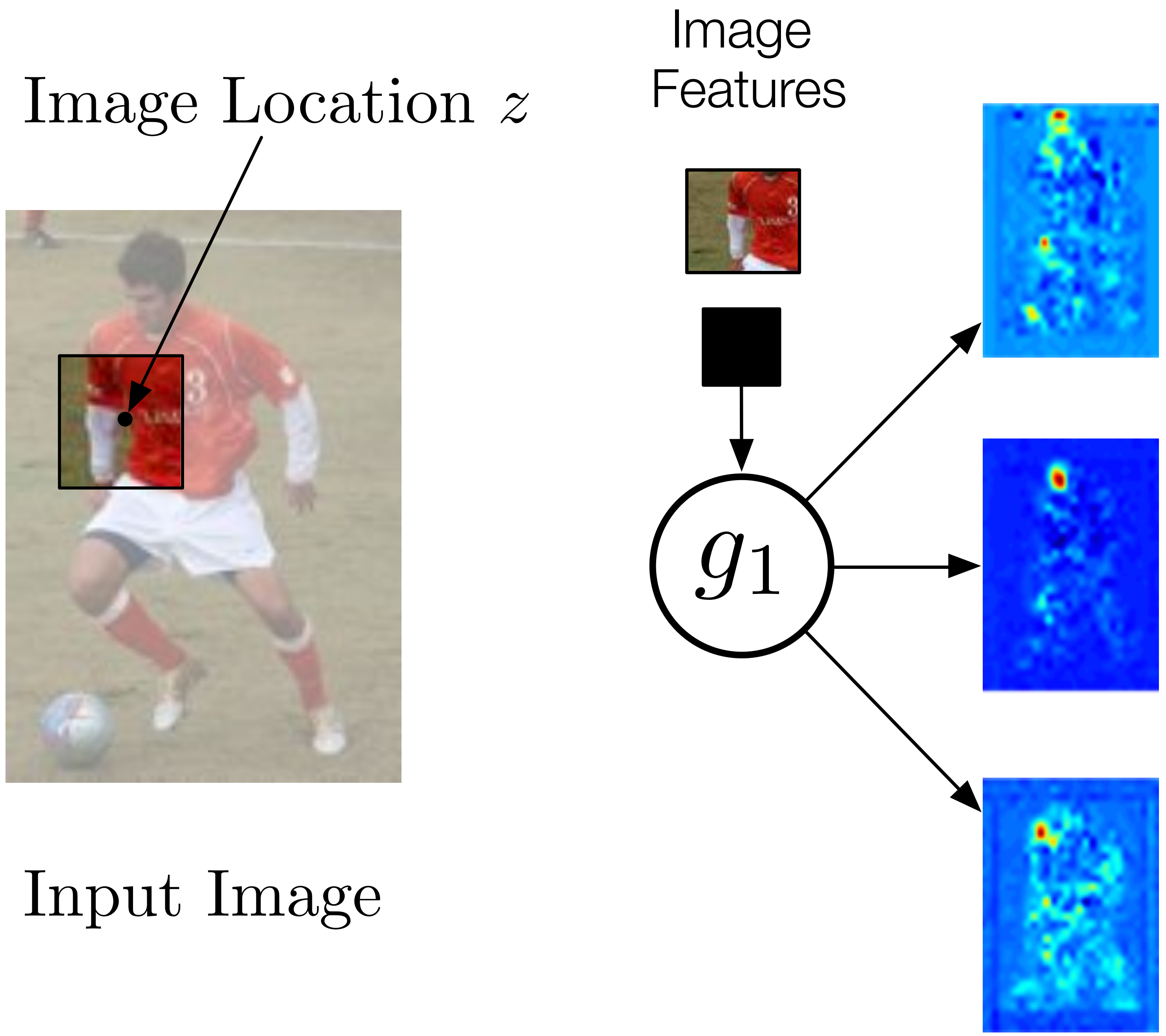


Parts have highly multi-modal appearance variation

Use a high-capacity supervised predictor capable of handling multi-modal data

Part Detection using Local Image Evidence

Multi-class classification of each patch into one of *P* part-types + background



Parts have highly multi-modal appearance variation

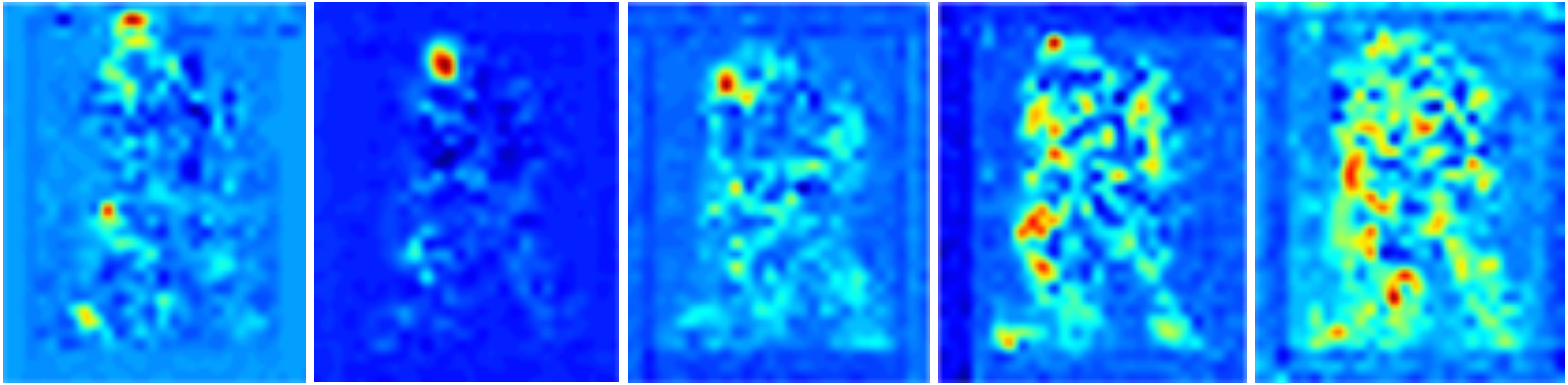
Use a high-capacity supervised predictor capable of handling multi-modal data

Boosted Random Forests
[Breiman, 2001] [Friedman, 2001]
[Caruana et al., 2009]

Local Image Evidence is Weak

Multi-class classification of each patch into one of *P* part-types + background

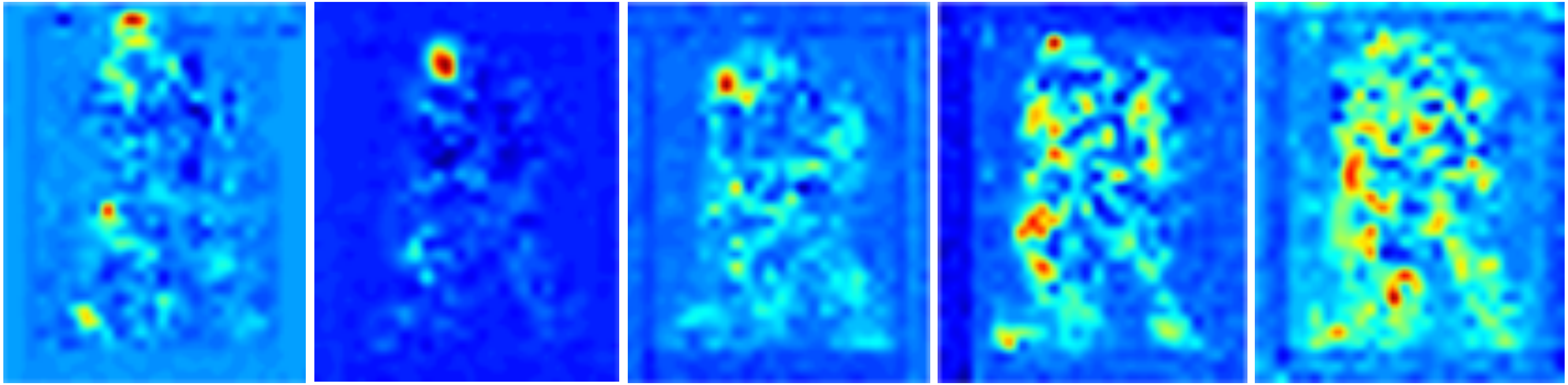
Head Neck L-Shoulder L-Elbow L-Wrist



Local Image Evidence is Weak

Multi-class classification of each patch into one of *P* part-types + background

Head Neck L-Shoulder L-Elbow L-Wrist

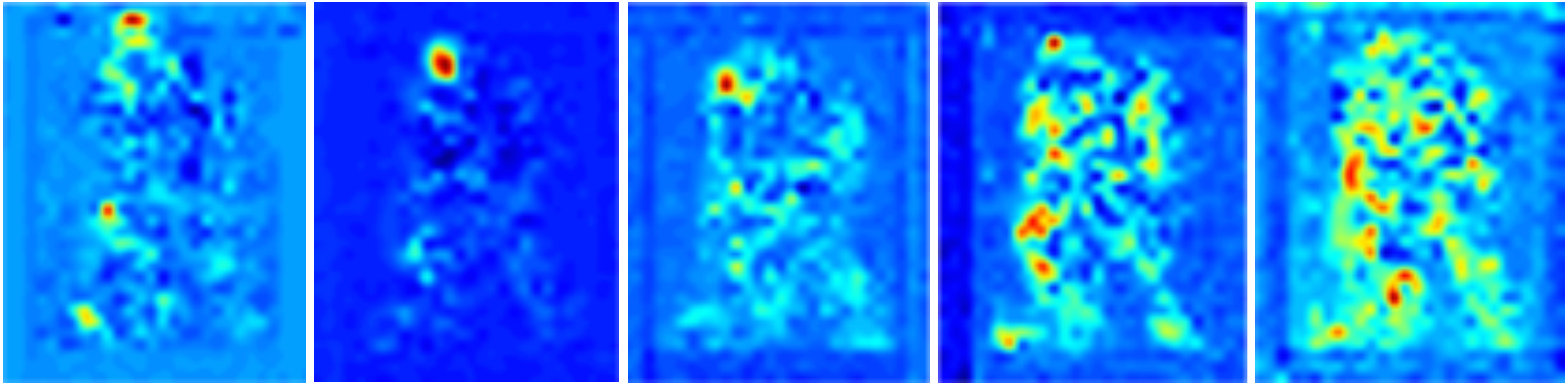


Local image evidence is weak

Local Image Evidence is Weak

Multi-class classification of each patch into one of *P* part-types + background

Head | Neck | L-Shoulder | L-Elbow | L-Wrist

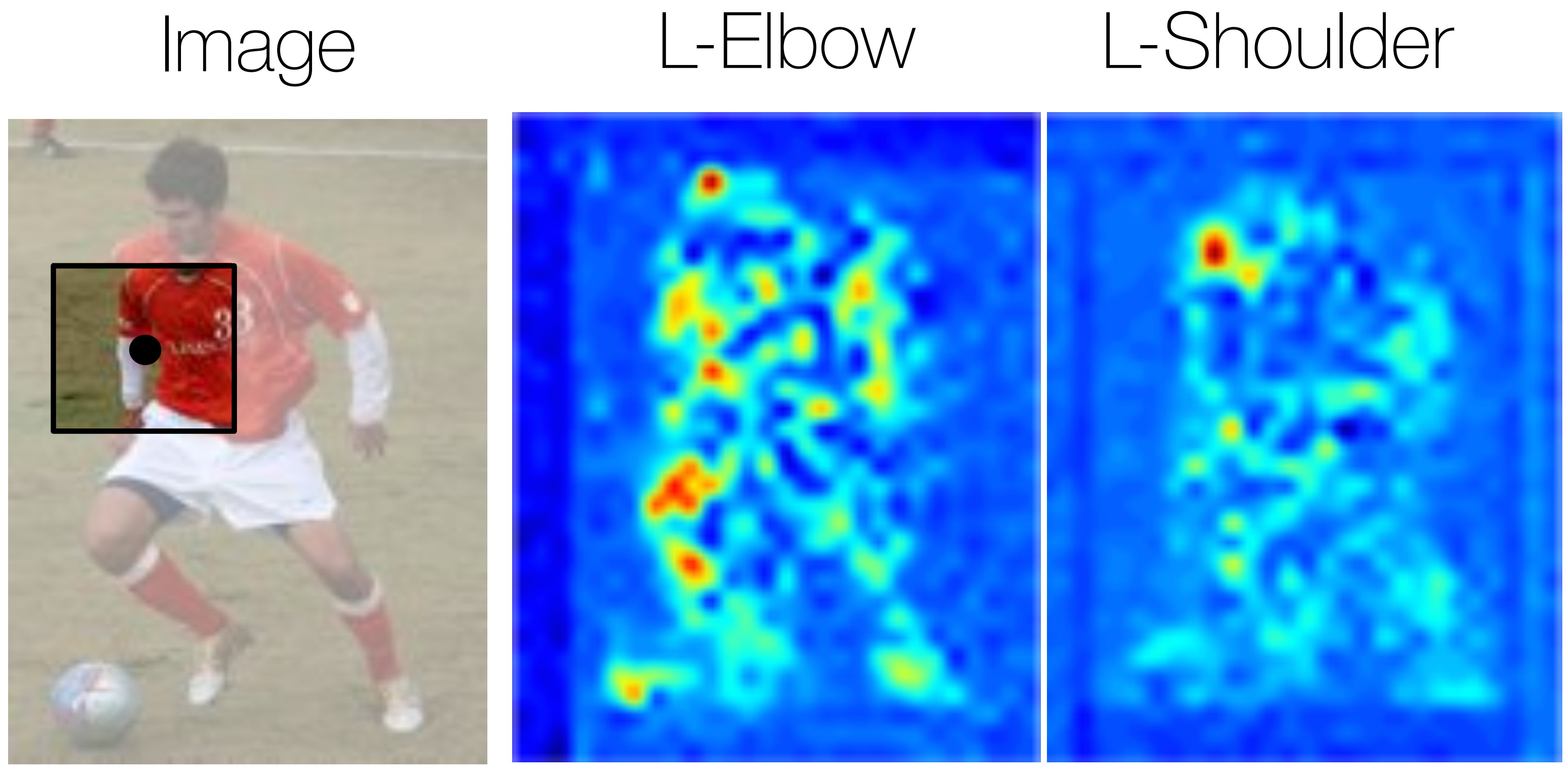


Local image evidence is weak

Certain parts are easier to detect than others

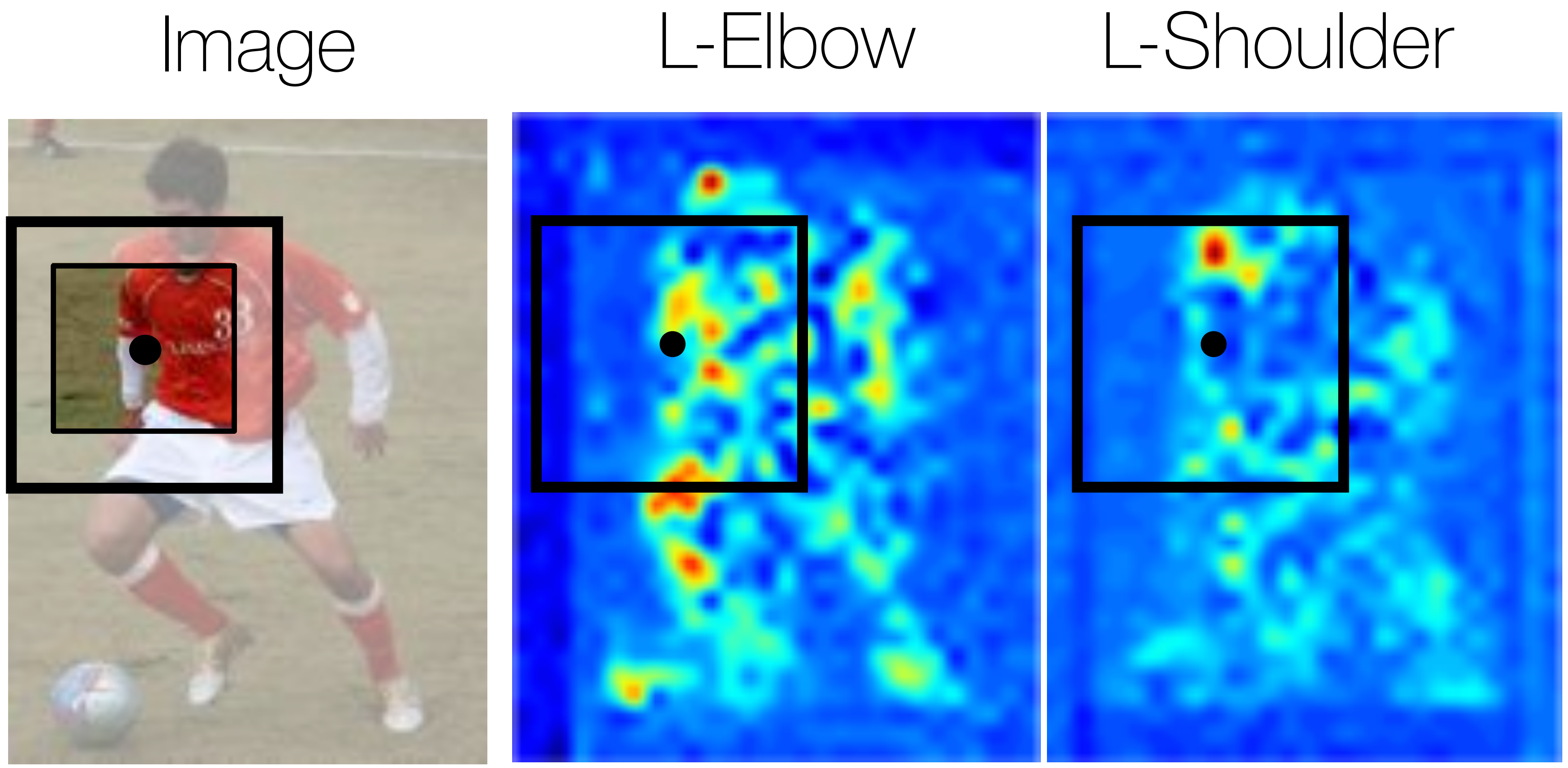
Part Context is a Strong Cue

Part detection confidences provide spatial context cues



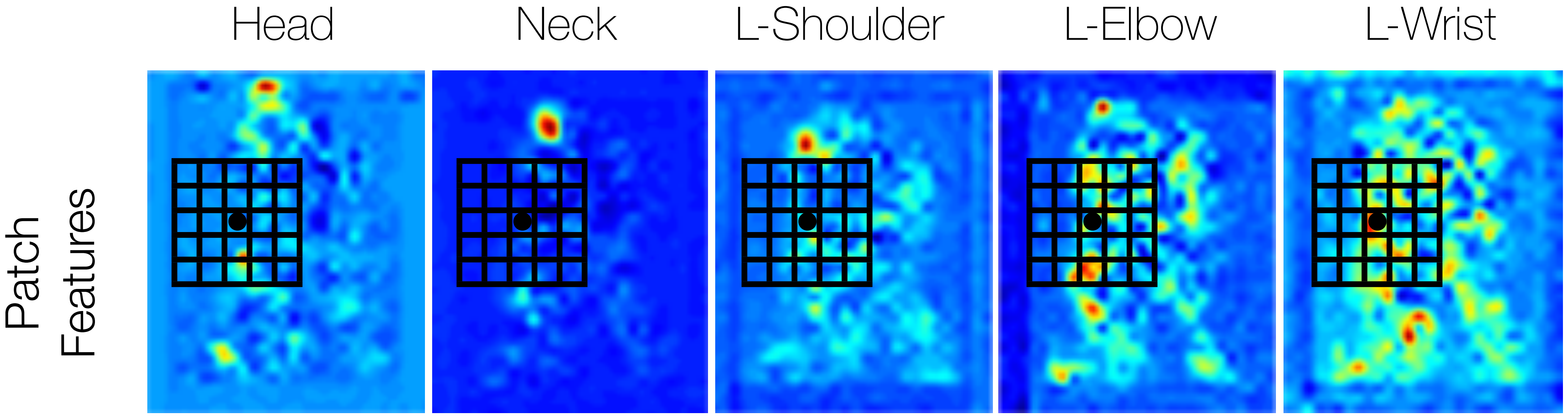
Part Context is a Strong Cue

Part detection confidences provide spatial context cues



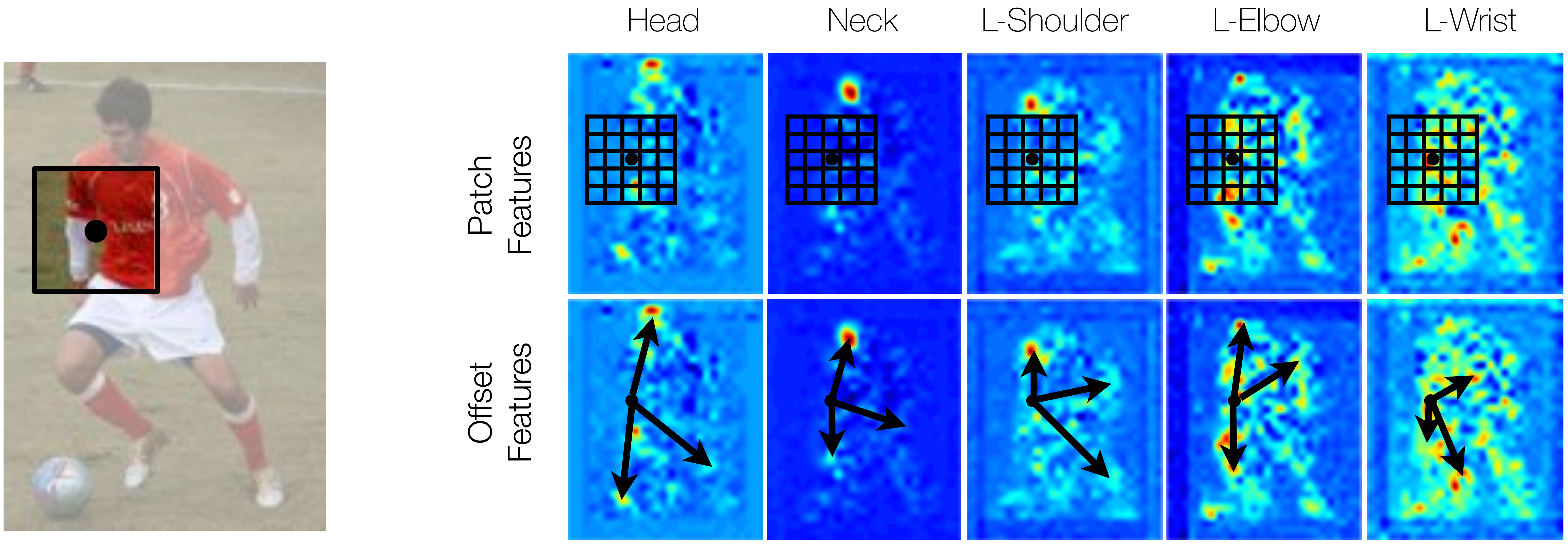
Part Context is a Strong Cue

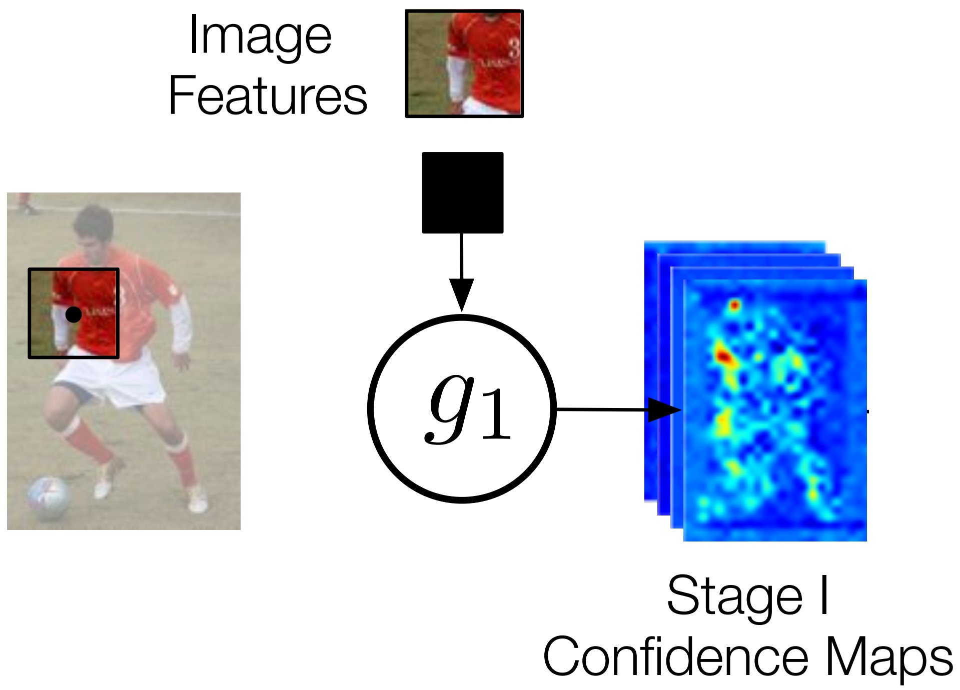
Context features summarize responses of a previous prediction stage



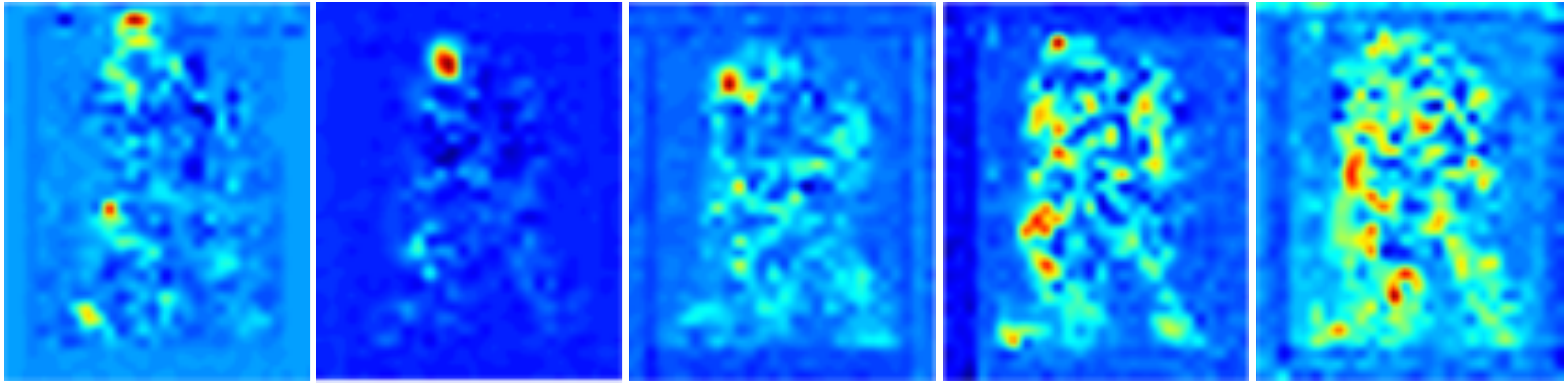
Part Context is a Strong Cue

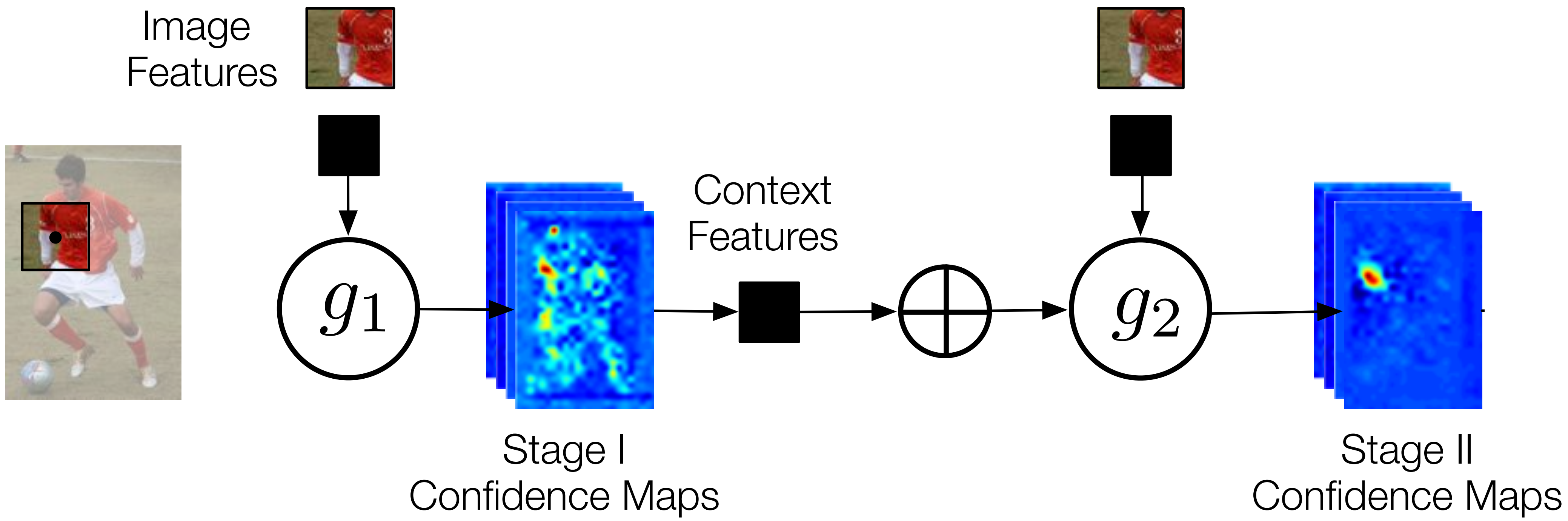
Context features summarize responses of a previous prediction stage



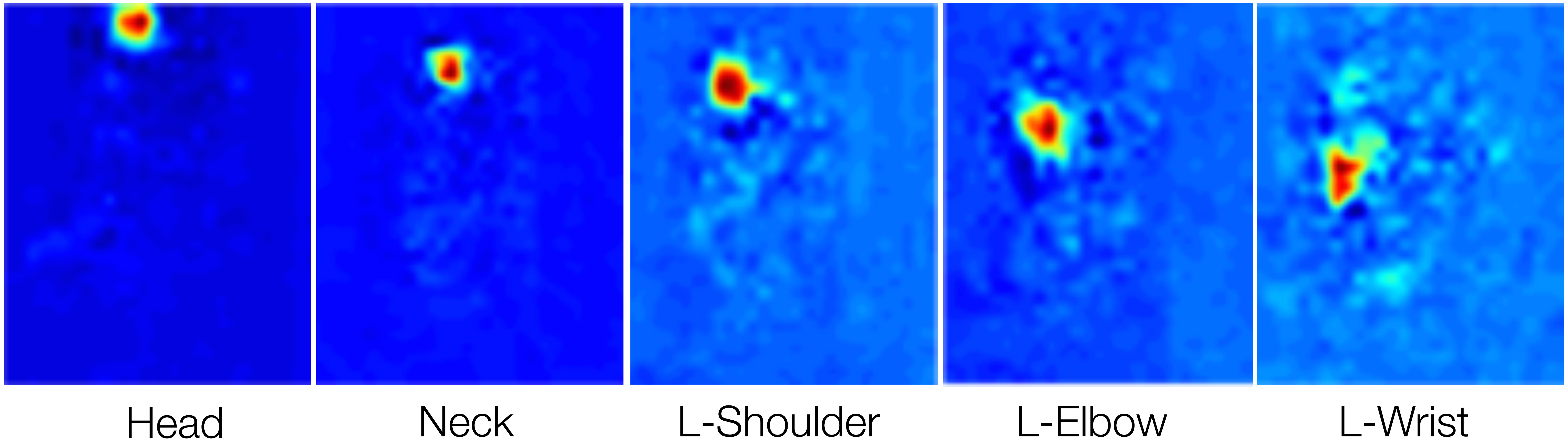


Stage I Confidence


Head | Neck | L-Shoulder | L-Elbow | L-Wrist



Stage II Confidence



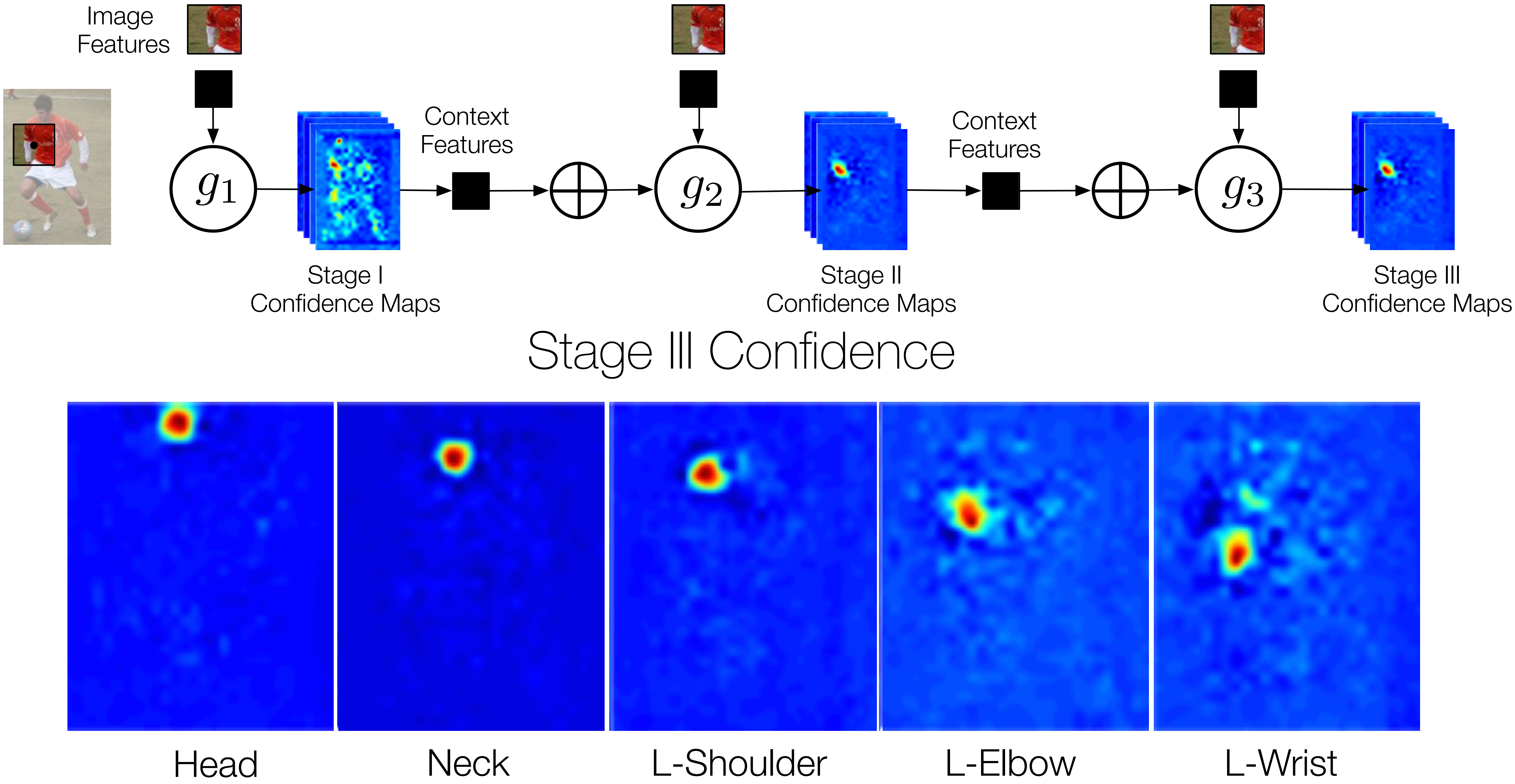


Image
Features
g_1
Context
Features
g_2
Context
Features
g_3
Stage I
Confidence Maps
Stage II
Confidence Maps
Stage III
Confidence Maps
Stage III Confidence
Head
Neck
L-Shoulder
L-Elbow
L-Wrist

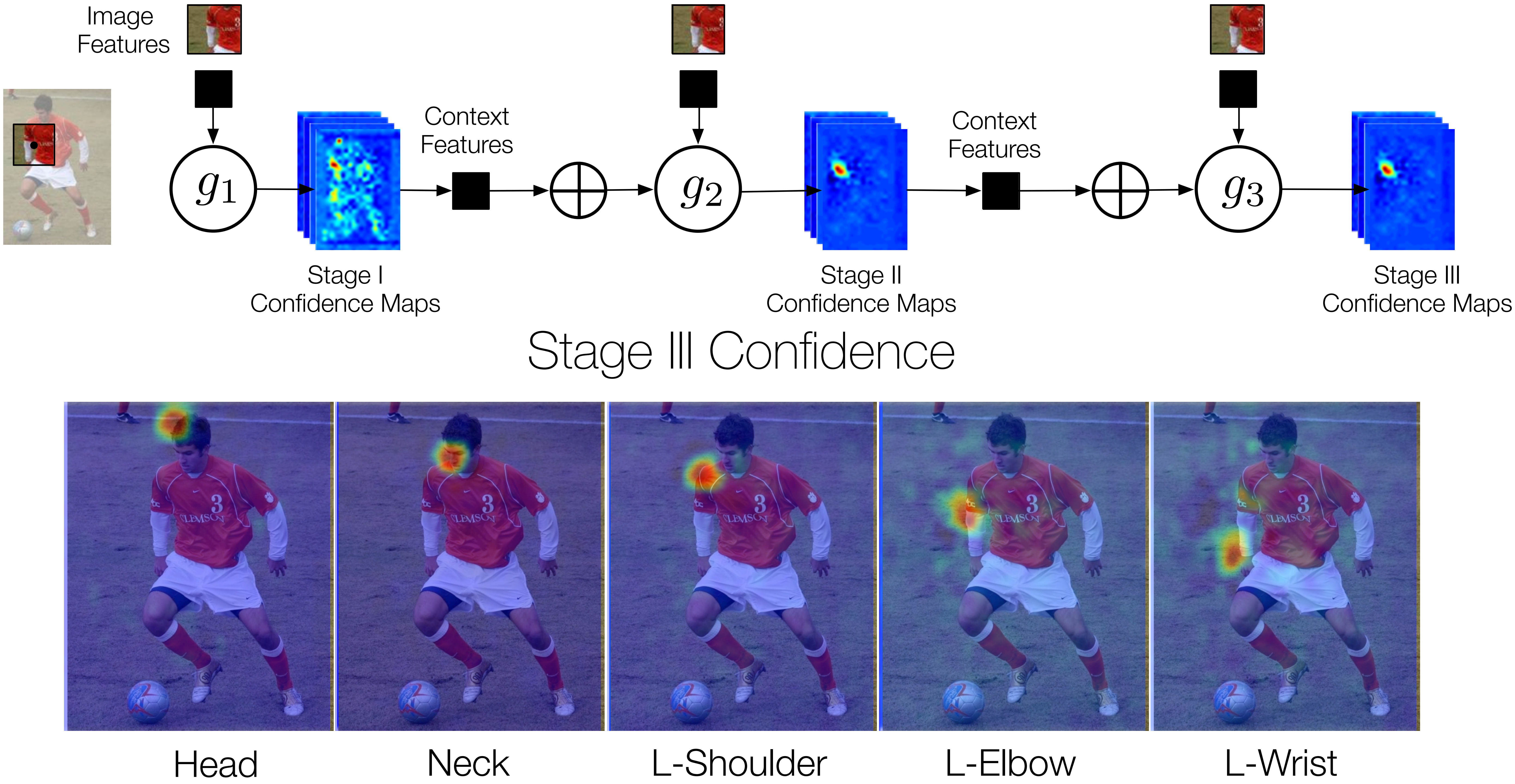
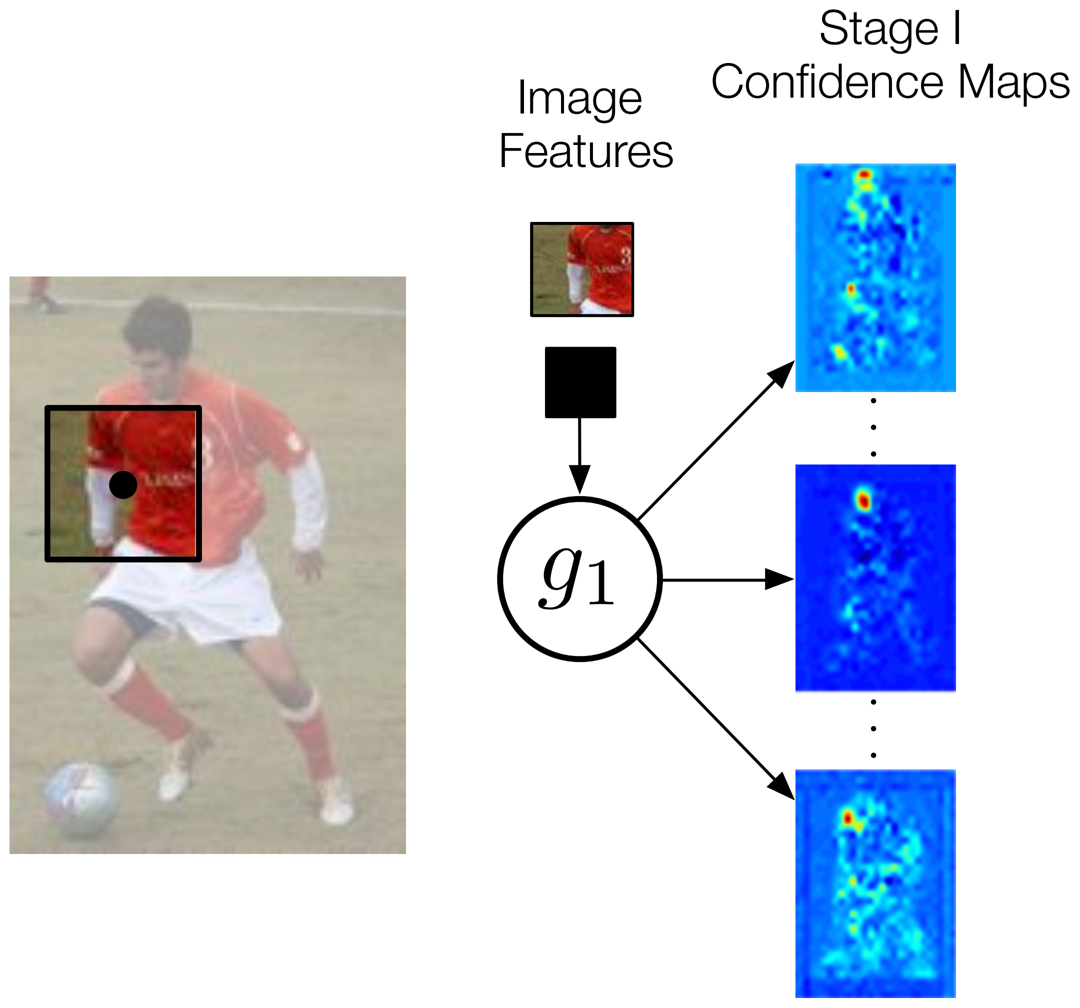


Image Features
Context Features
Context Features
Stage I Confidence Maps
Stage II Confidence Maps
Stage III Confidence Maps
Stage III Confidence
Head
Neck
L-Shoulder
L-Elbow
L-Wrist

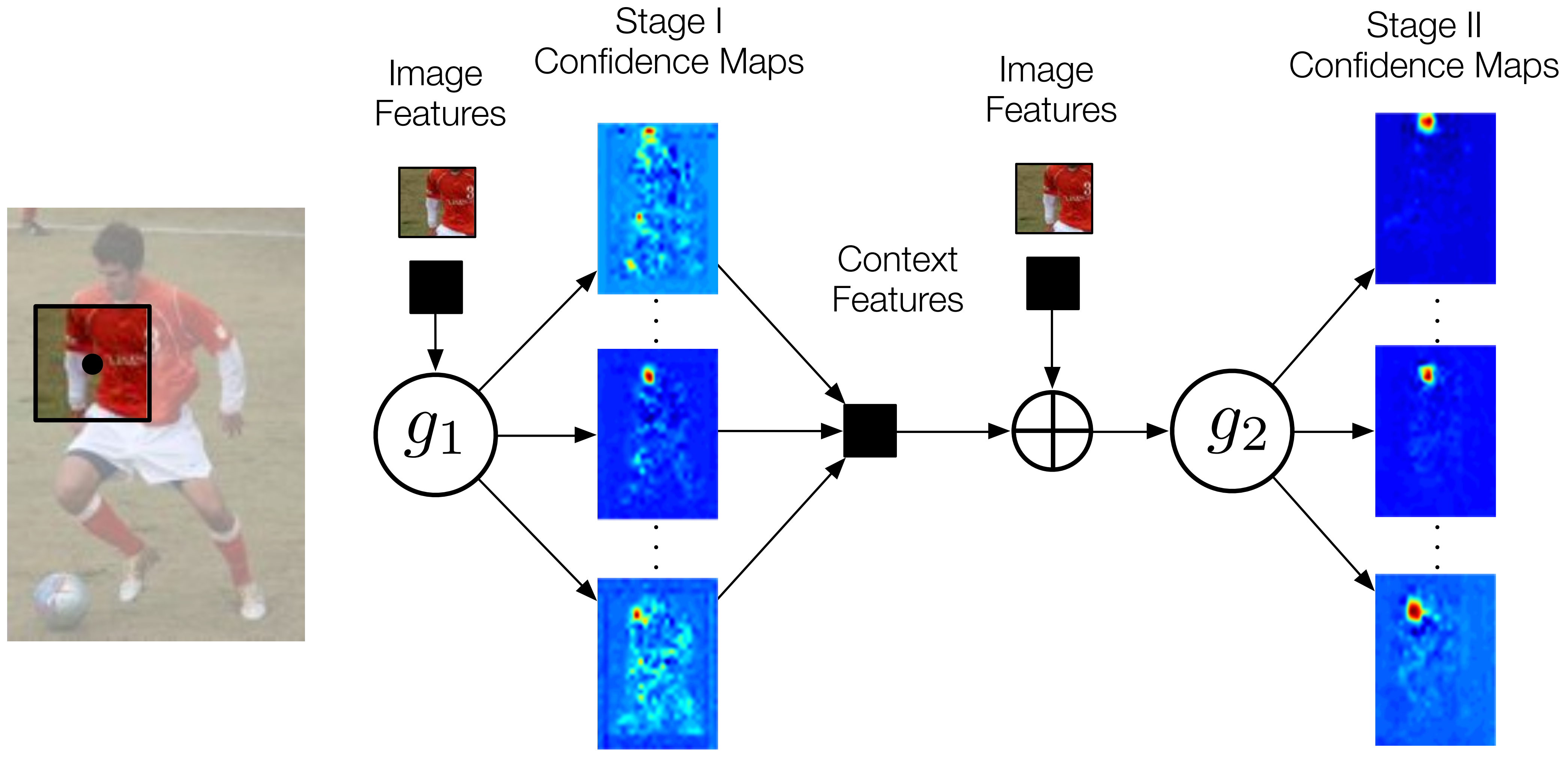
Inference Machines for Pose Estimation

Reduces structured prediction to a sequence of simple classification problems



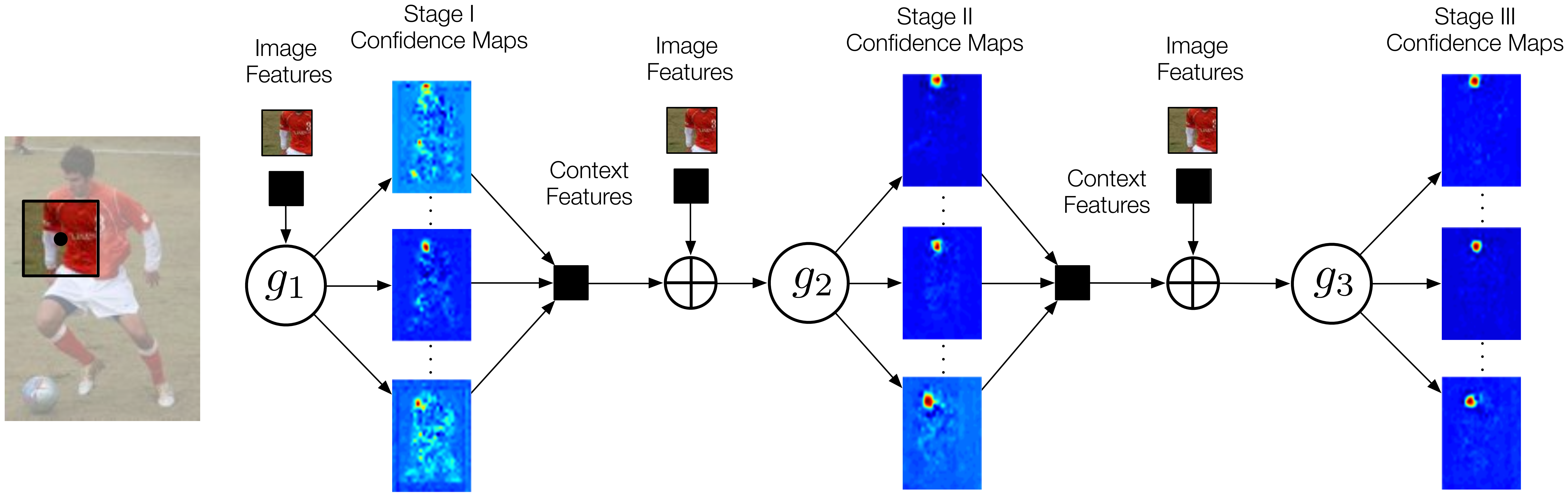
Inference Machines for Pose Estimation

Reduces structured prediction to a sequence of simple classification problems



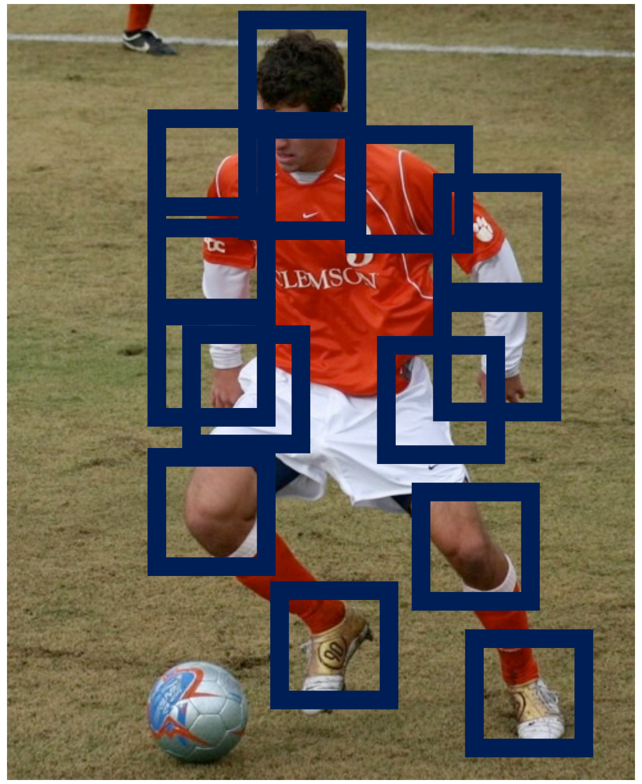
Inference Machines for Pose Estimation

Reduces structured prediction to a sequence of simple classification problems

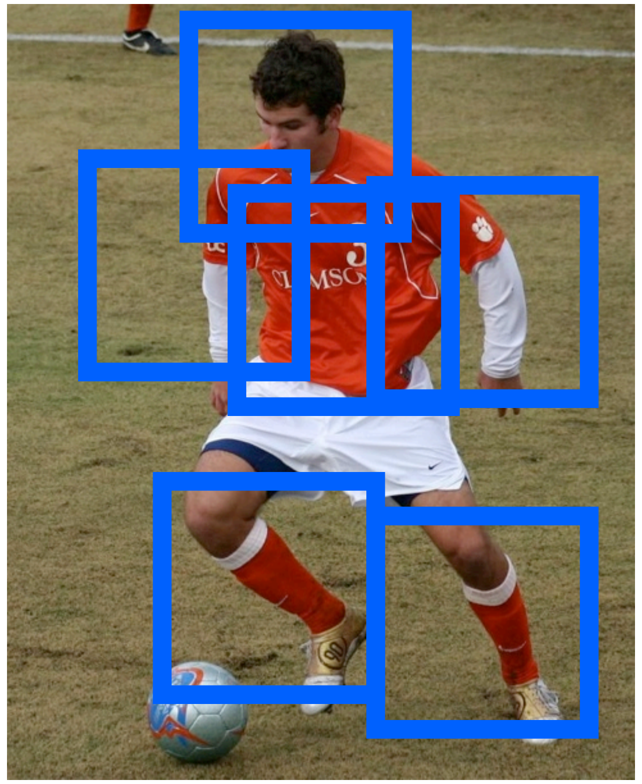


Larger Composite Parts are Easier to Detect

Level 1 parts



Level 2 parts

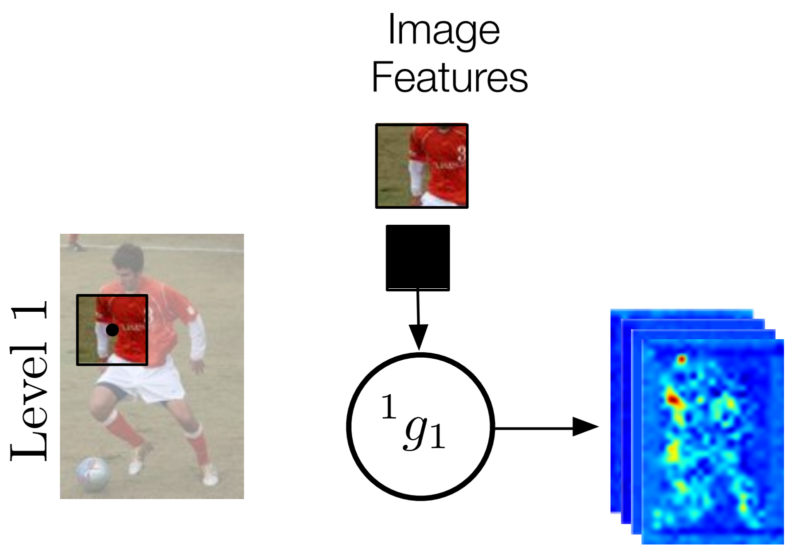


Level 3 parts

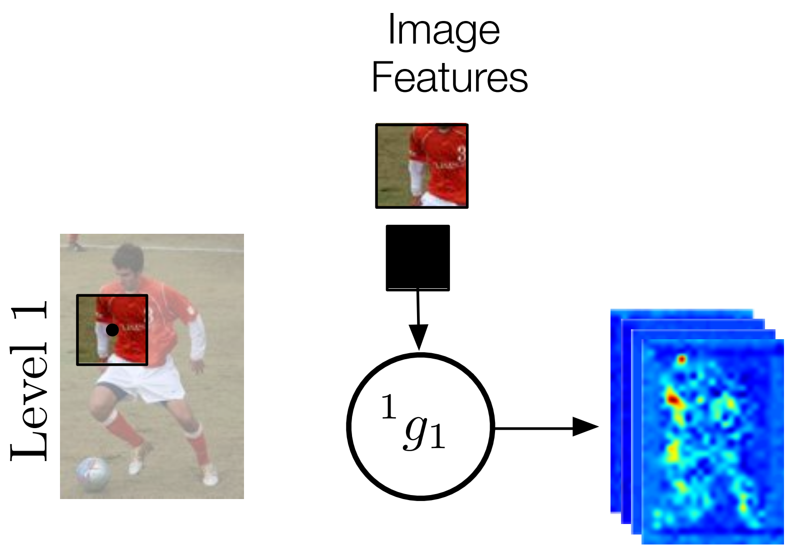
[Bourdev et al., CVPR 2009]
[Sun et al., CVPR 2012]
[Duan et al., BMVC 2012]
[Singh et al., ECCV 2012]
[Pishchulin et al., CVPR 2013] etc.

Incorporating a Part Hierarchy

Incorporating a Part Hierarchy

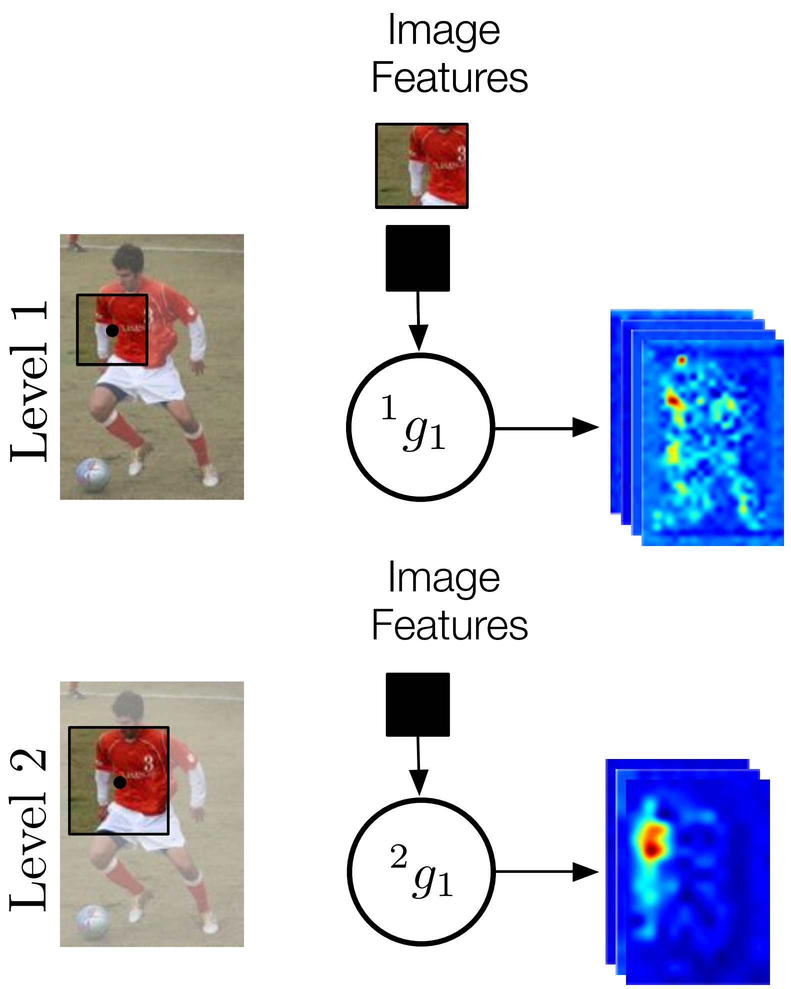


Incorporating a Part Hierarchy



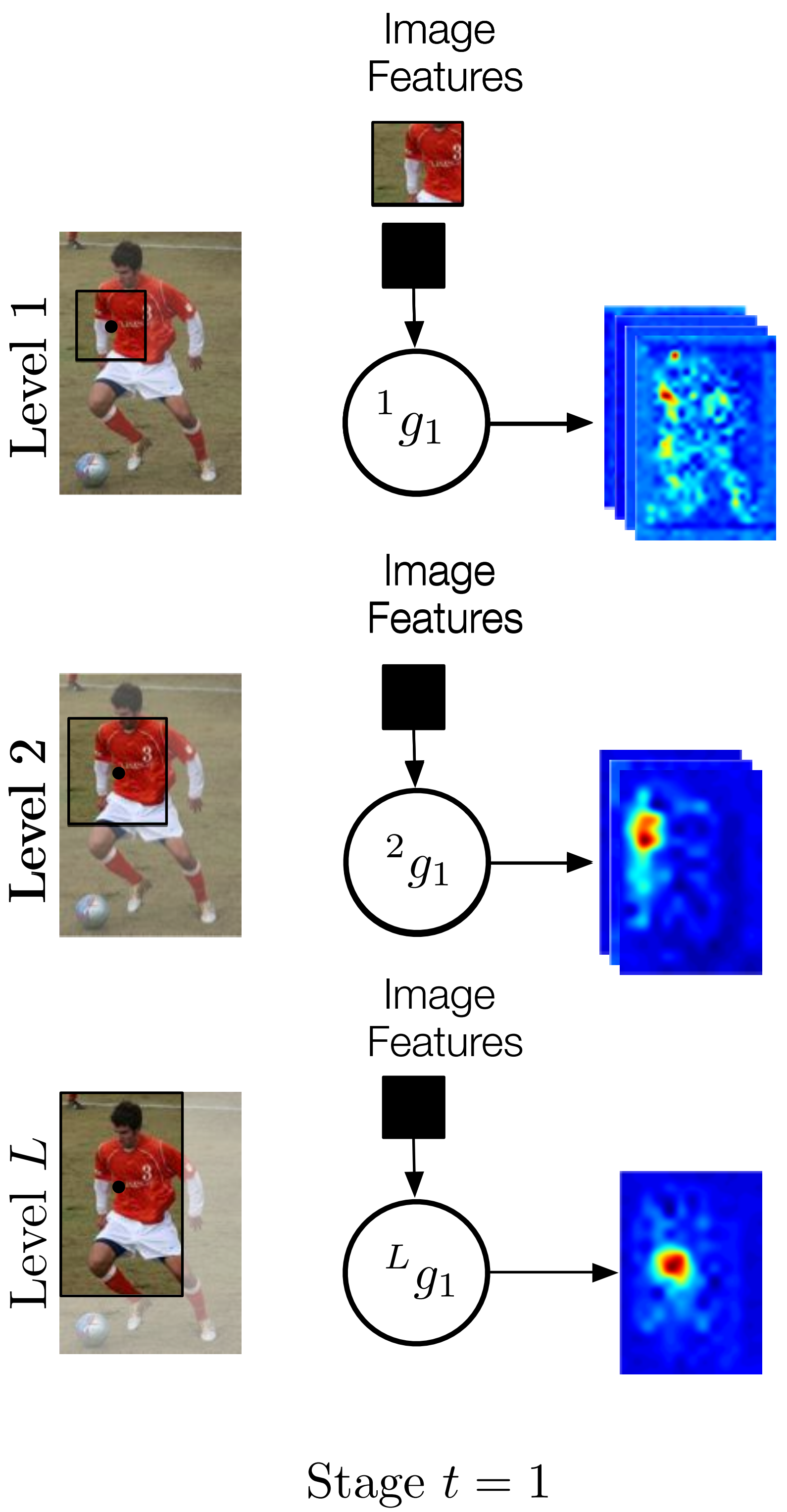
Each level of the hierarchy uses a separate predictor

Incorporating a Part Hierarchy



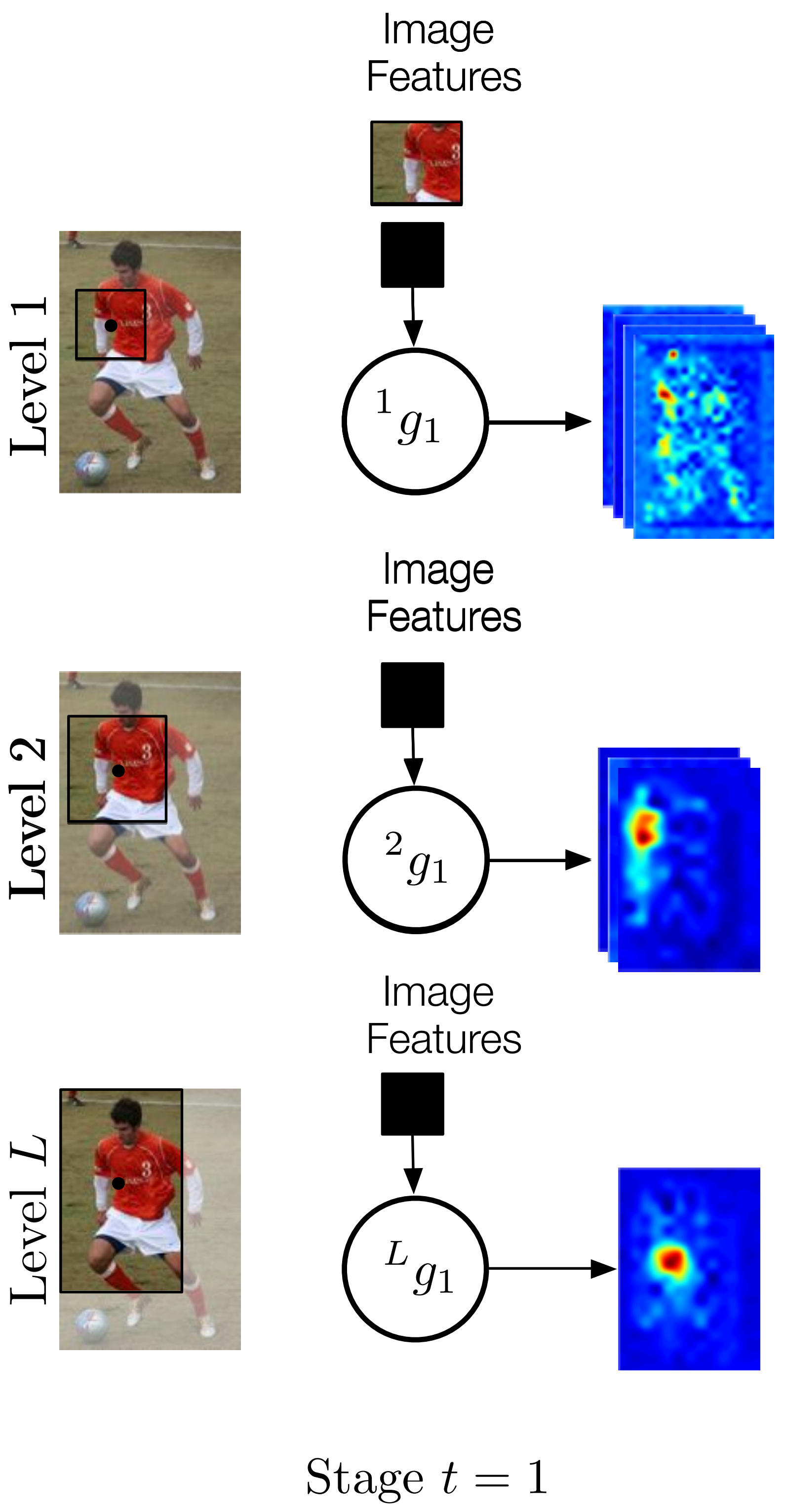
Each level of the hierarchy uses a separate predictor

Incorporating a Part Hierarchy

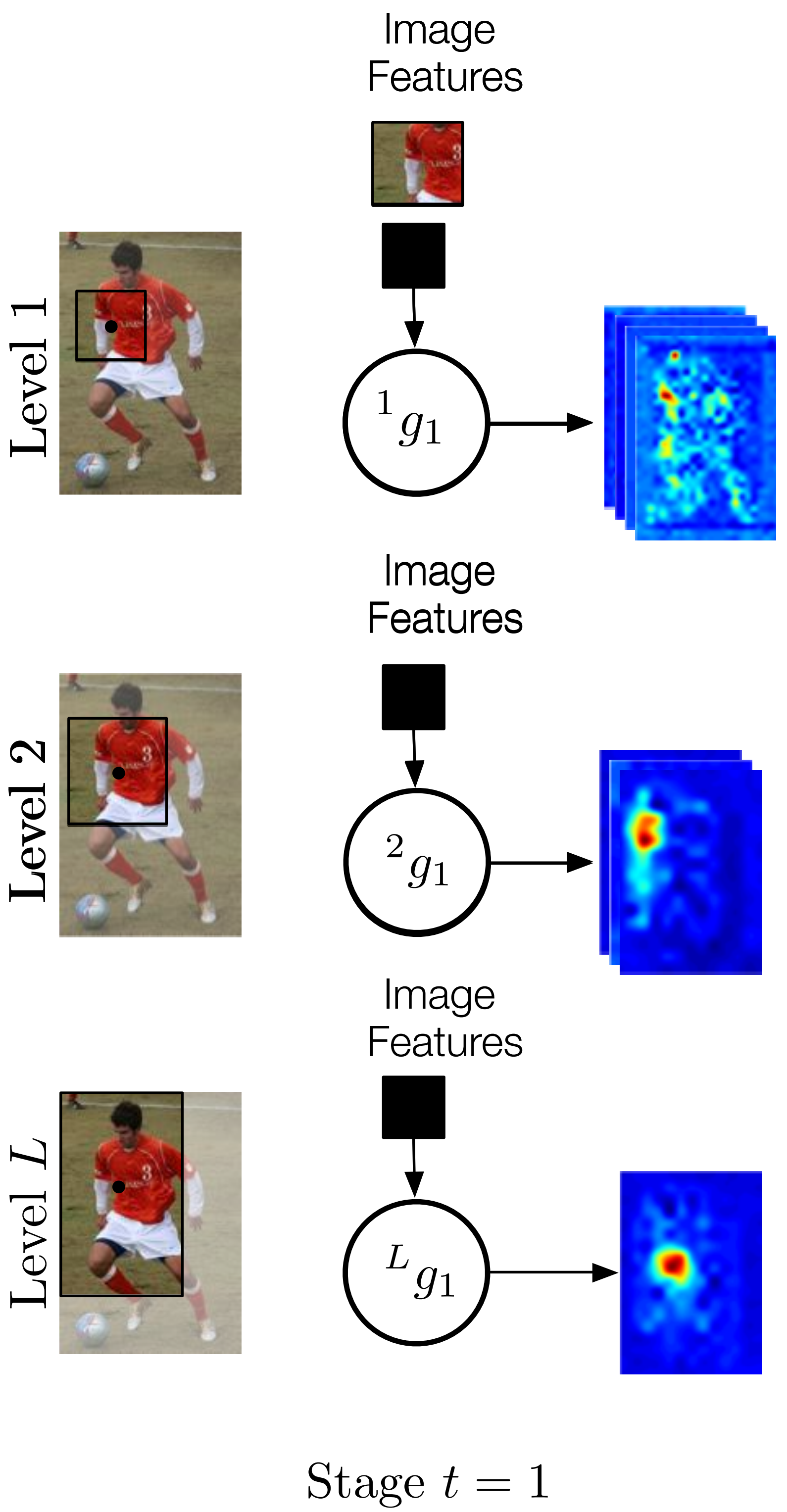


Each level of the hierarchy uses a separate predictor

Incorporating a Part Hierarchy

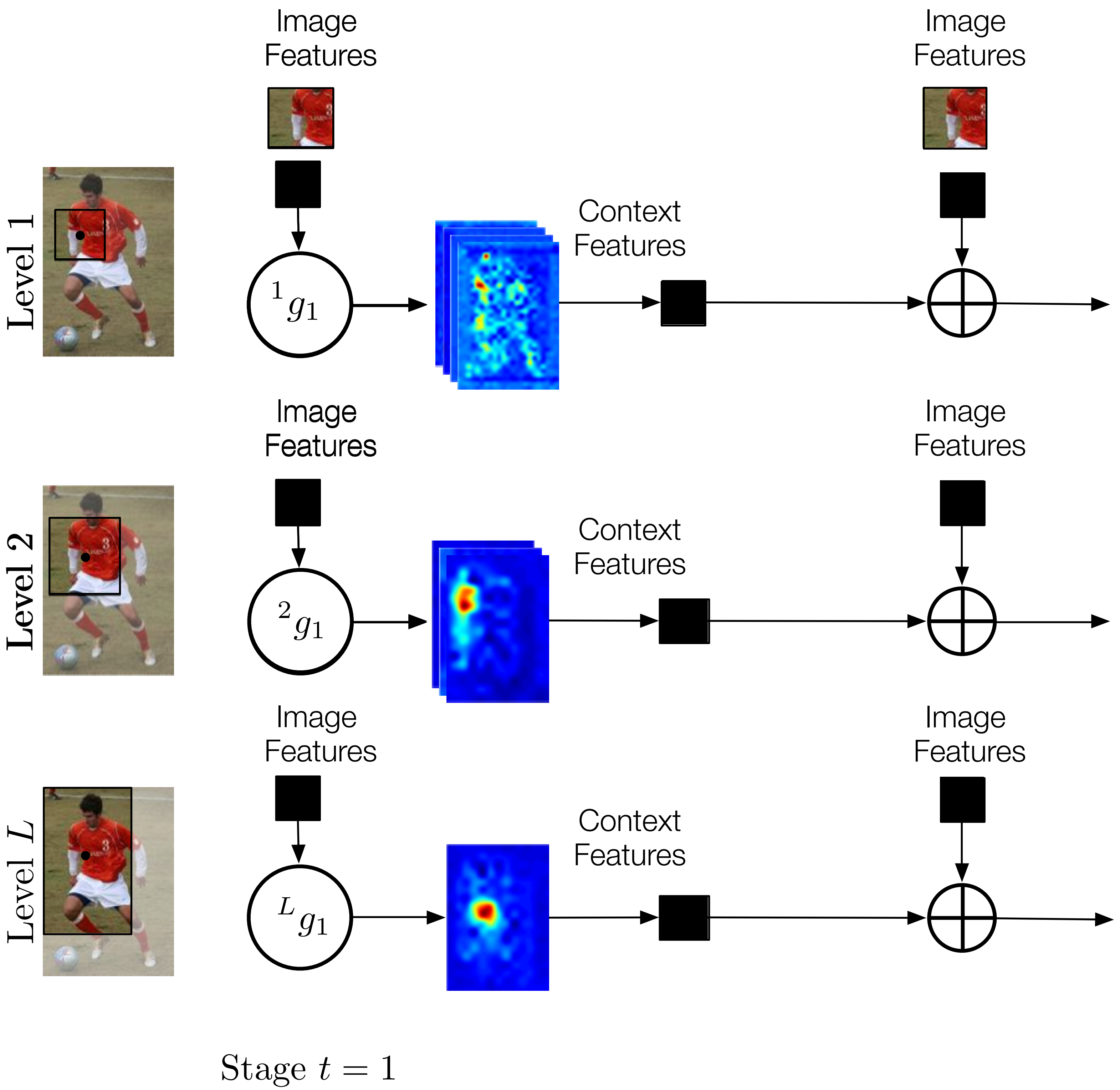


Incorporating a Part Hierarchy



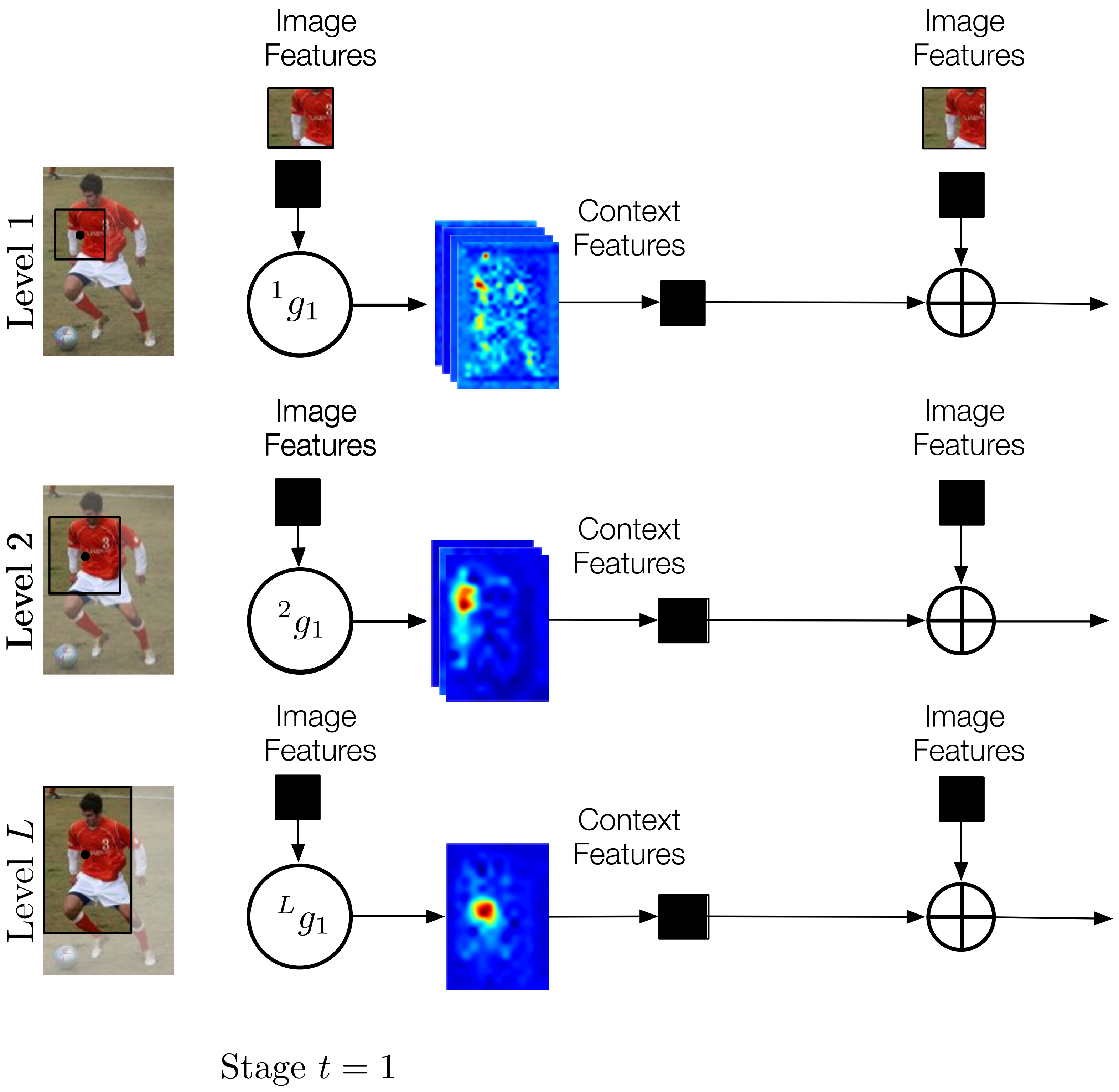
Context Features are computed on the outputs of the previous stage

Incorporating a Part Hierarchy

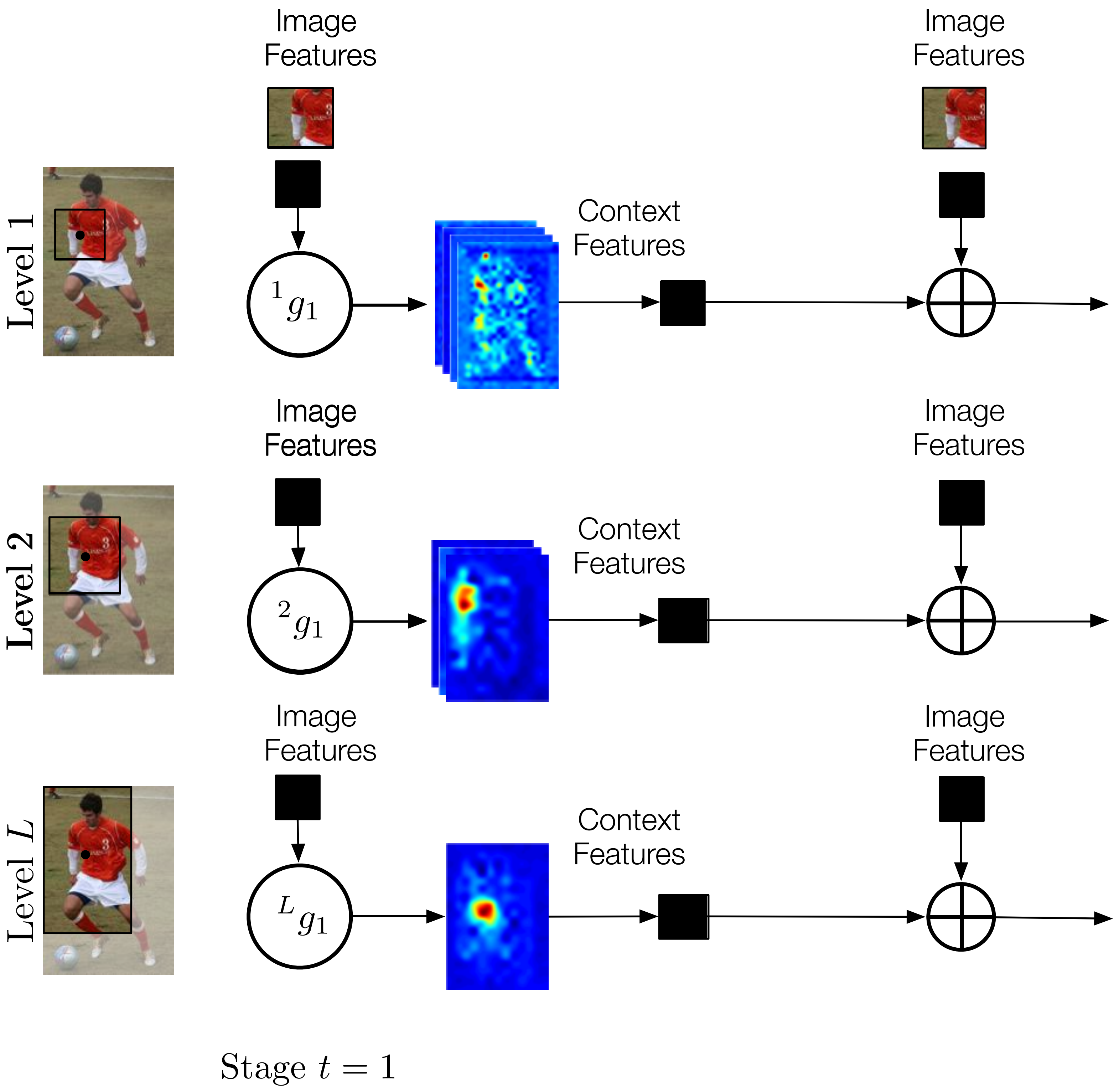


Context Features are computed on the outputs of the previous stage

Incorporating a Part Hierarchy

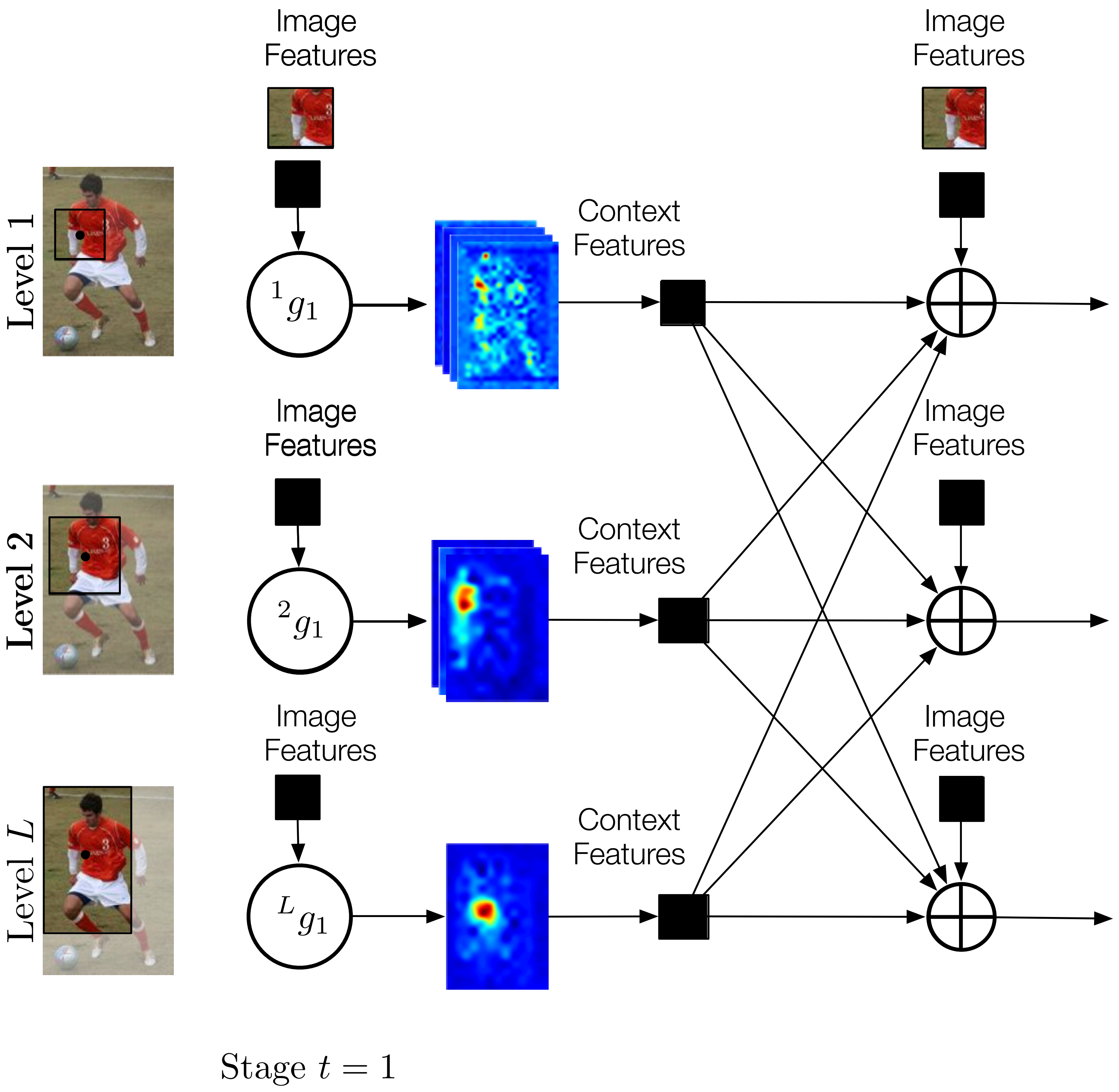


Incorporating a Part Hierarchy



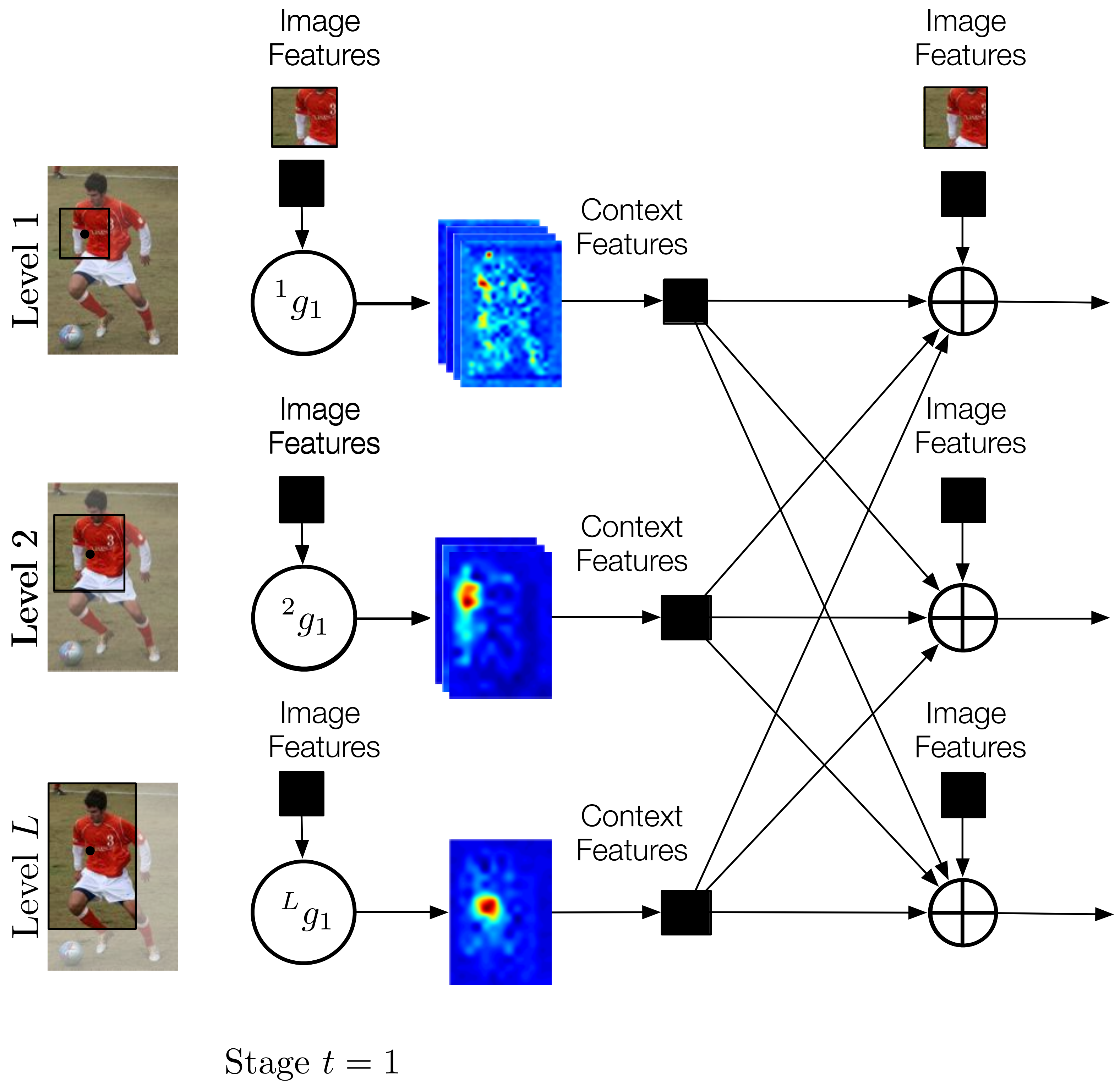
Spatial context information is passed across layers via context features.

Incorporating a Part Hierarchy

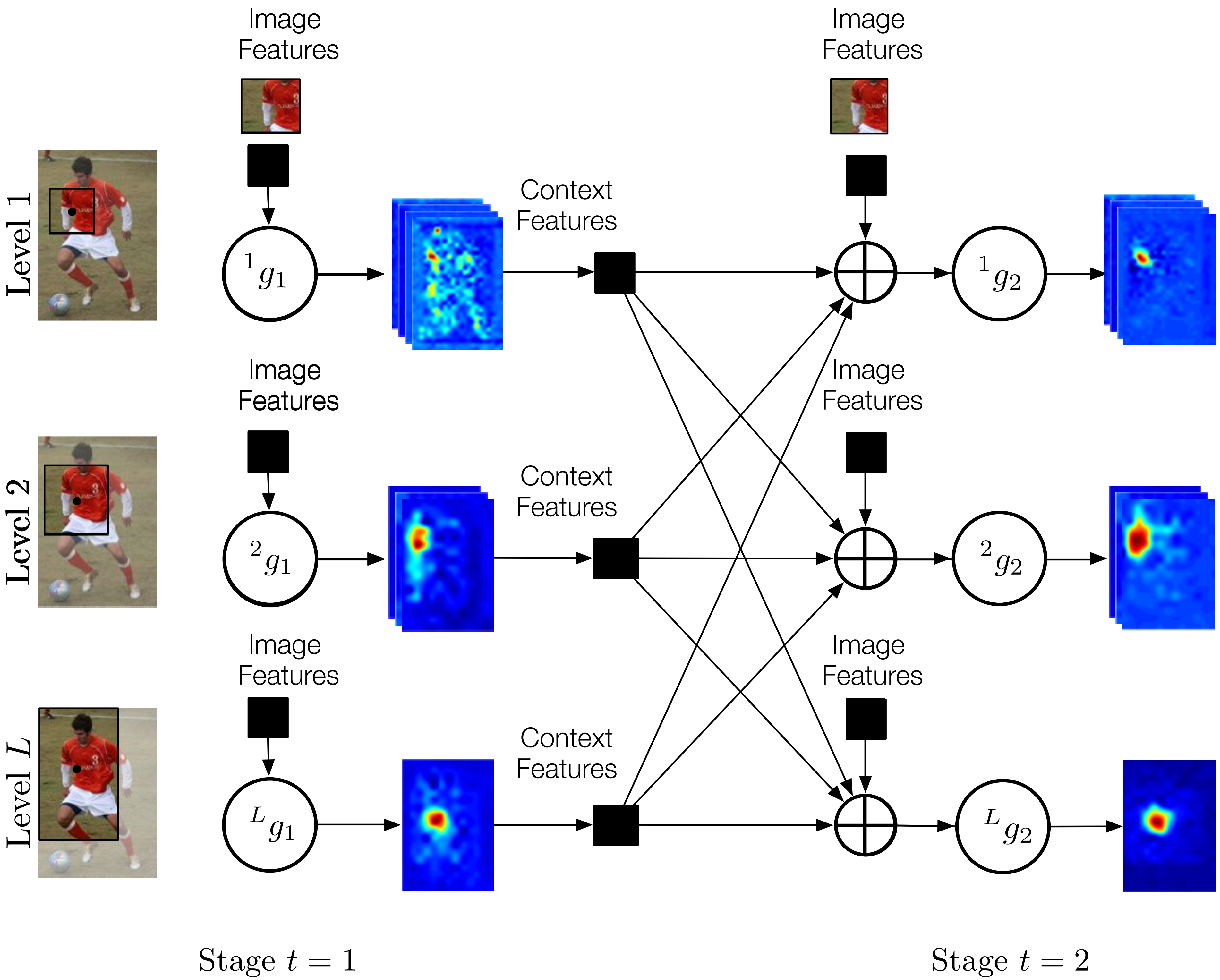


Spatial context information is passed across layers via context features.

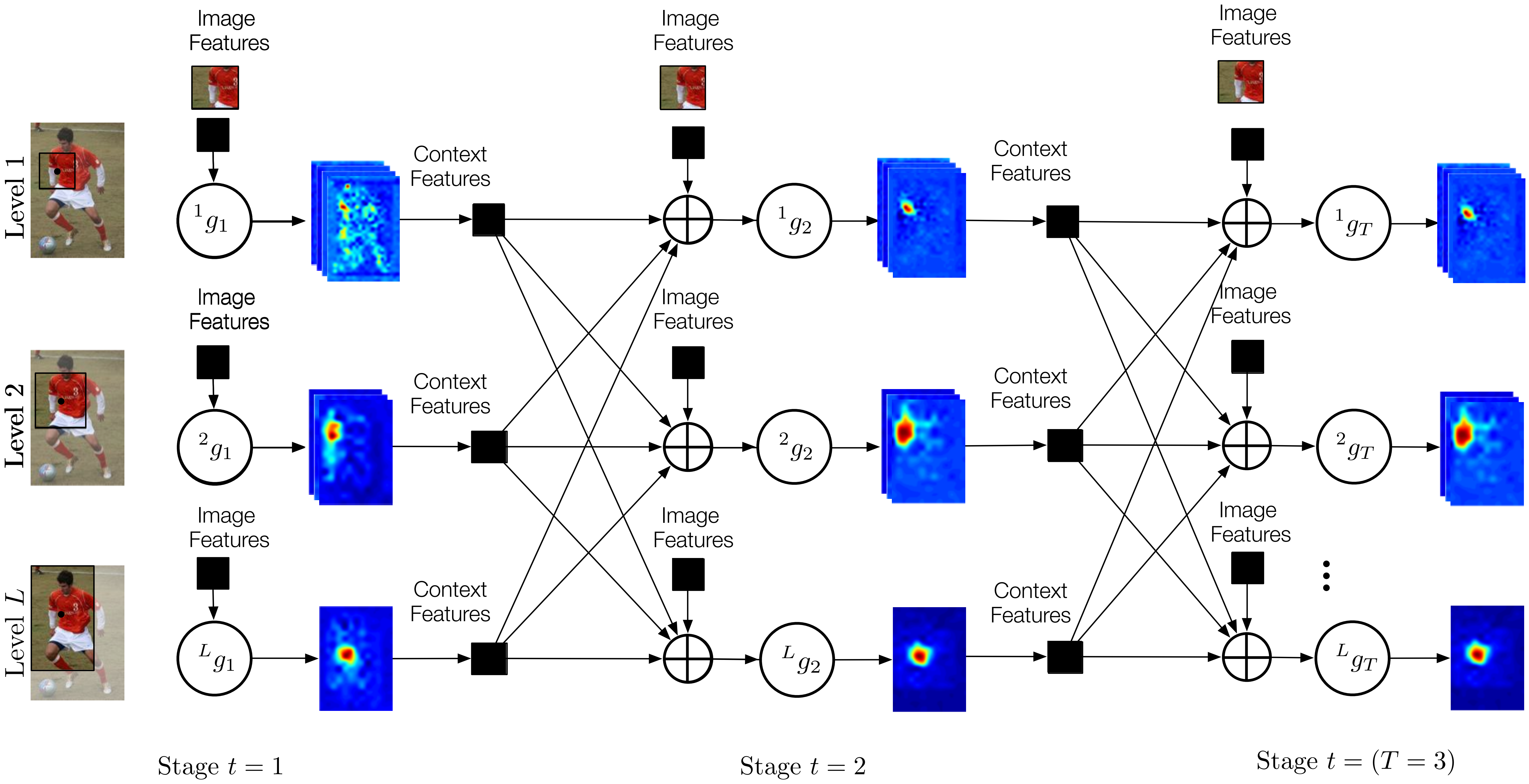
Incorporating a Part Hierarchy

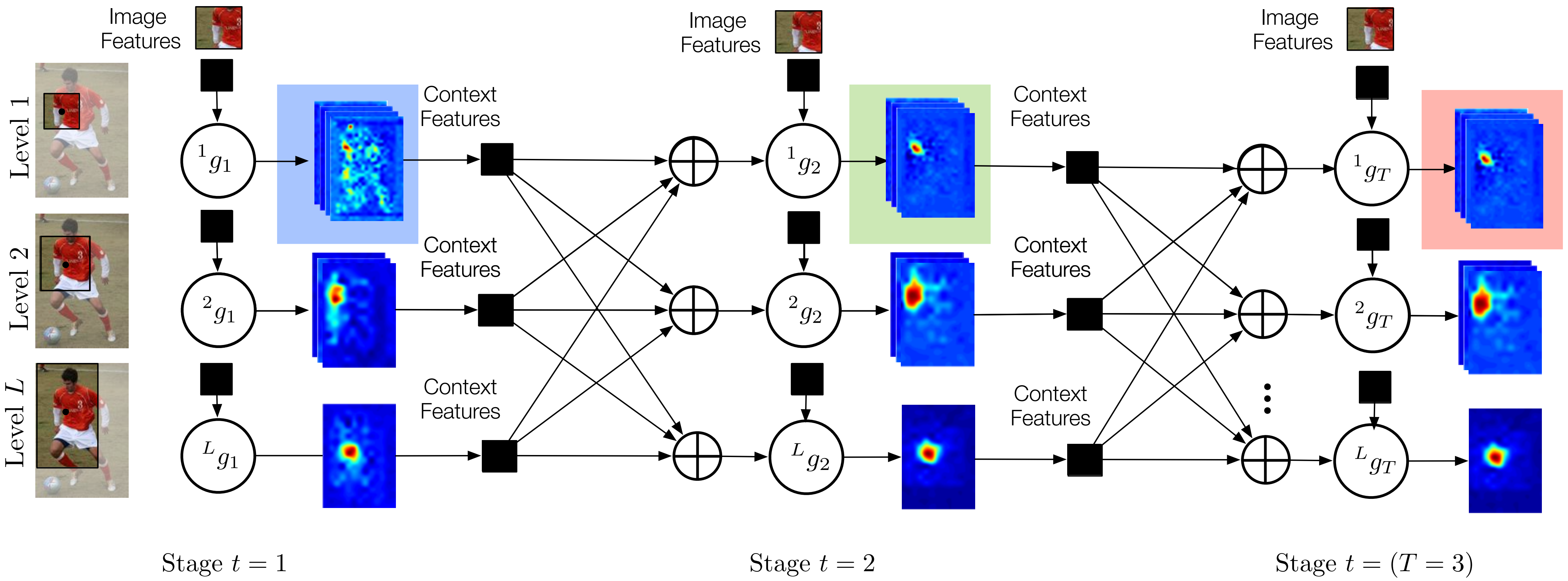


Incorporating a Part Hierarchy

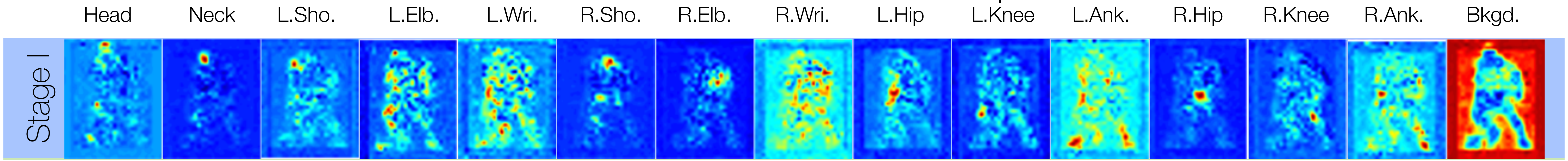


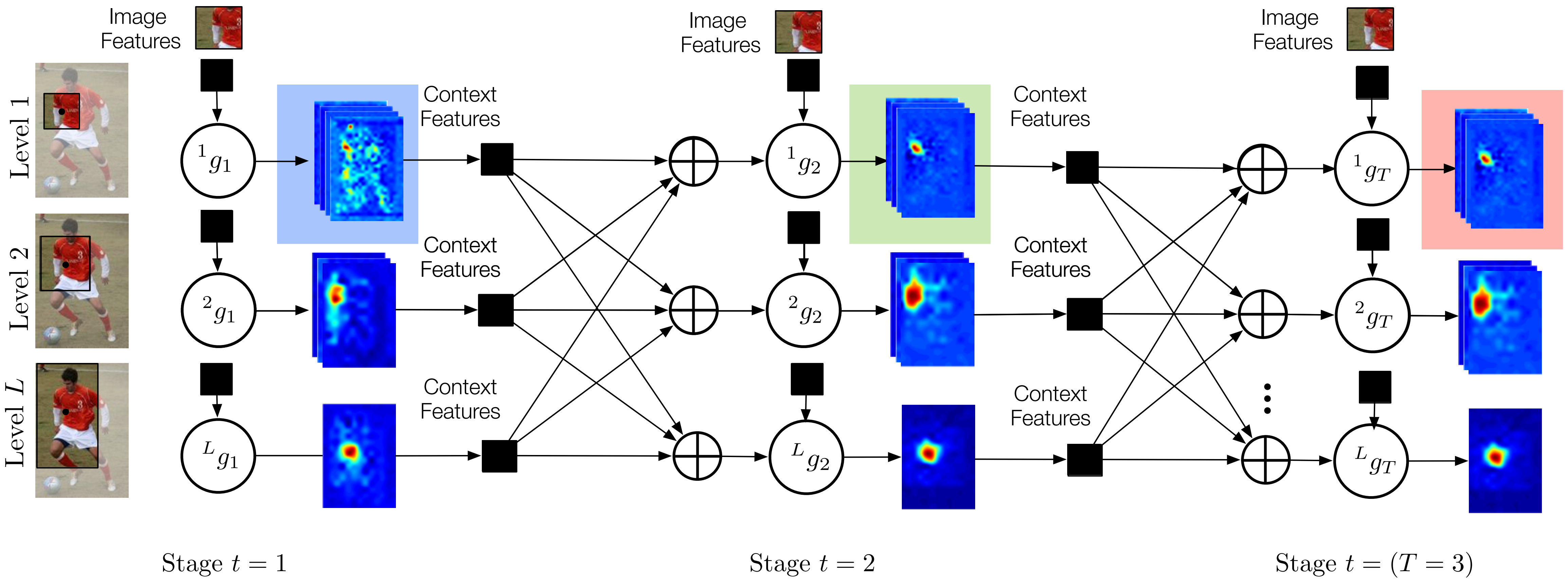
Incorporating a Part Hierarchy



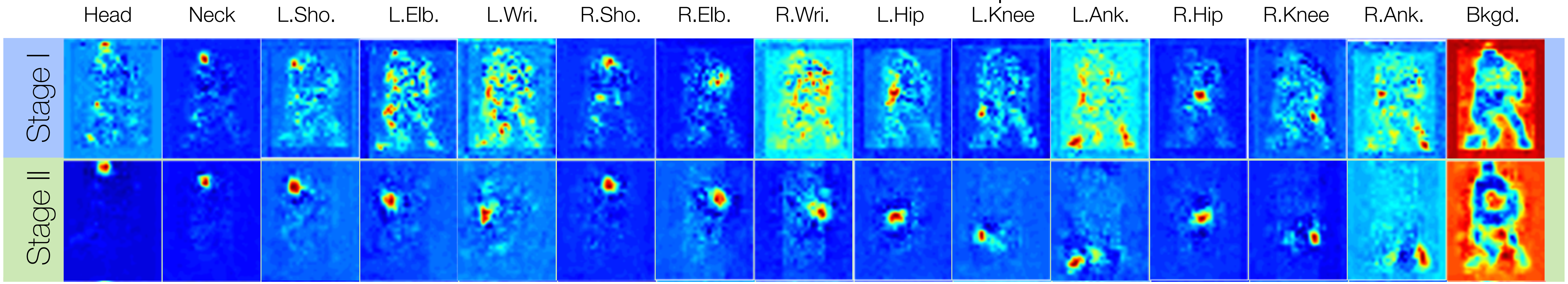


Level l Confidence Maps





Level I Confidence Maps



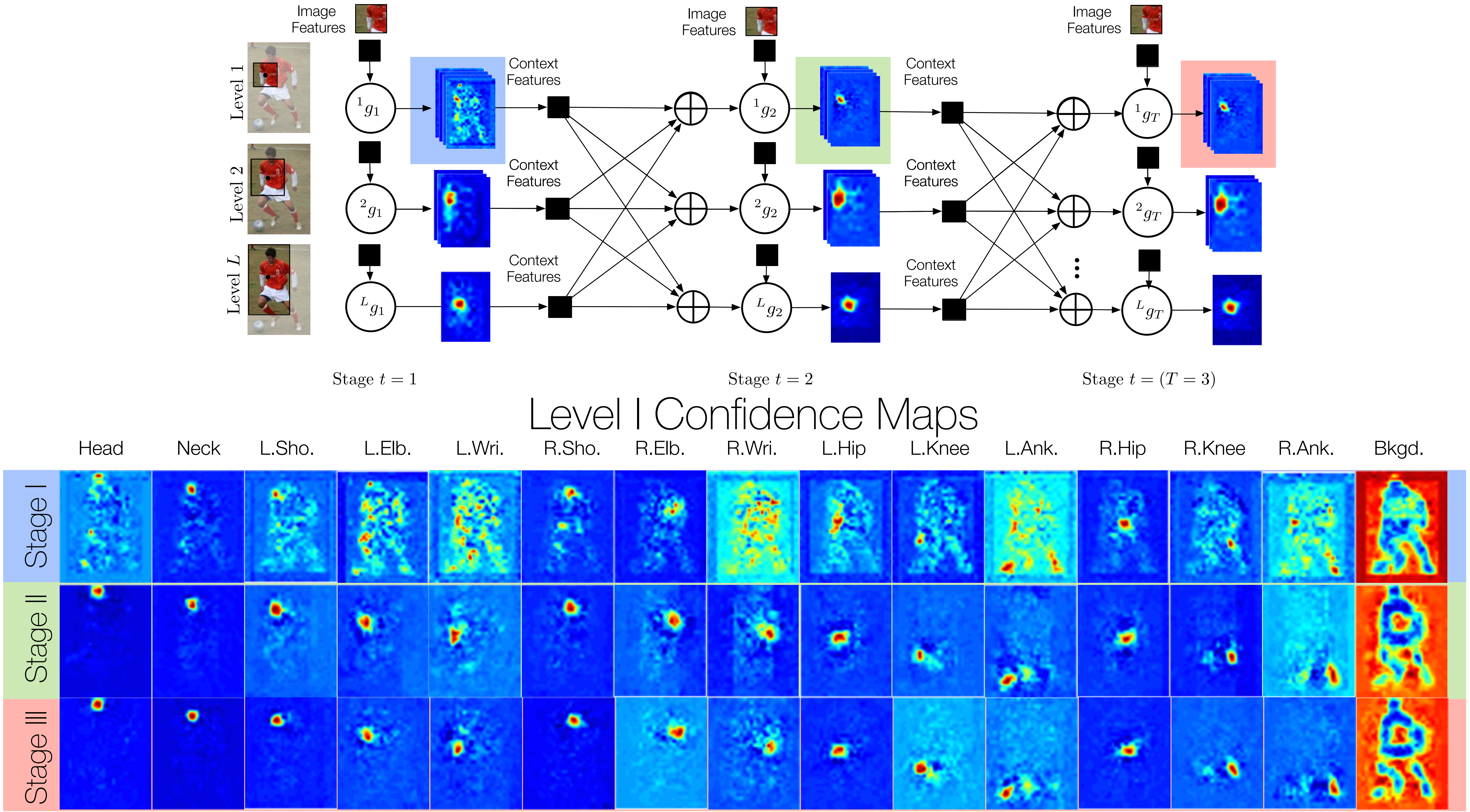


Image Features
Context Features
Level 1
Level 2
Level L
$^1g_1$
$^2g_1$
$^Lg_1$
$^1g_2$
$^2g_2$
$^Lg_2$
^1g_T
^2g_T
^Lg_T
Stage $t = 1$
Stage $t = 2$
Stage $t = (T = 3)$
Level l Confidence Maps
Head
Neck
L.Sho.
L.Elb.
L.Wri.
R.Sho.
R.Elb.
R.Wri.
L.Hip
L.Knee
L.Ank.
R.Hip
R.Knee
R.Ank.
Bkgd.
Stage I
Stage II
Stage III

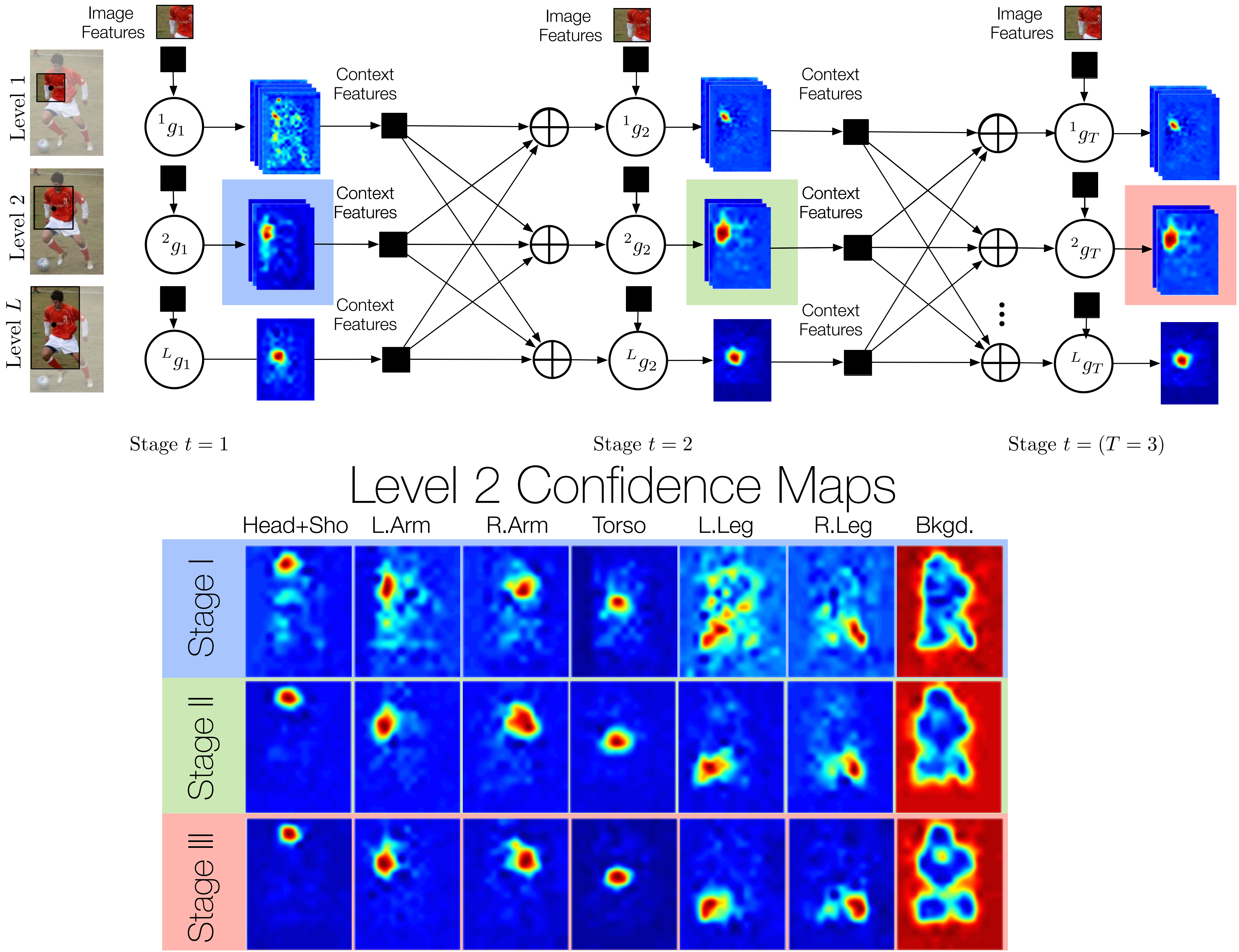
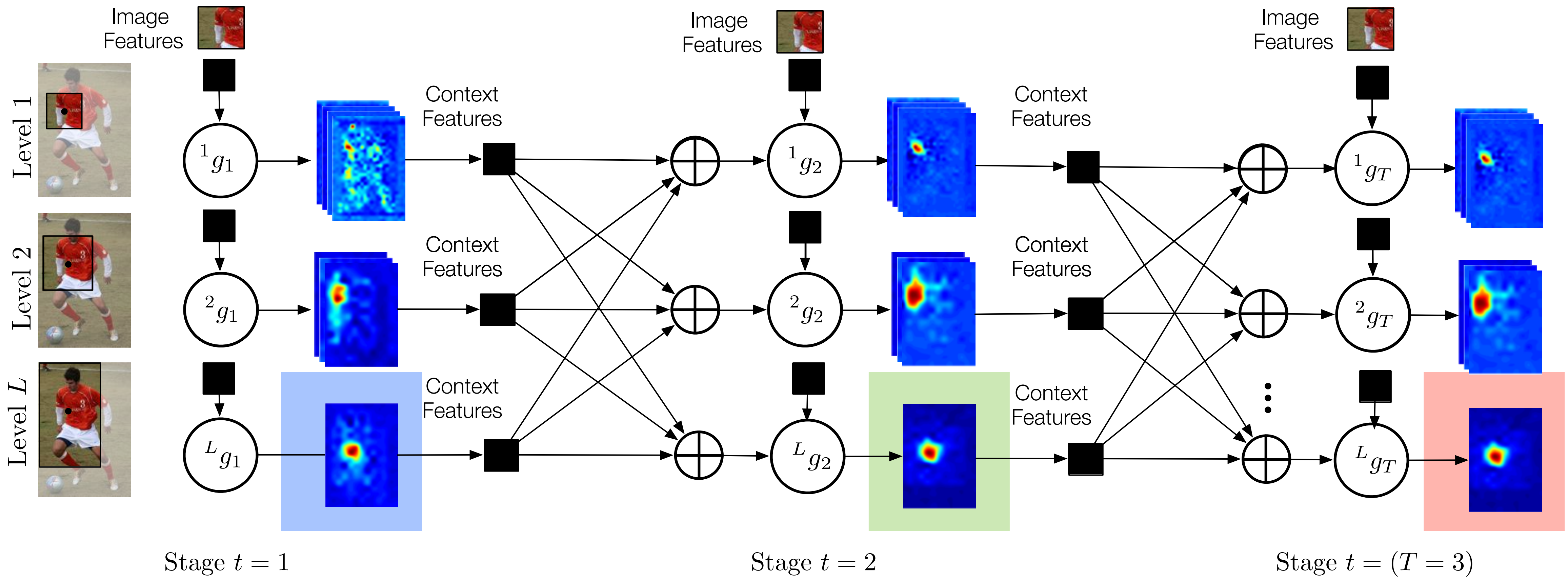
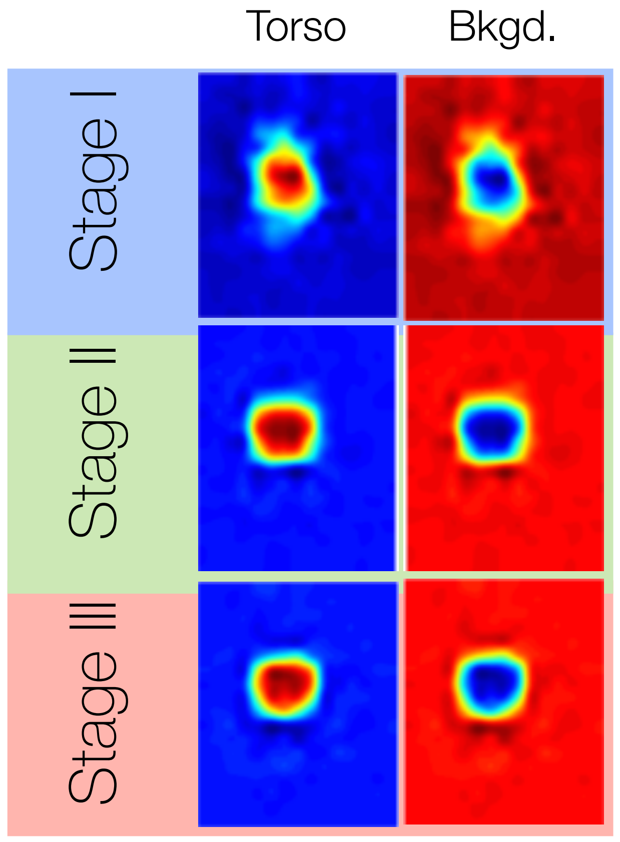
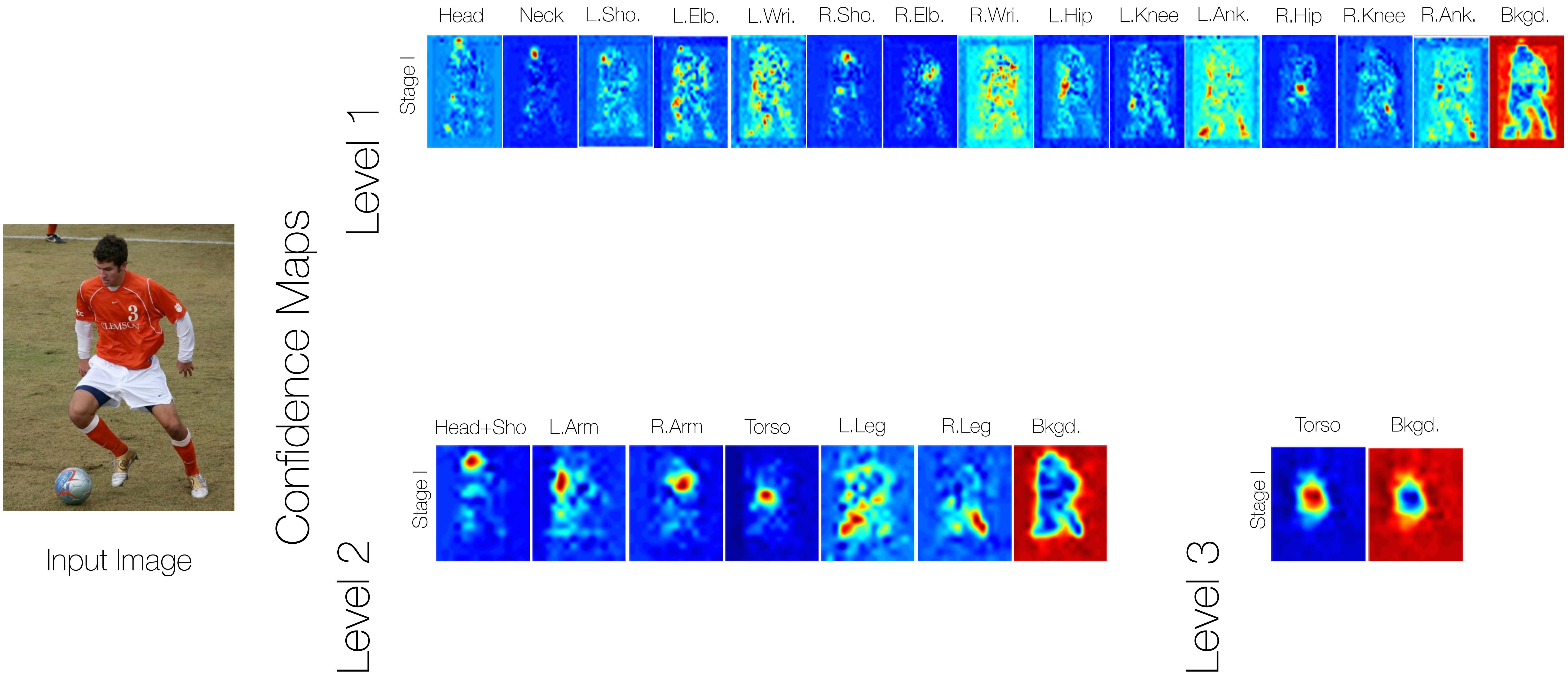


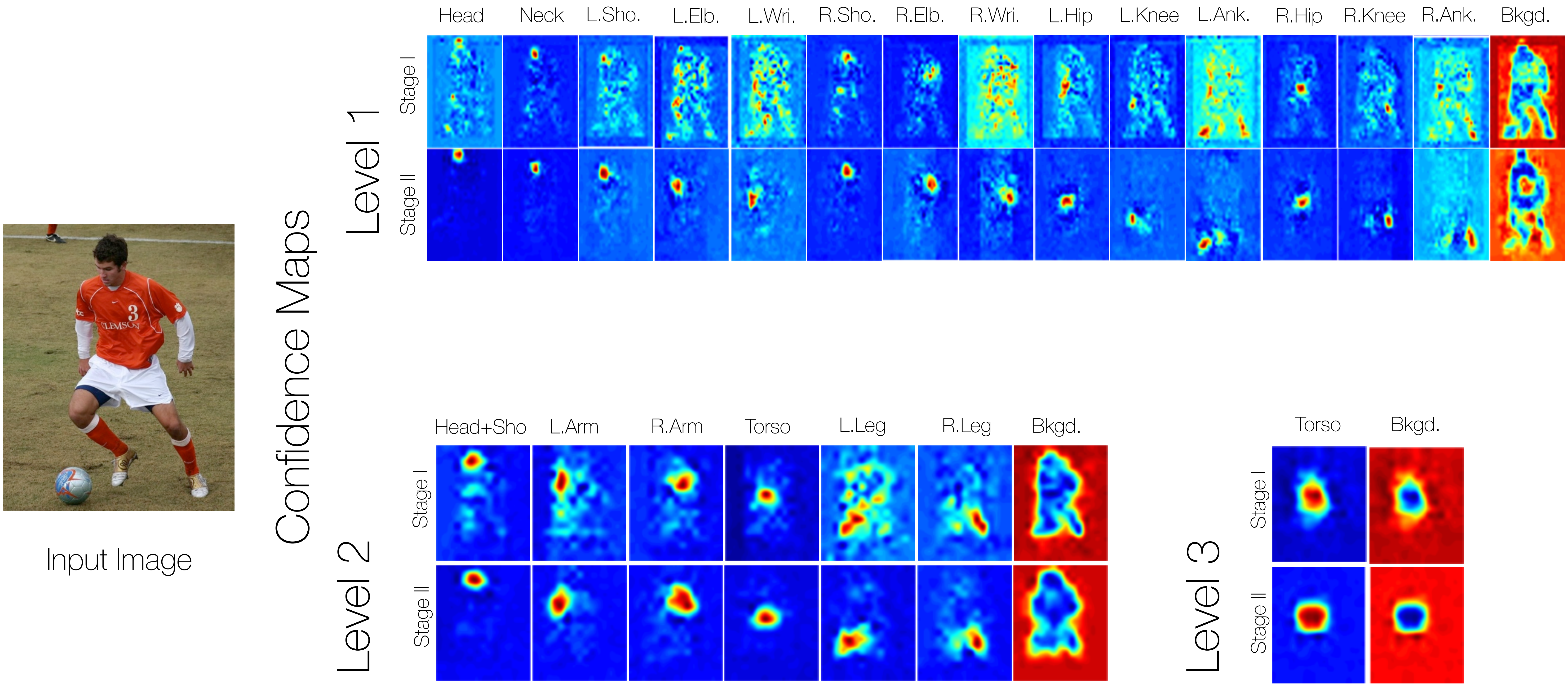
Image Features
Level 1
Level 2
Level L
Context Features
${}^{1}g_1$
${}^{2}g_1$
${}^{L}g_1$
${}^{1}g_2$
${}^{2}g_2$
${}^{L}g_2$
${}^{1}g_T$
${}^{2}g_T$
${}^{L}g_T$
Stage $t = 1$
Stage $t = 2$
Stage $t = (T = 3)$
Level 2 Confidence Maps
Head+Sho
L.Arm
R.Arm
Torso
L.Leg
R.Leg
Bkgd.
Stage I
Stage II
Stage III

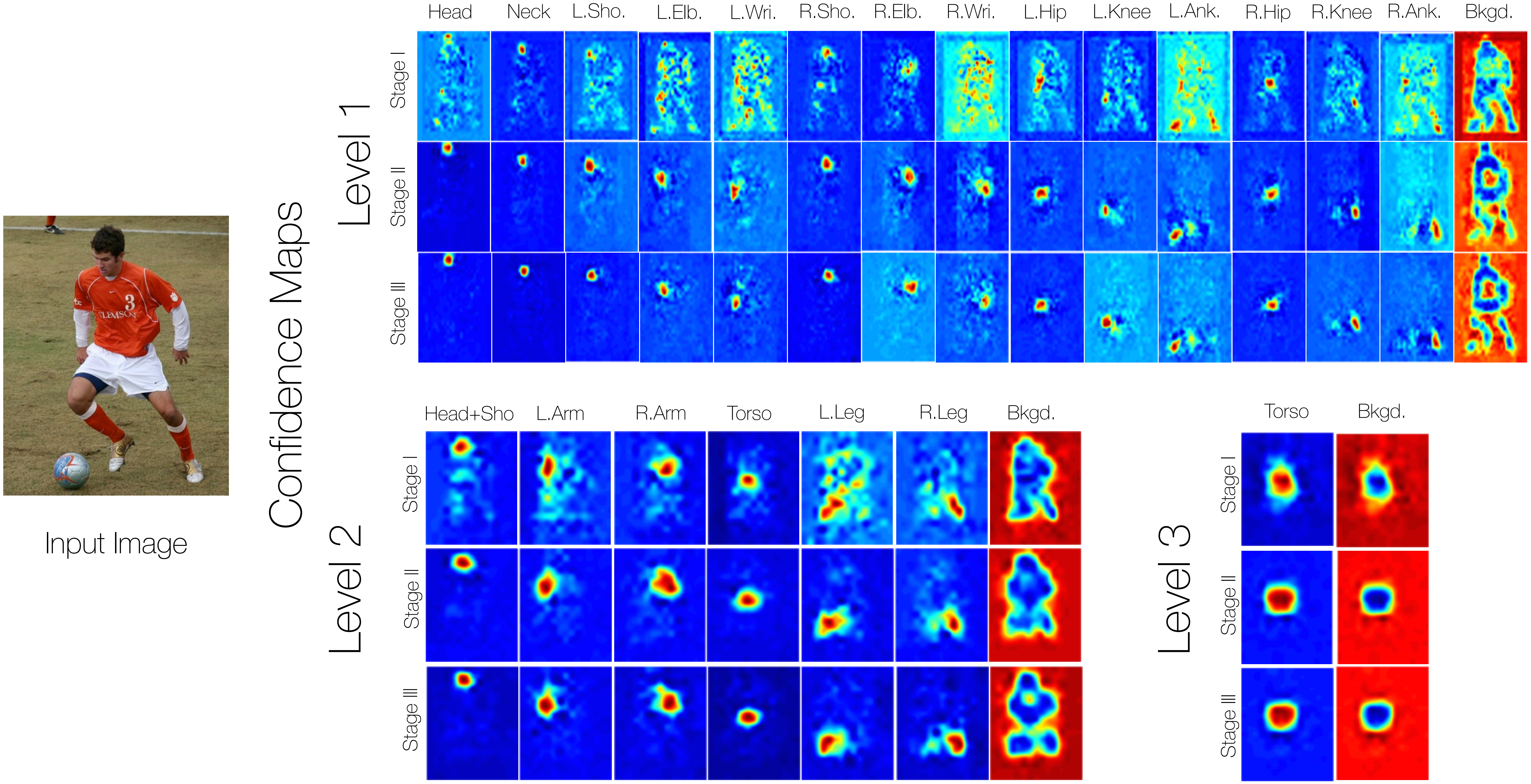


Level 3 Confidence Maps




Input Image
Confidence Maps
Level 1
Stage I
Head
Neck
L.Sho.
L.Elb.
L.Wri.
R.Sho.
R.Elb.
R.Wri.
L.Hip
L.Knee
L.Ank.
R.Hip
R.Knee
R.Ank.
Bkgd.
Level 2
Stage I
Head+Sho
L.Arm
R.Arm
Torso
L.Leg
R.Leg
Bkgd.
Level 3
Stage I
Torso
Bkgd.


Input Image
Confidence Maps
Level 1
Head
Neck
L.Sho.
L.Elb.
L.Wri.
R.Sho.
R.Elb.
R.Wri.
L.Hip
L.Knee
L.Ank.
R.Hip
R.Knee
R.Ank.
Bkgd.
Stage I
Stage II
Level 2
Head+Sho
L.Arm
R.Arm
Torso
L.Leg
R.Leg
Bkgd.
Stage I
Stage II
Level 3
Torso
Bkgd.
Stage I
Stage II


Input Image
Confidence Maps
Level 1
Head
Neck
L.Sho.
L.Elb.
L.Wri.
R.Sho.
R.Elb.
R.Wri.
L.Hip
L.Knee
L.Ank.
R.Hip
R.Knee
R.Ank.
Bkgd.
Stage I
Stage II
Stage III
Level 2
Head+Sho
L.Arm
R.Arm
Torso
L.Leg
R.Leg
Bkgd.
Stage I
Stage II
Stage III
Level 3
Torso
Bkgd.
Stage I
Stage II
Stage III

Temporal Sequence

(No temporal consistency enforced)

Predicted Poses

Stage I

Stage II

Stage III

Level 1

Head L.Elb. L.Wri. R.Elb. R.Wri. L.Knee L.Ank. R.Knee R.Ank. Bkgd.

Stage I

Stage II

Stage III

Level 2

Head+Sho L.Arm R.Arm Torso L.Leg R.Leg Bkgd.

Stage I

Stage II

Stage III

Level 3

Torso Bkgd.

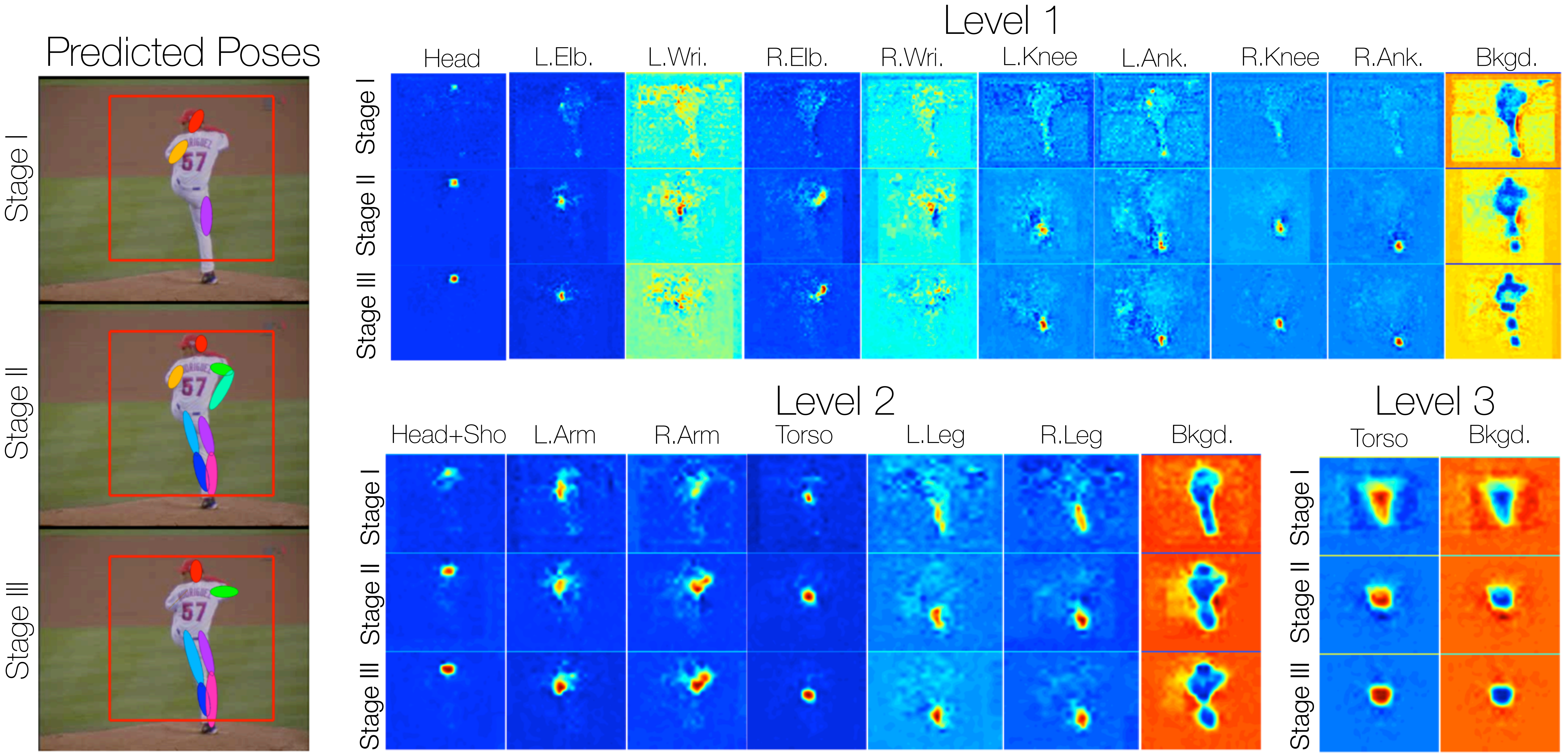
Stage I

Stage II

Stage III

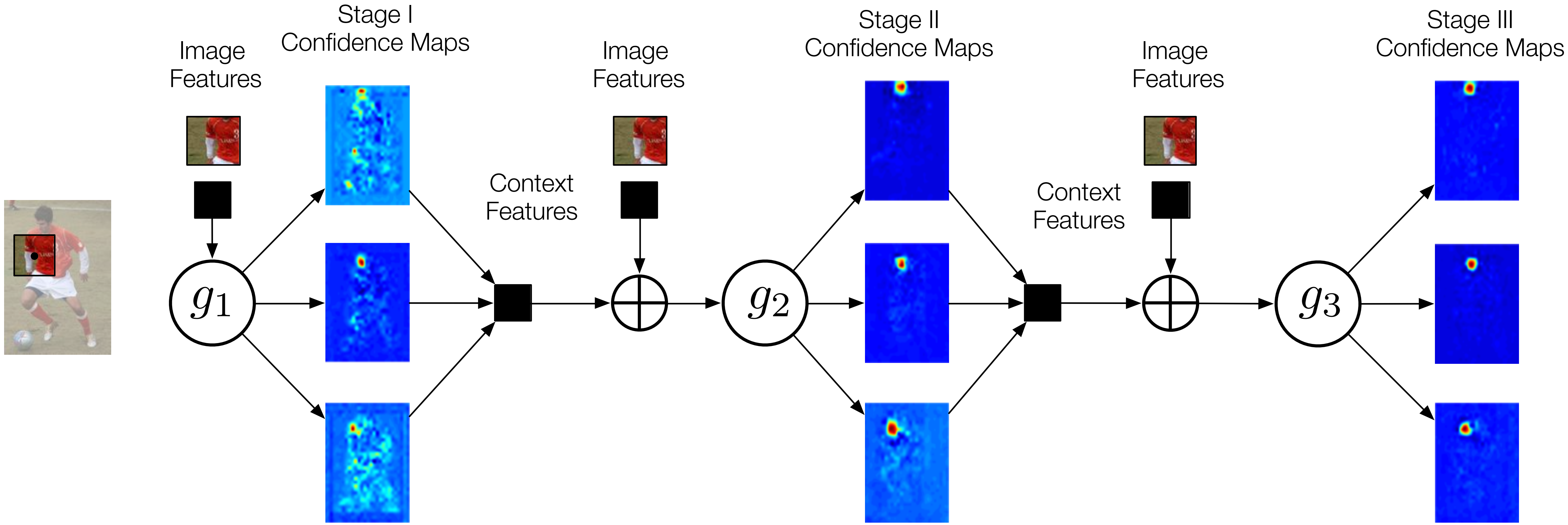
Temporal Sequence

(No temporal consistency enforced)



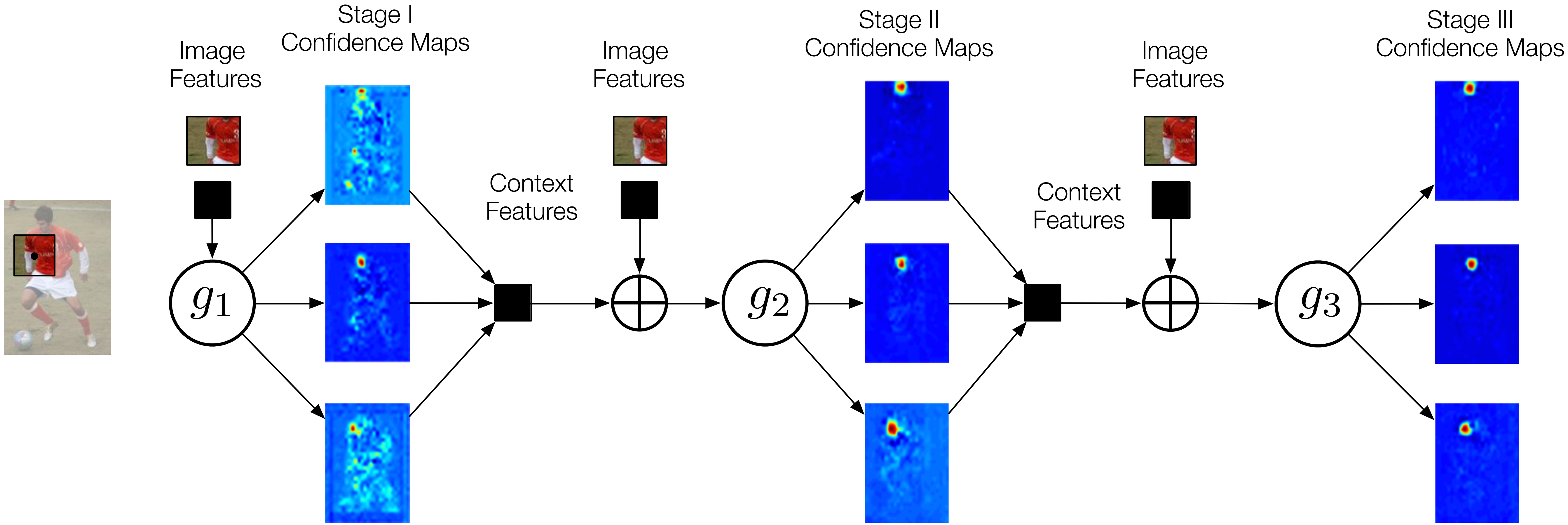
Pose Machines

Reduces structured prediction to a sequence of simple classification problems



Pose Machines

Reduces structured prediction to a sequence of simple classification problems



In Natural Language Processing

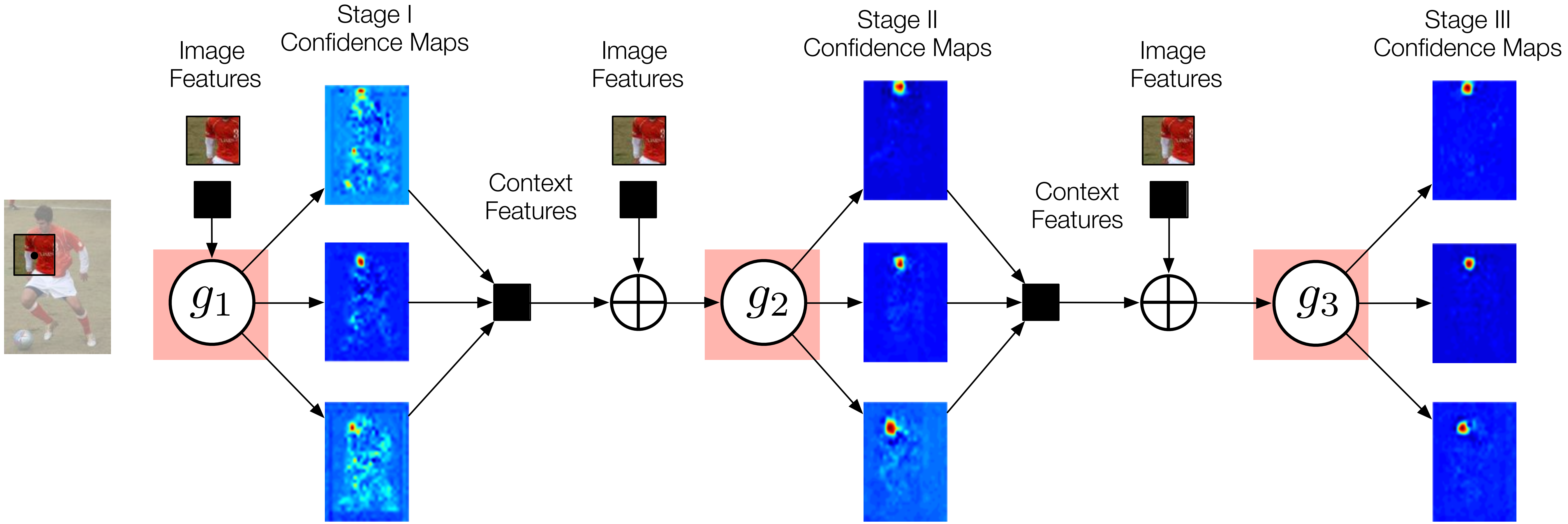
[Cohen and Carvalho, 2005] [Daume III et al., 2006]

In Computer Vision

[Kou et al., 2007] [Tu and Bai, 2008] [Munoz et al., 2010]

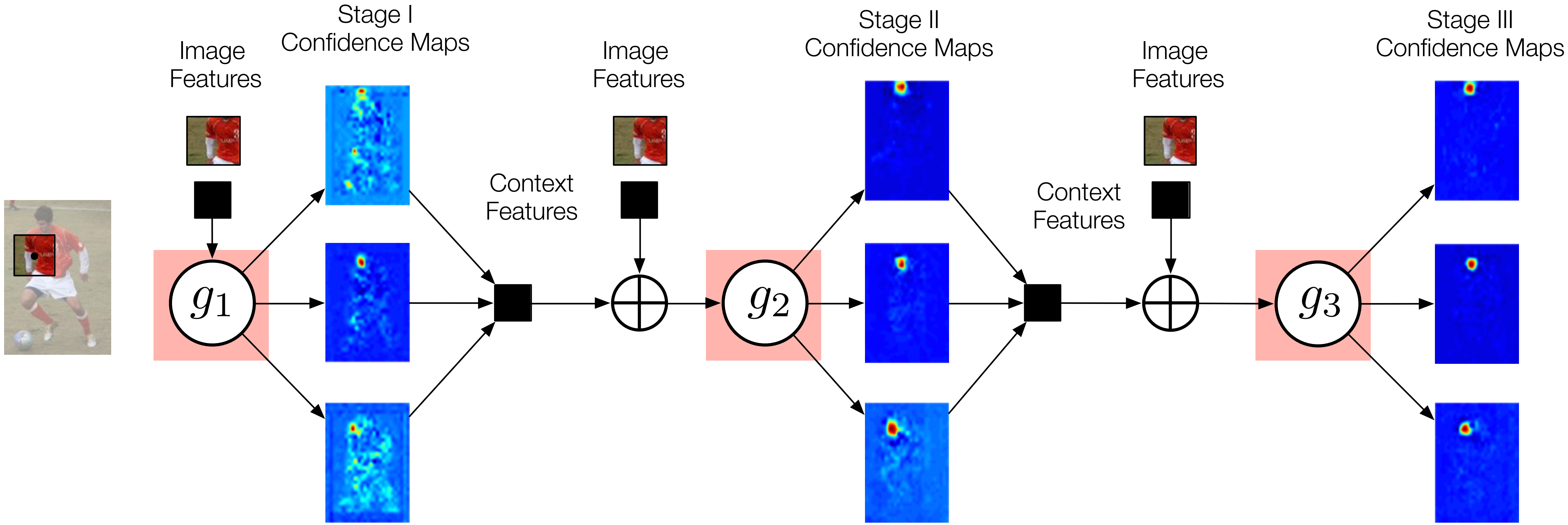
Pose Machines

Reduces structured prediction to a sequence of simple classification problems



Pose Machines

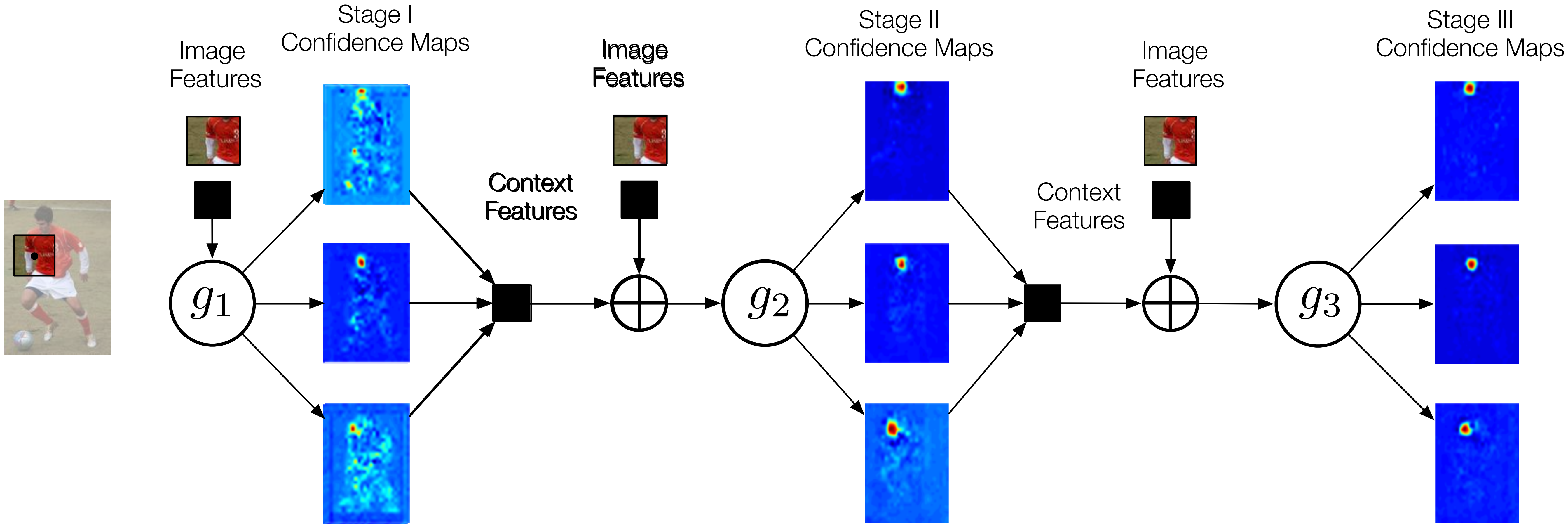
Reduces structured prediction to a sequence of simple classification problems



Training reduces to training multiple supervised classifiers

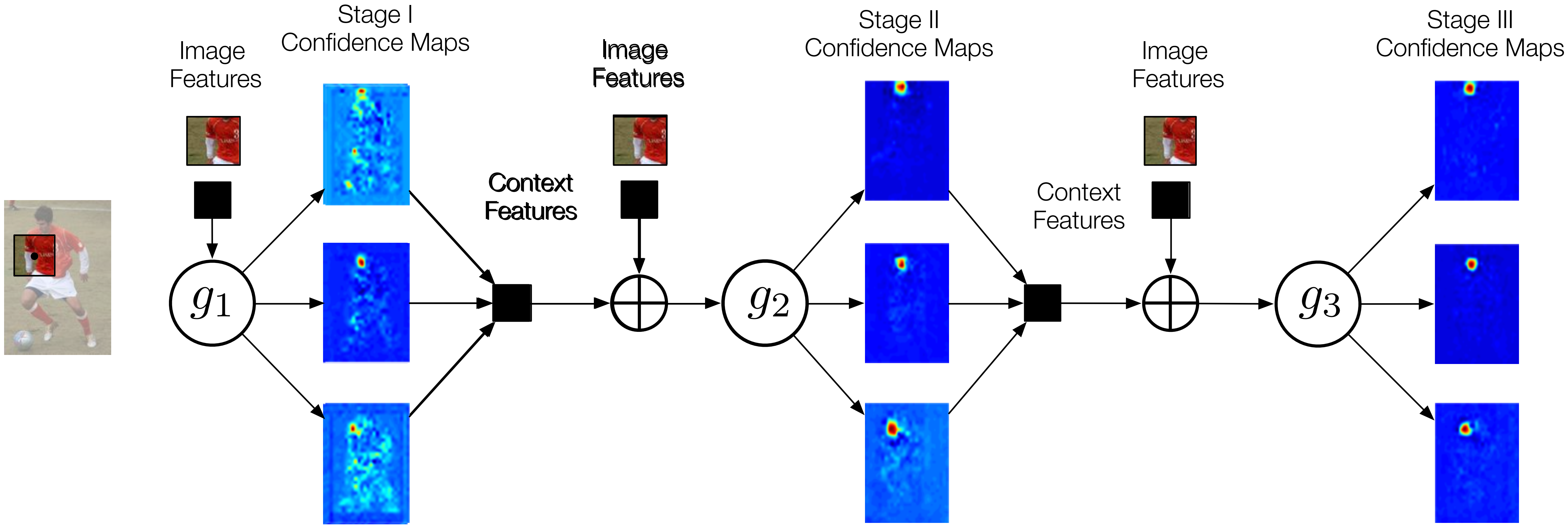
Pose Machines

Reduces structured prediction to a sequence of simple classification problems



Pose Machines

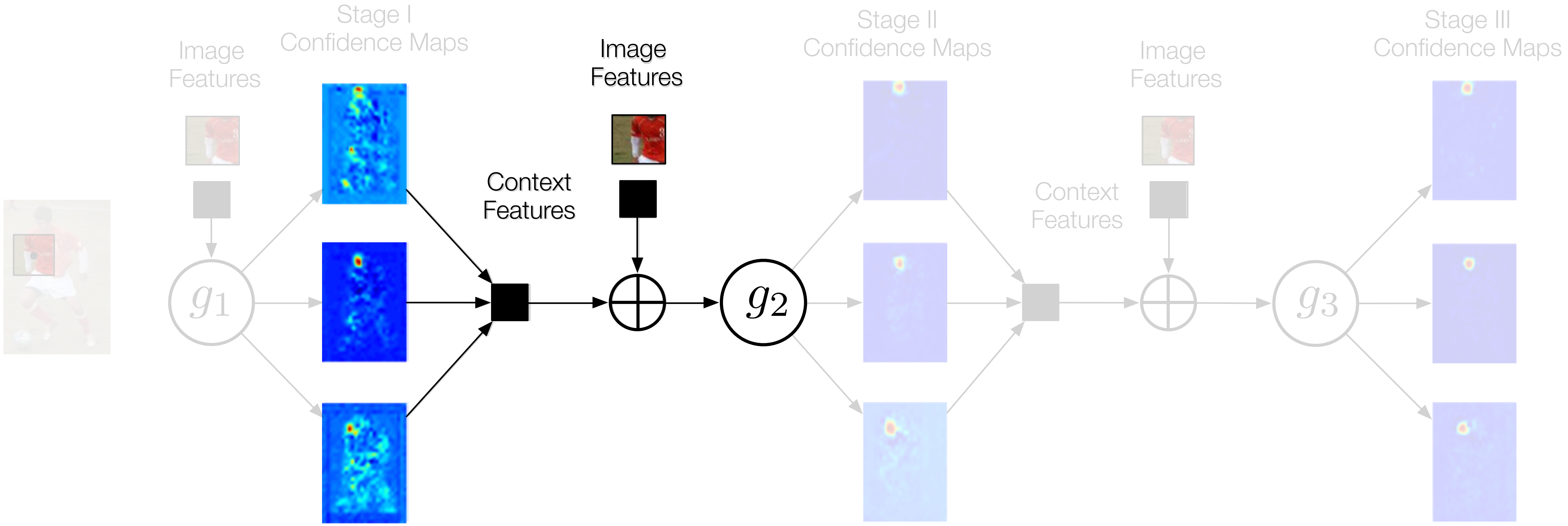
Reduces structured prediction to a sequence of simple classification problems



Spatial model is learned implicitly by the classifiers in a data-driven fashion

Pose Machines

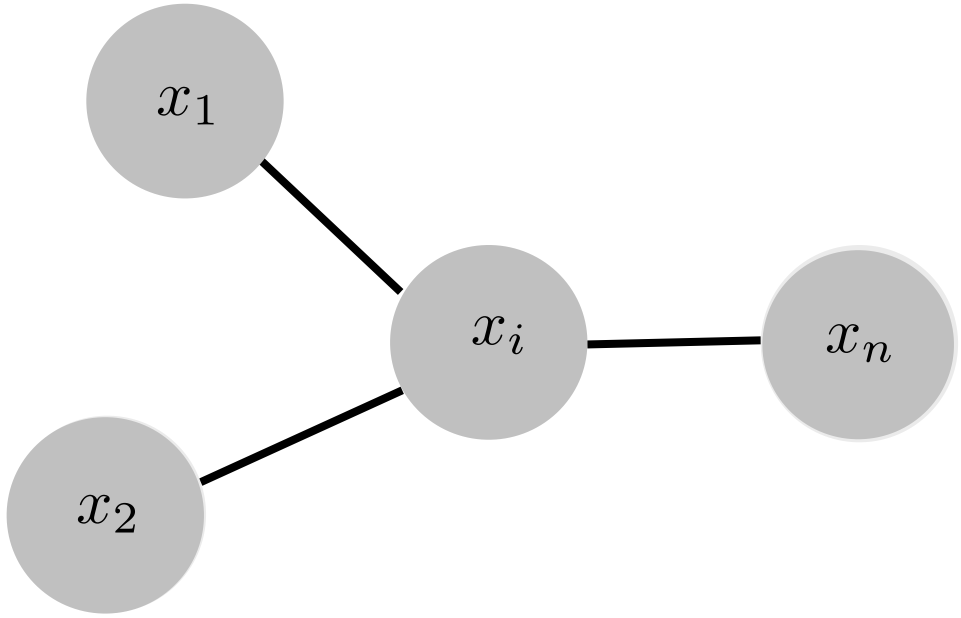
Reduces structured prediction to a sequence of simple classification problems



Spatial model is learned implicitly by the classifiers in a data-driven fashion

Inference Machines for Pose Estimation

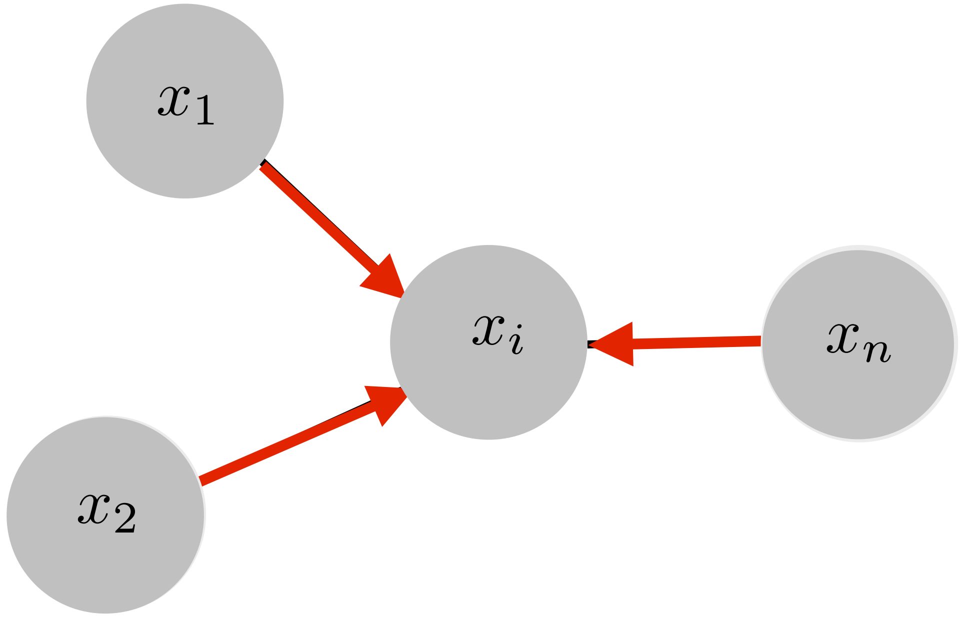
Unrolling message passing inference in graphical models



[Munoz et al., *ECCV 2010,* Ross et al., *CVPR 2011*]

Inference Machines for Pose Estimation

Unrolling message passing inference in graphical models

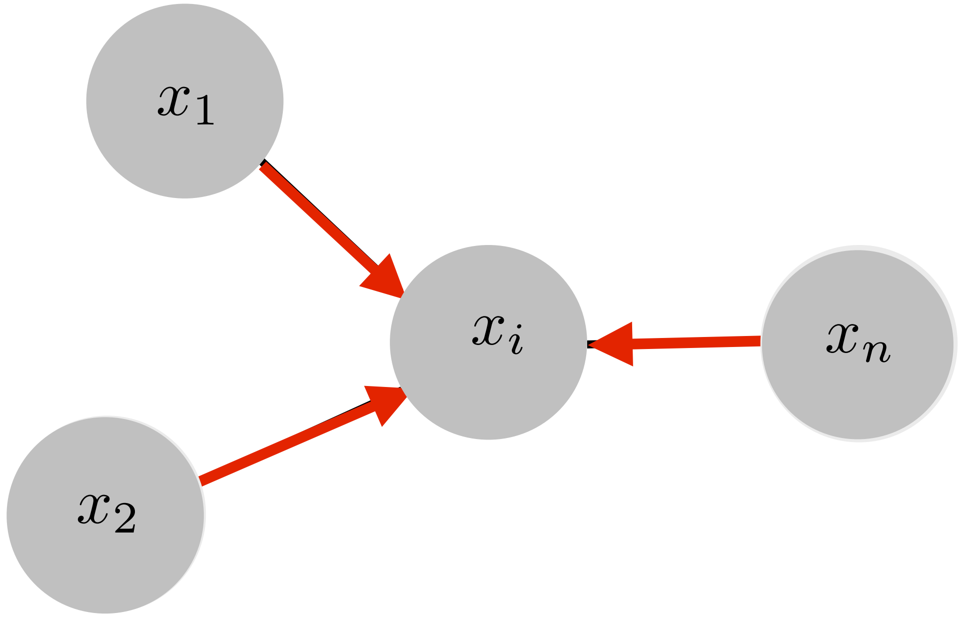


$$b(x_i) \propto \prod_{j \in \mathcal{N}_i} m_{j \to i}(x_i)$$

[Munoz et al., *ECCV 2010,* Ross et al., *CVPR 2011*]

Inference Machines for Pose Estimation

Unrolling message passing inference in graphical models



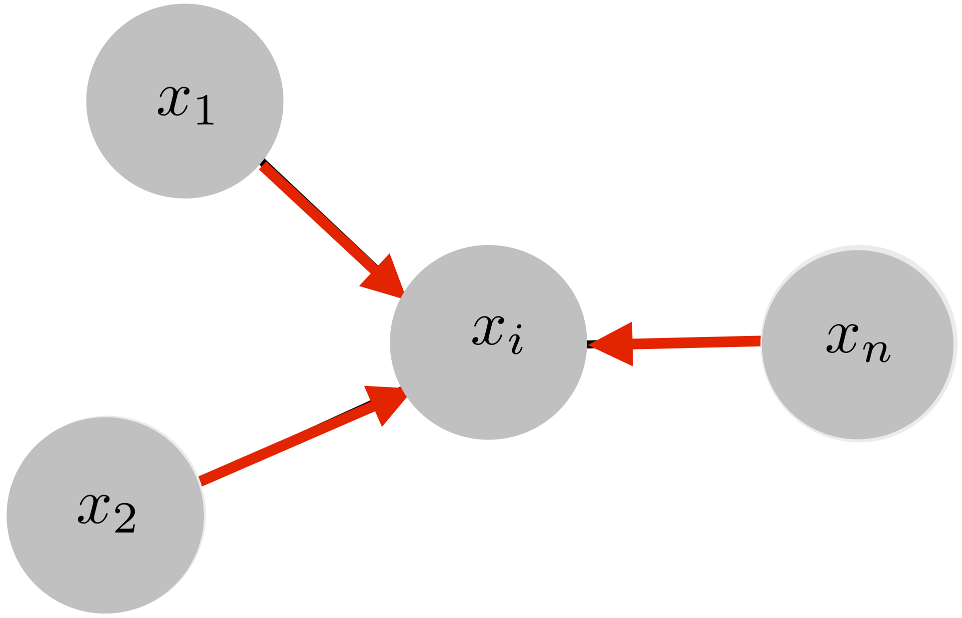
$$b(x_i) \propto \prod_{j \in \mathcal{N}_i} m_{j \to i}(x_i)$$

Message passing in graphical model inference can be thought of as sequential prediction

[Munoz et al., *ECCV 2010*, Ross et al., *CVPR 2011*]

Inference Machines for Pose Estimation

Unrolling message passing inference in graphical models



Replace product with classifier

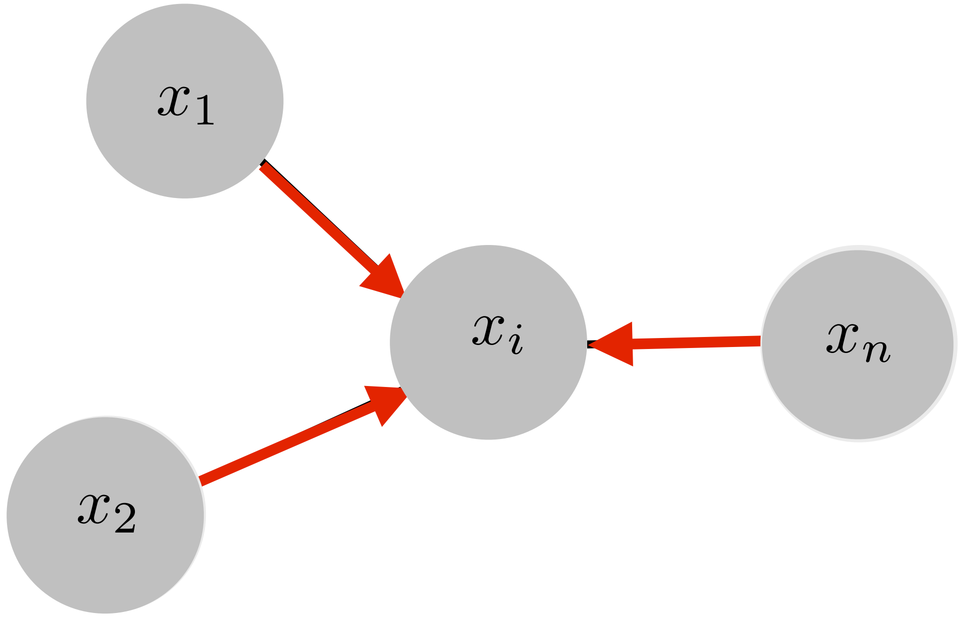
$$b(x_i) \propto g \quad m_{j \to i}(x_i)$$

Message passing in graphical model inference can be thought of as sequential prediction

[Munoz et al., *ECCV 2010*, Ross et al., *CVPR 2011*]

Inference Machines for Pose Estimation

Unrolling message passing inference in graphical models



Replace product with classifier

$$b(x_i) \propto g(\{\psi_j(x_i)\}_{j \in \mathcal{N}_i})$$

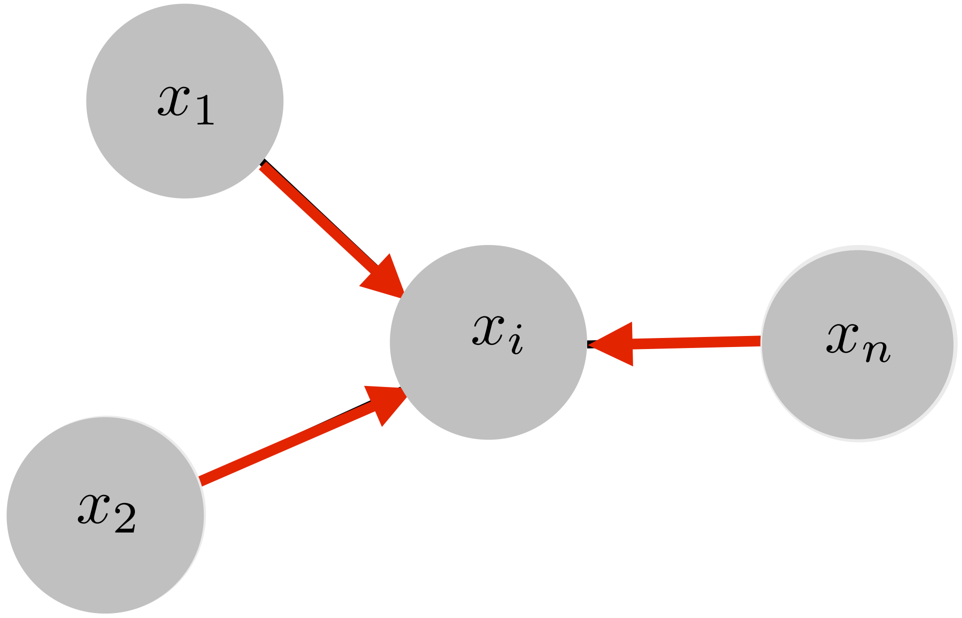
Messages consist of context feature computations

Message passing in graphical model inference can be thought of as sequential prediction

[Munoz et al., *ECCV 2010*, Ross et al., *CVPR 2011*]

Inference Machines for Pose Estimation

Unrolling message passing inference in graphical models



Replace product with classifier

$$b(x_i) \propto g(\{\psi_j(x_i)\}_{j \in \mathcal{N}_i})$$

Messages consist of context feature computations

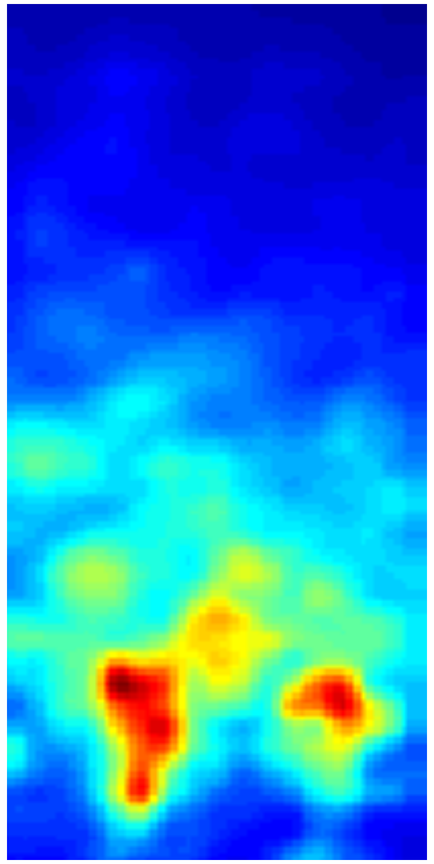
Models a fully connected graph. Information from parts in all levels are used for prediction

[Munoz et al., *ECCV 2010*, Ross et al., *CVPR 2011*]

Double Counting

Input Image

Estimated Pose



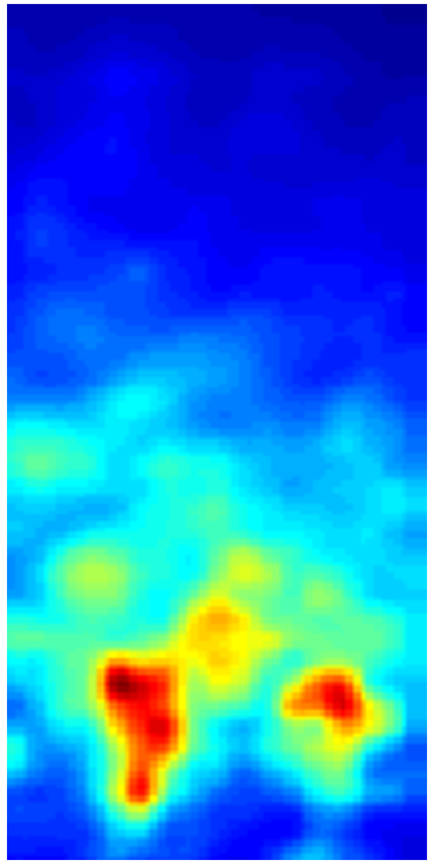
Max Marginal
(left ankle)

Tree Structured Model
[Yang and Ramanan, 2011]

Double Counting

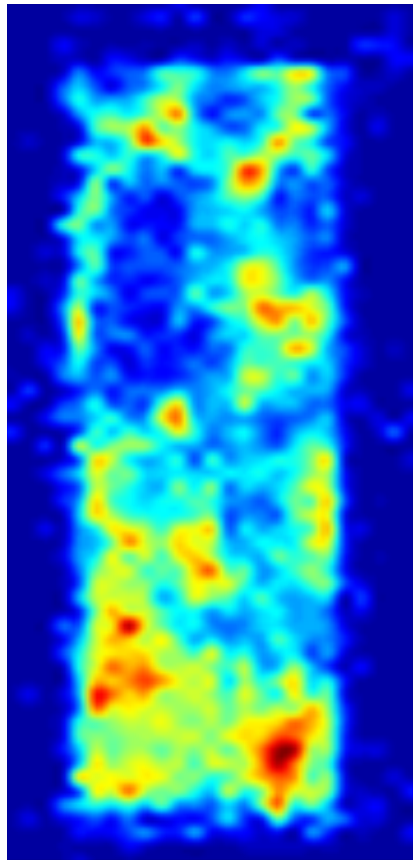
Input Image

Estimated Pose



Max Marginal
(left ankle)

Estimated Pose



Stage I
Confidence

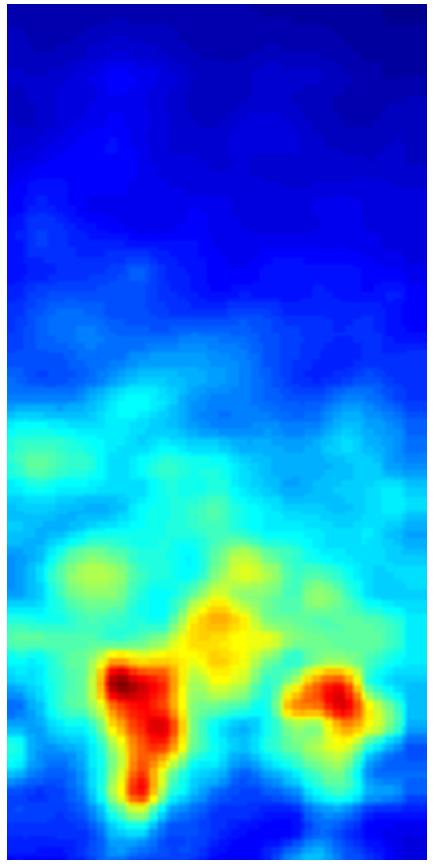
Tree Structured Model
[Yang and Ramanan, 2011]

Pose Machines

Double Counting

Input Image

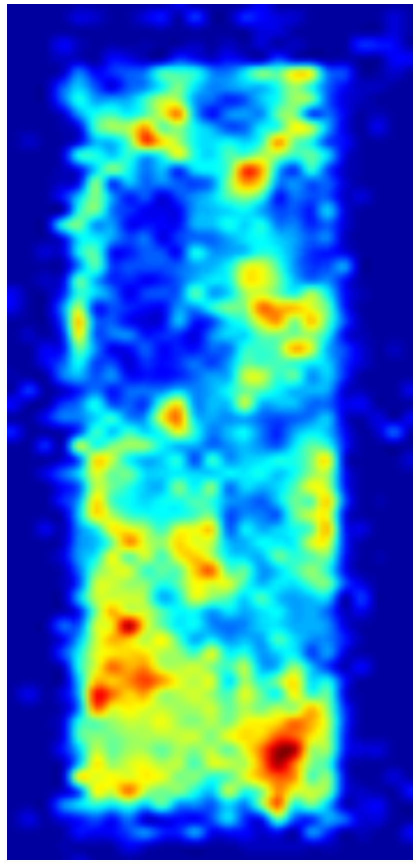
Estimated Pose



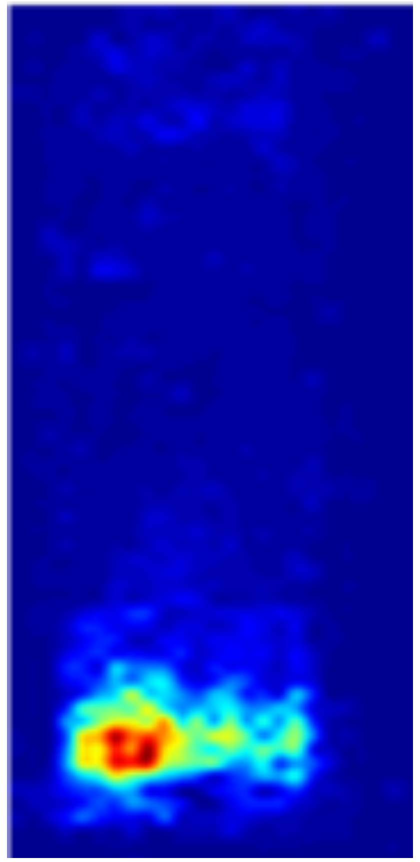
Max Marginal
(left ankle)

Tree Structured Model
[Yang and Ramanan, 2011]

Estimated Pose



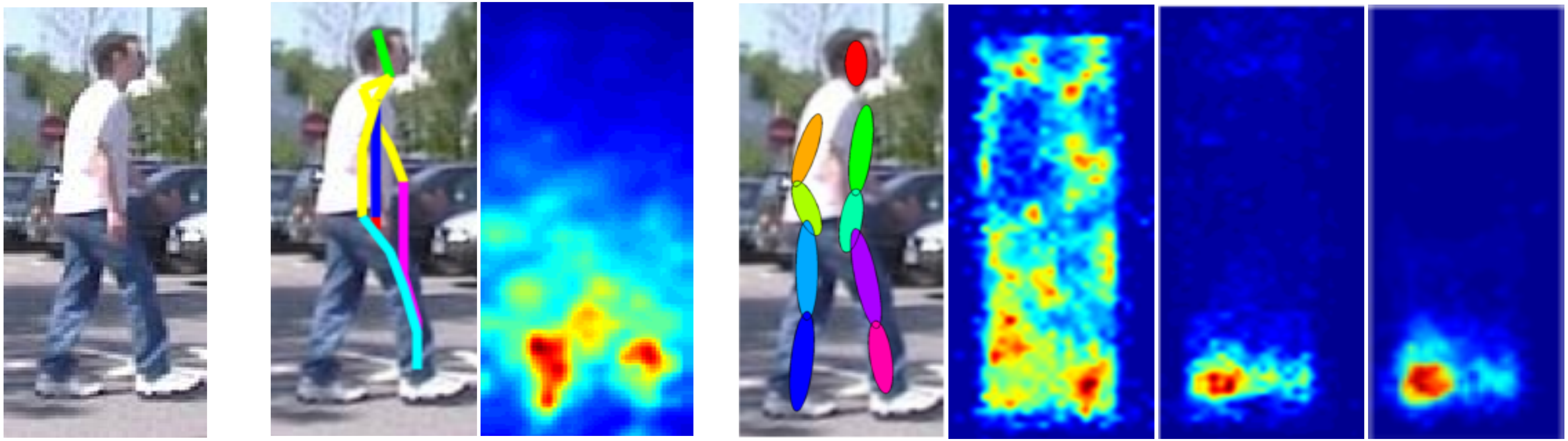
Stage I
Confidence



Stage II
Confidence

Pose Machines

Double Counting



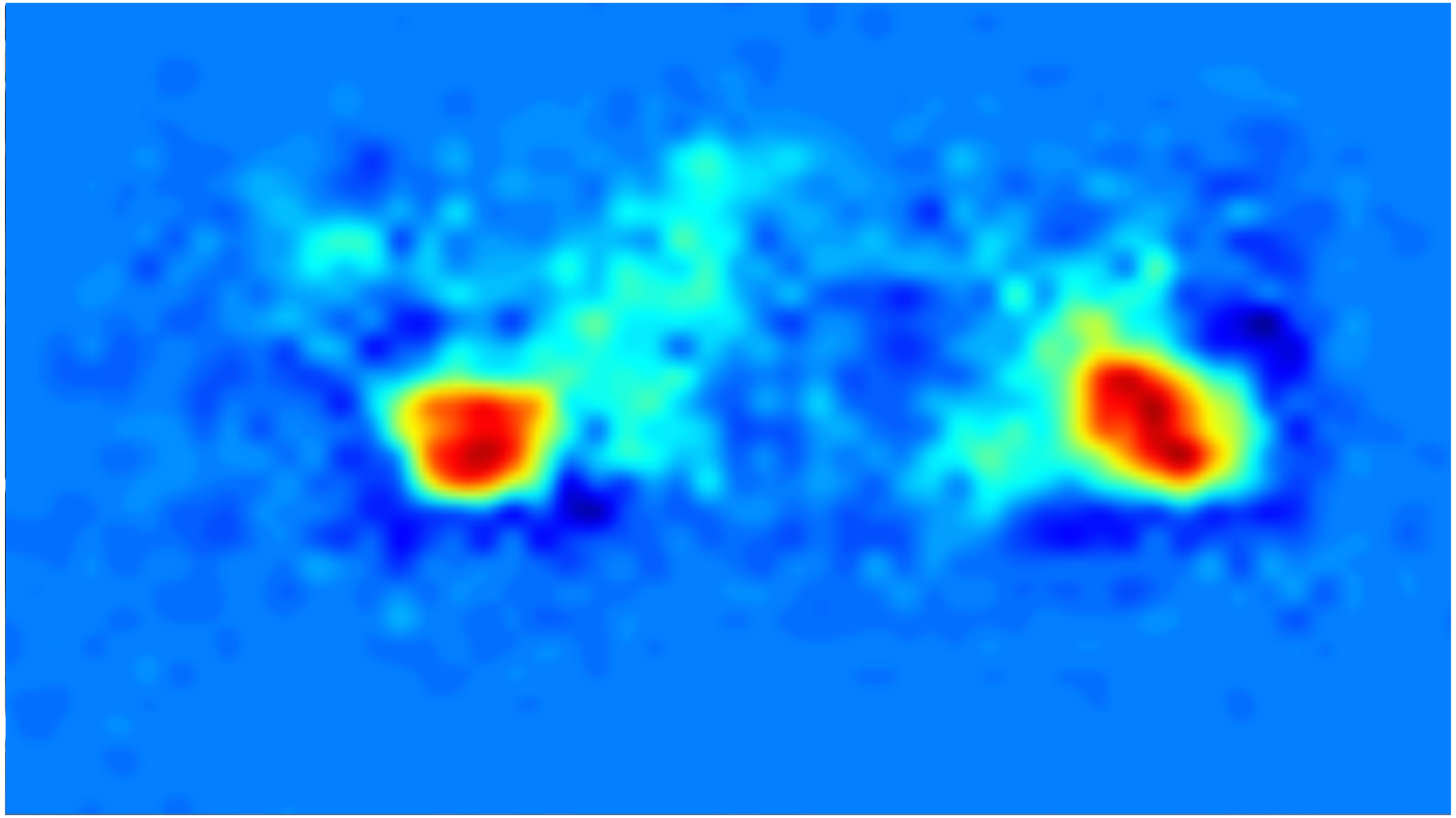
Input Image | Estimated Pose | Max Marginal (left ankle) | Estimated Pose | Stage I Confidence | Stage II Confidence | Stage III Confidence

Tree Structured Model
[Yang and Ramanan, 2011]

Pose Machines

Detection + Pose Estimation

Detection + Pose Estimation

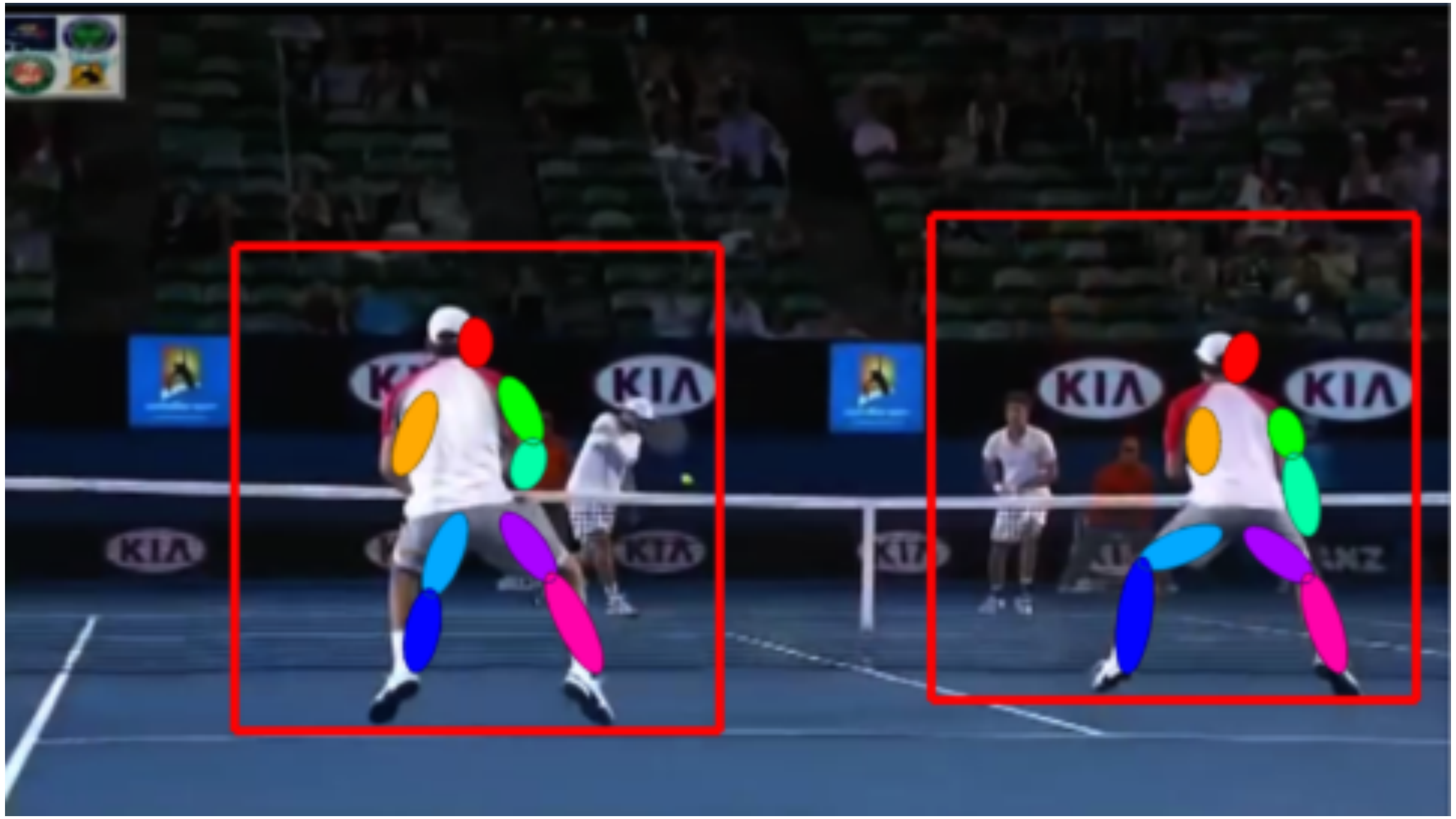


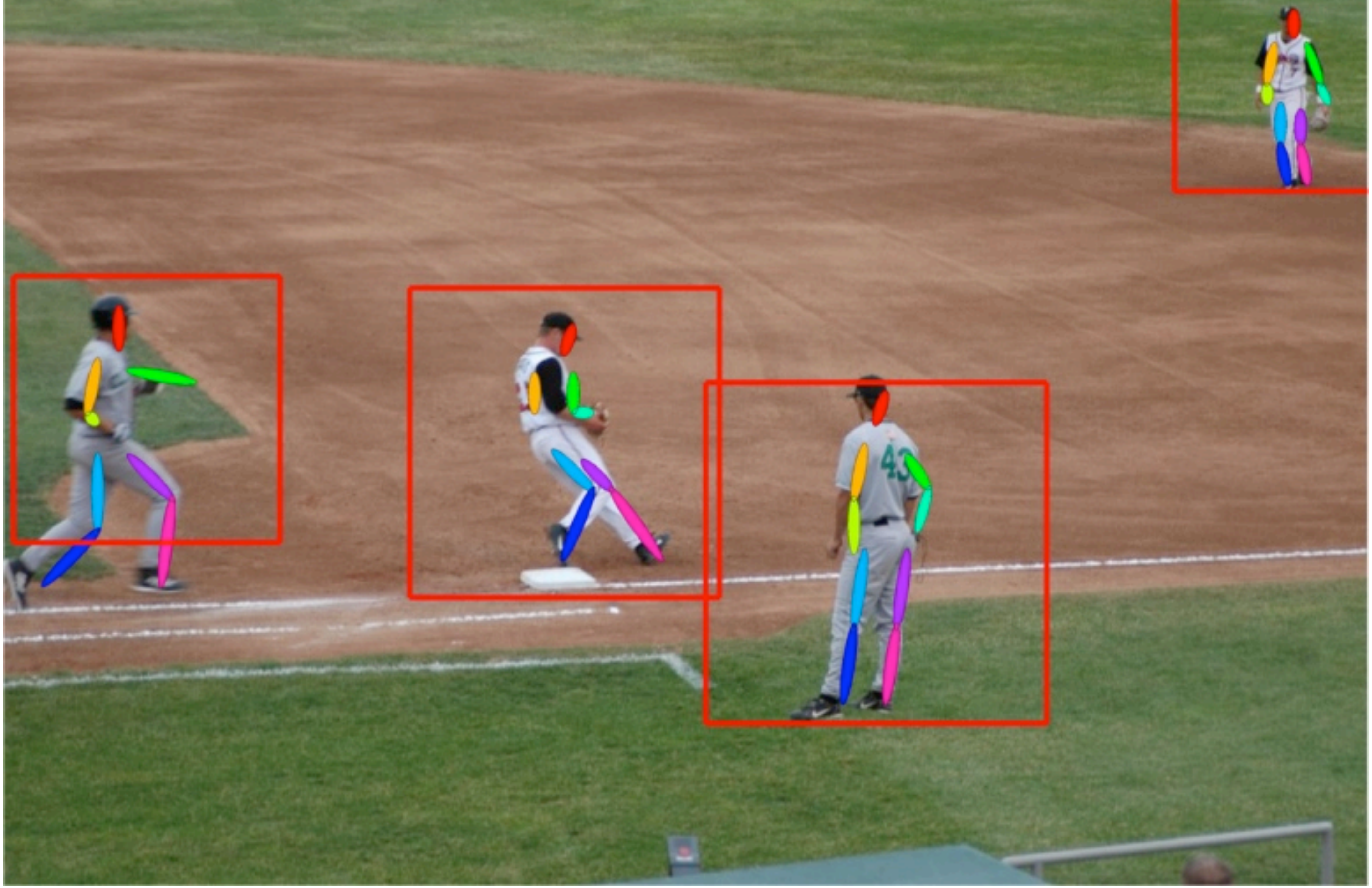
Confidence from Detection Level

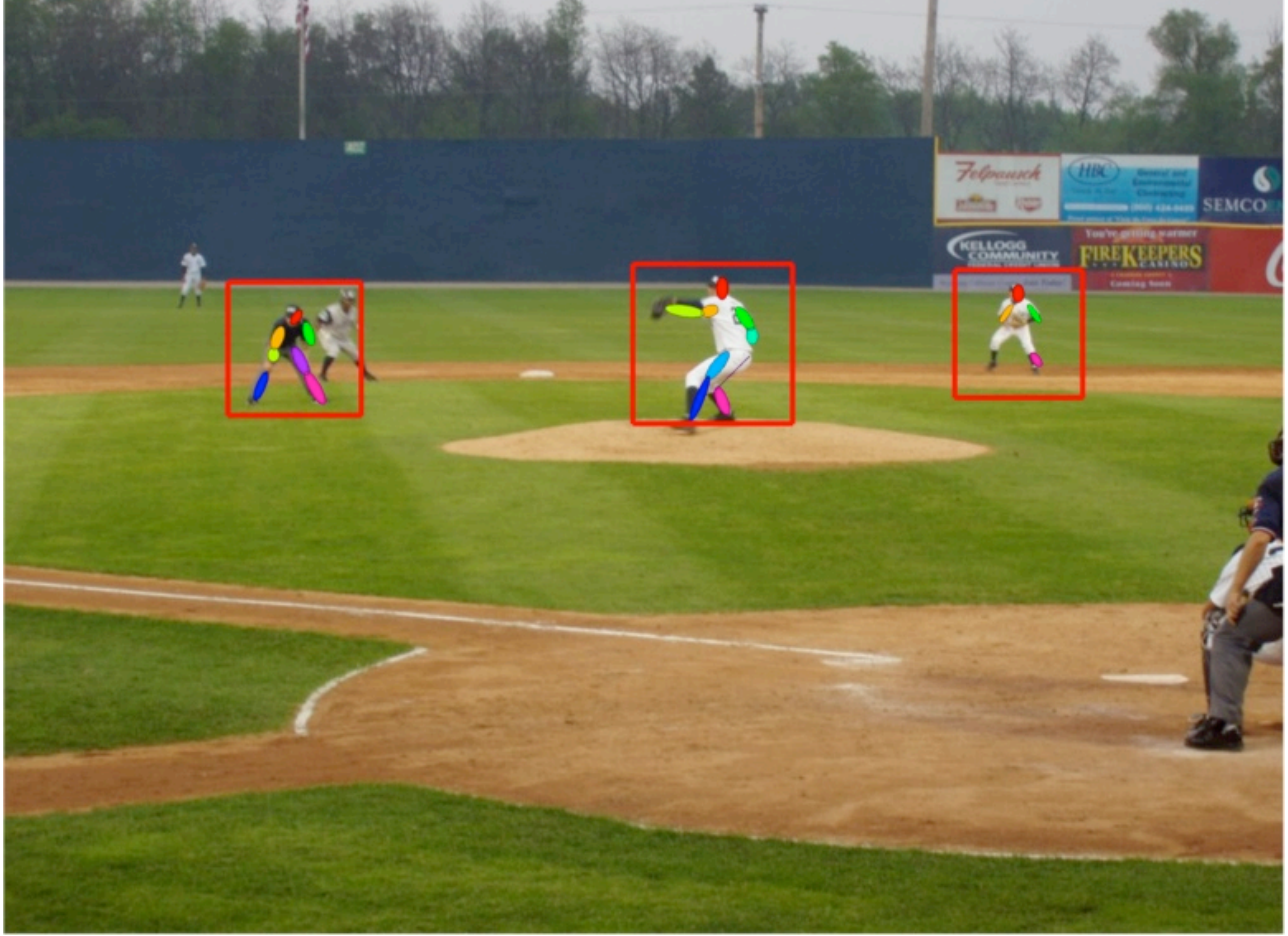
Detection + Pose Estimation

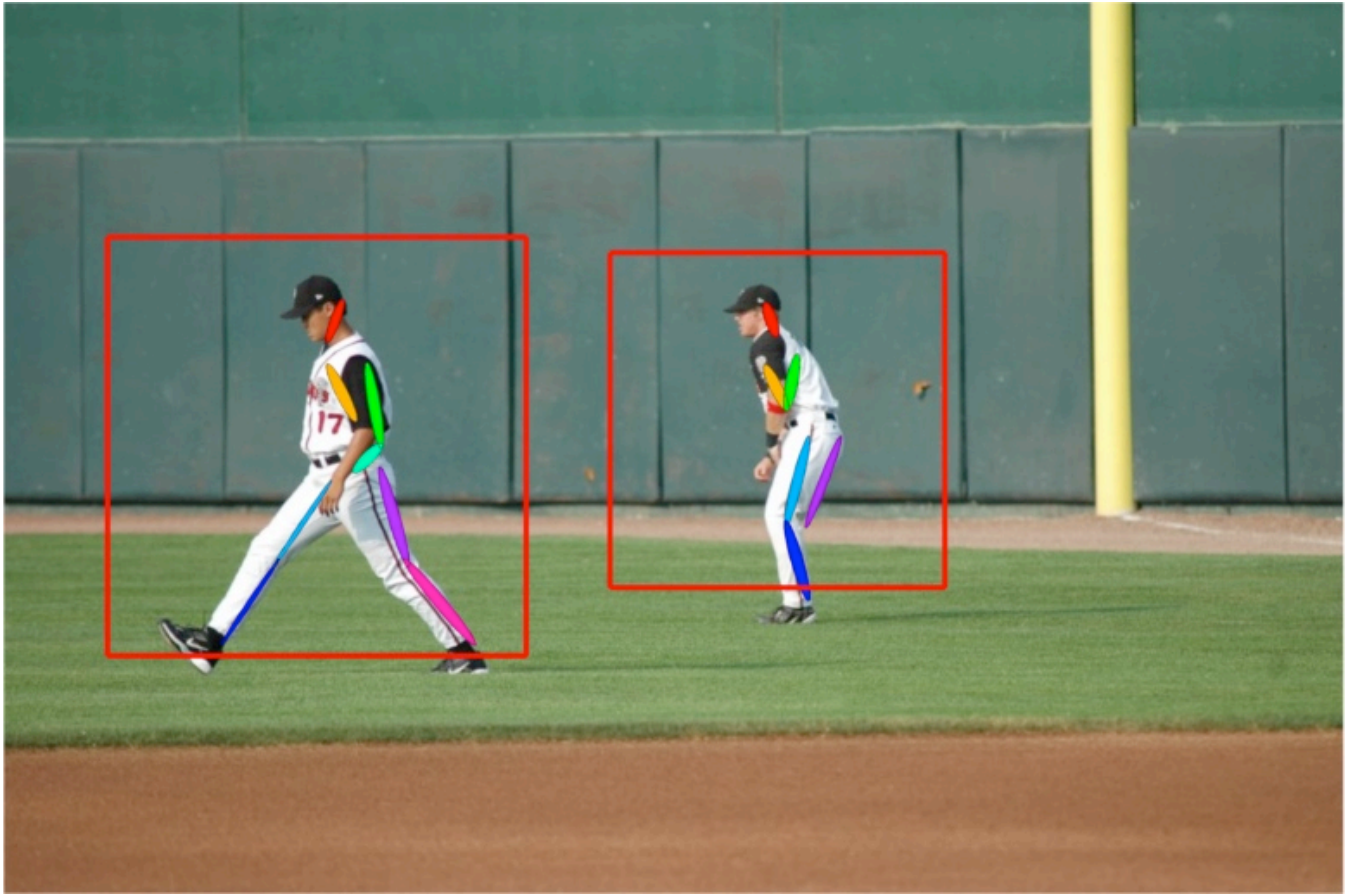
Confidence from Detection Level

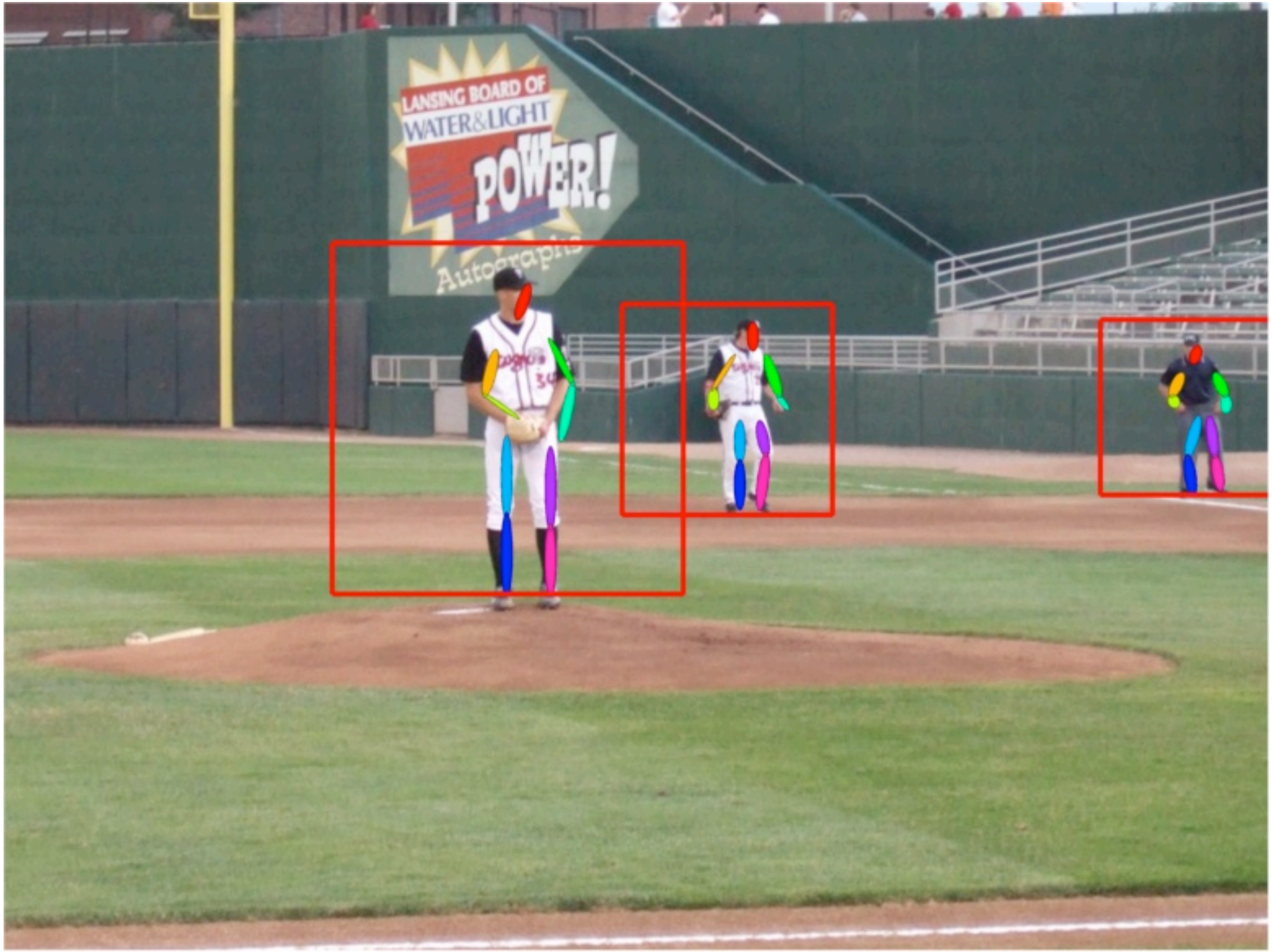
Detection + Pose Estimation

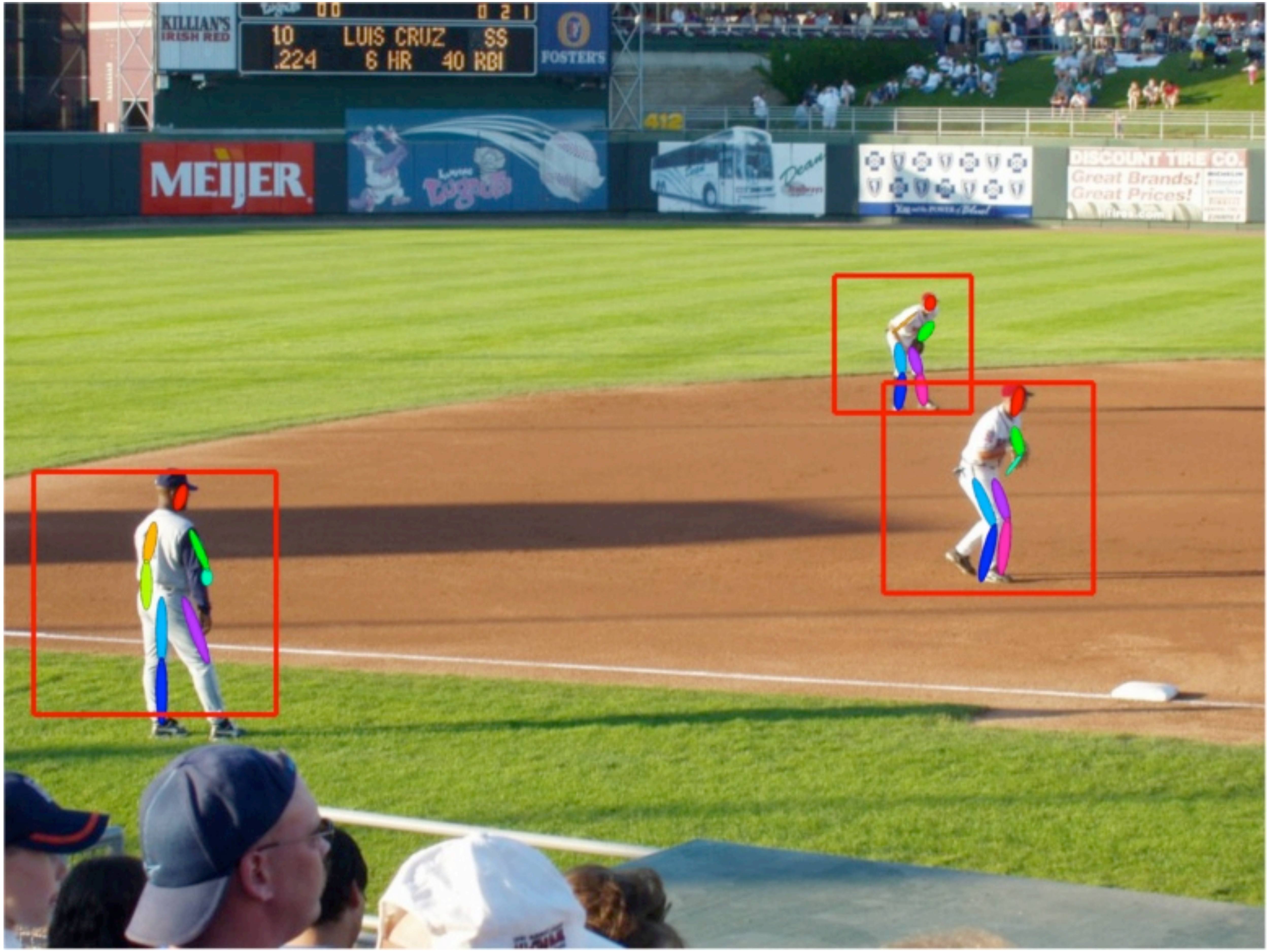





Felpausch
HBC
SEMCO
KELLOGG COMMUNITY
You're getting warmer
FIREKEEPERS CASINO
Coming Soon


17


LANSING BOARD OF
WATER&LIGHT
POWER!
Autographs


KILLIAN'S IRISH RED
10 LUIS CRUZ SS
.224 6 HR 40 RBI
FOSTER'S
MEIJER
412
DISCOUNT TIRE CO.
Great Brands!
Great Prices!

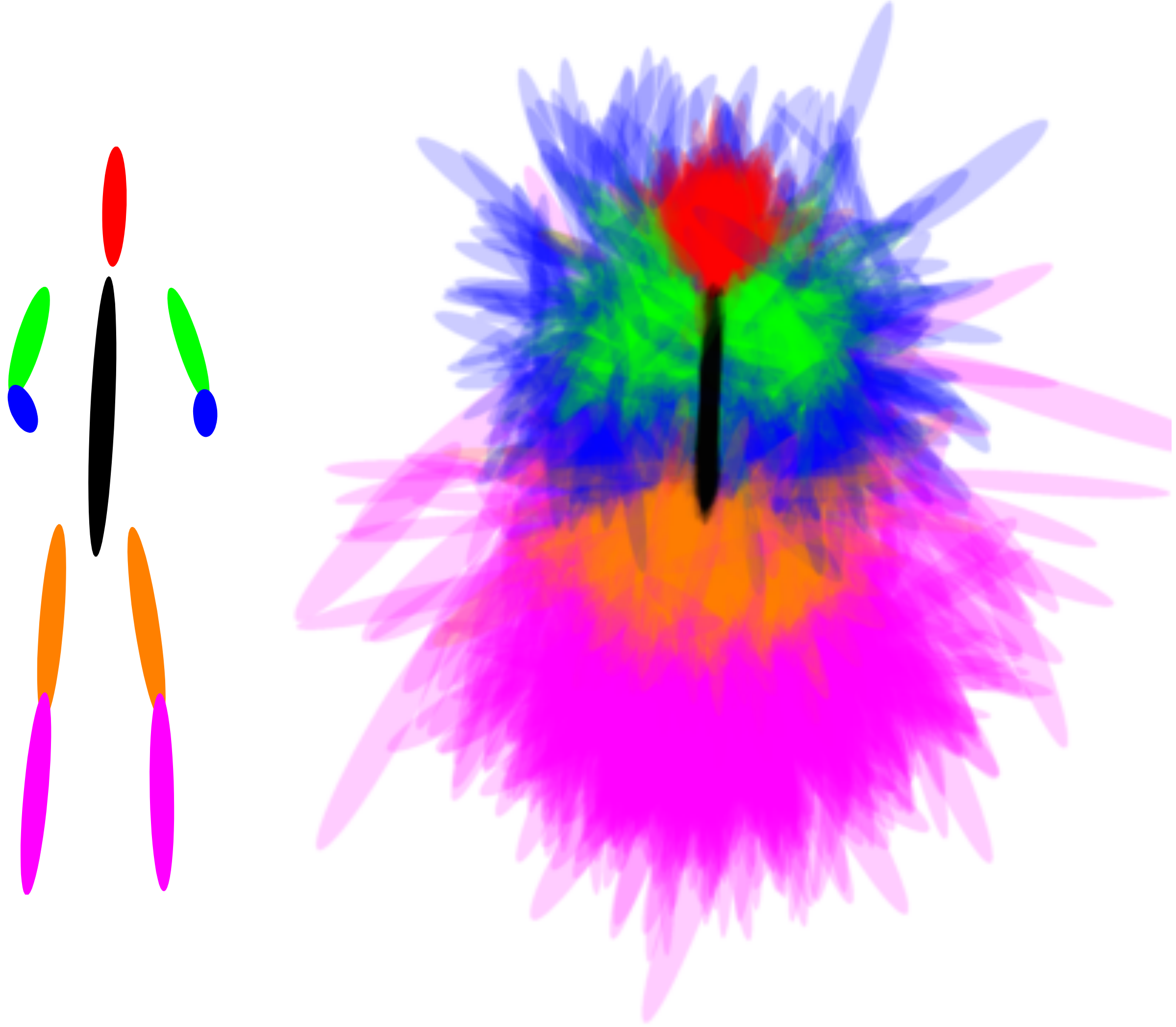
Evaluation: Datasets

LEEDS Sports Dataset

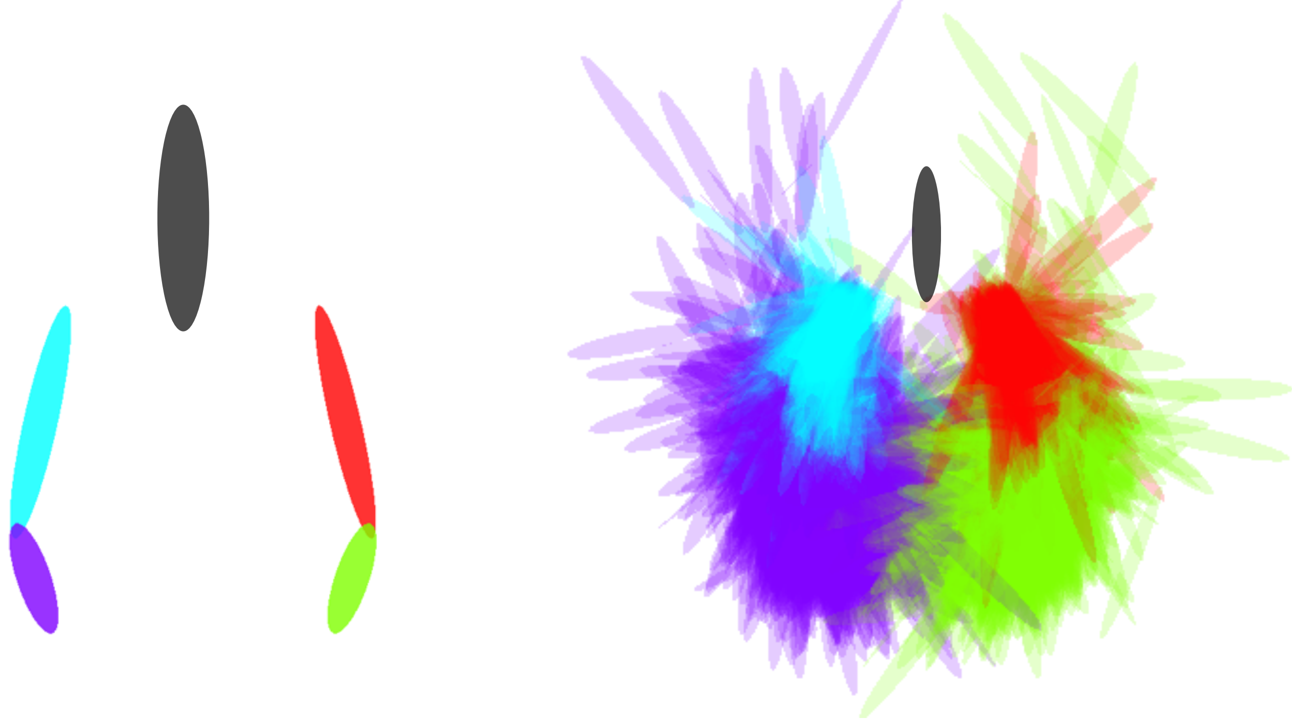
FLIC Dataset

Evaluation: Datasets

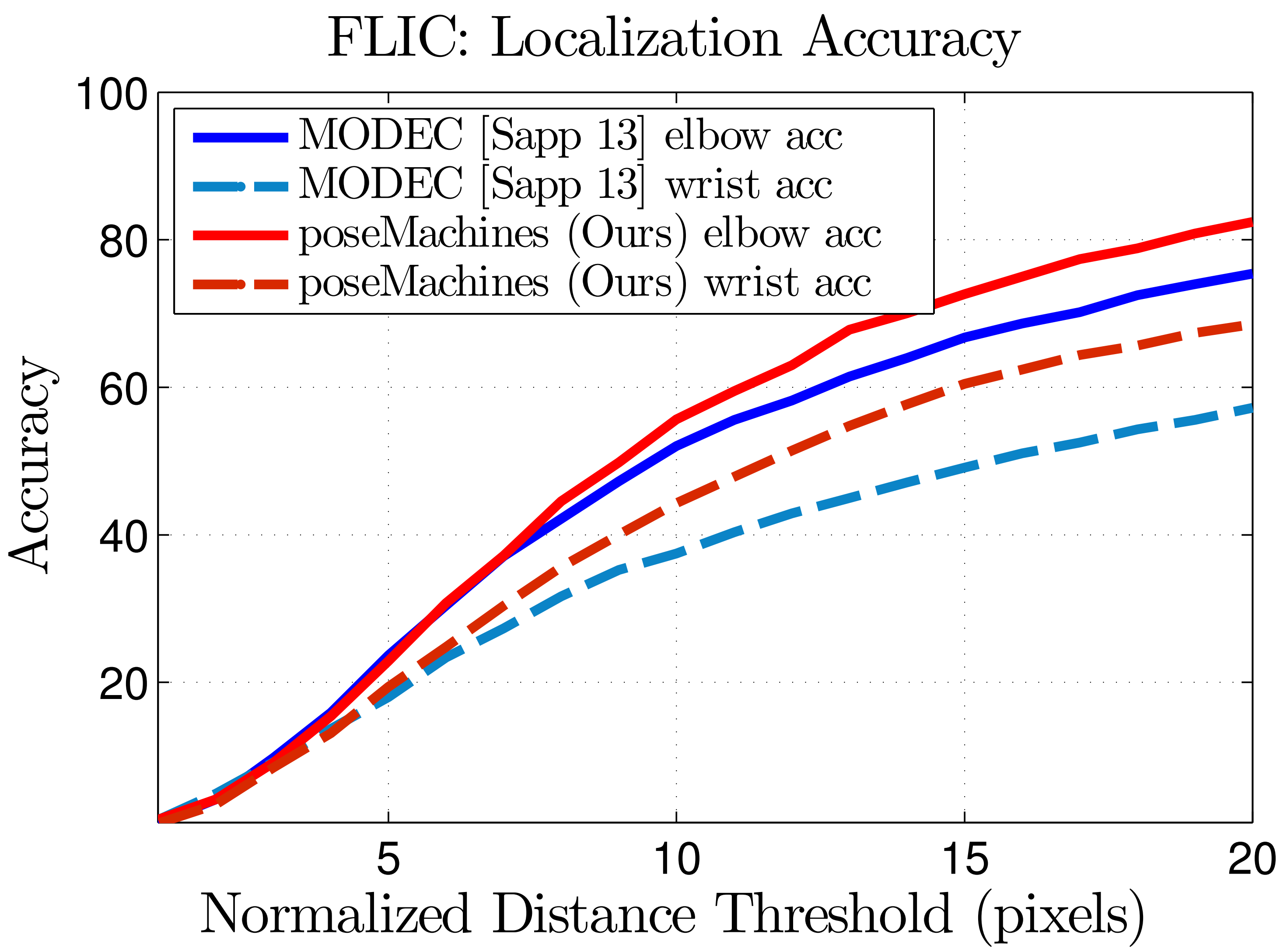
LEEDS Sports Dataset

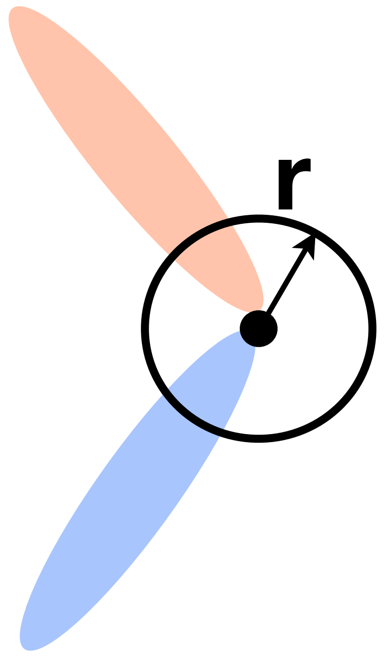


FLIC Dataset



Evaluation: FLIC





Percentage Detected Joints

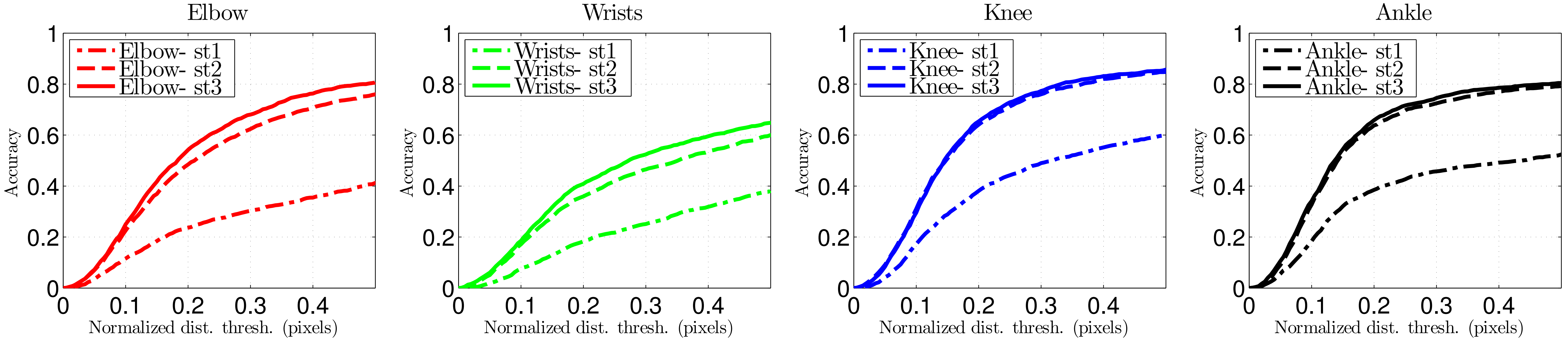
Evaluation: FLIC

Evaluation: LEEDS

Analysis

Performance variation with number of stages

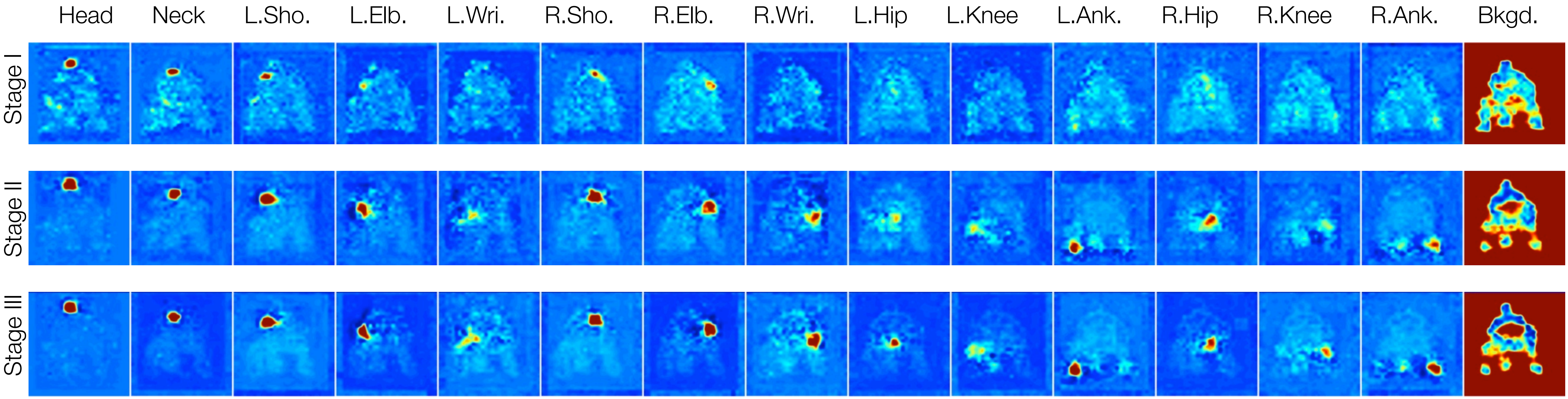
LEEDS



Ablative Spatial Analysis

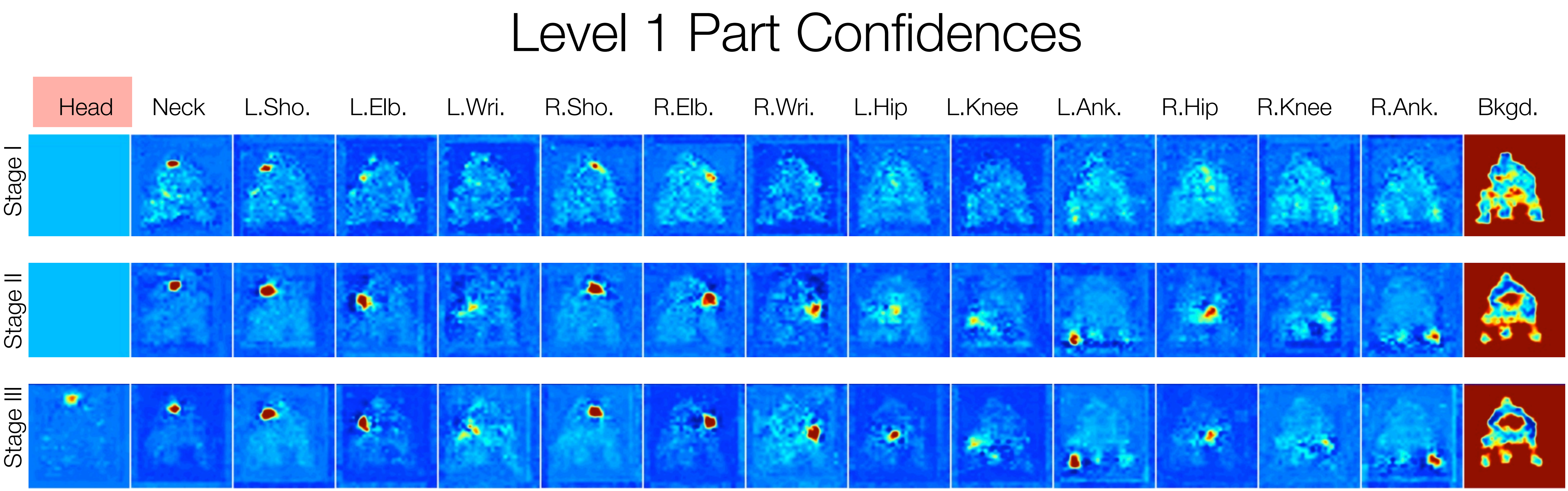
Level 1 Part Confidences

Predicted Pose



Ablative Spatial Analysis

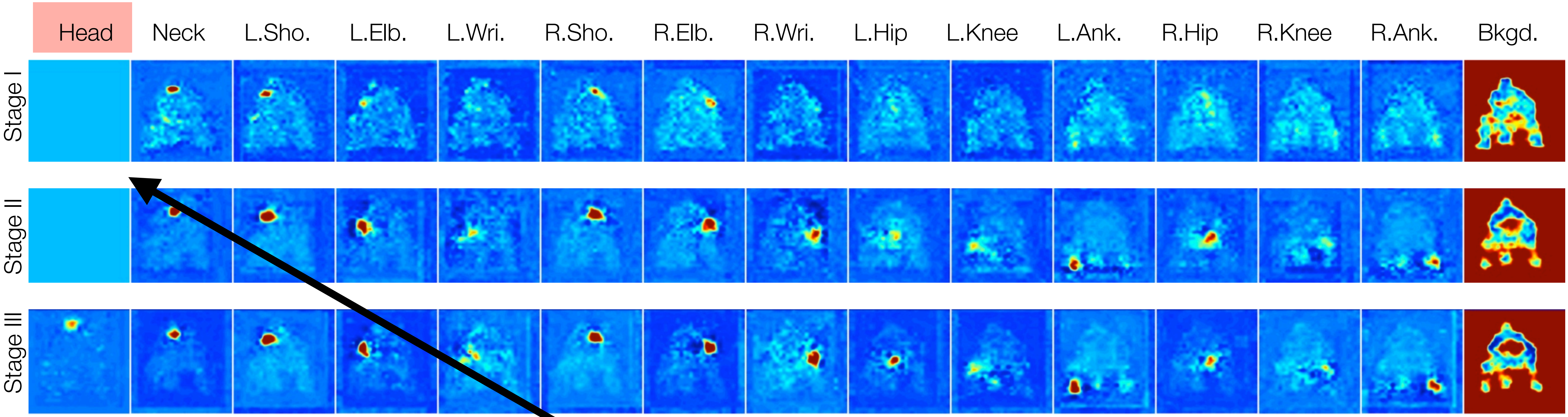
Predicted Pose



Ablative Spatial Analysis

Predicted Pose

Level 1 Part Confidences

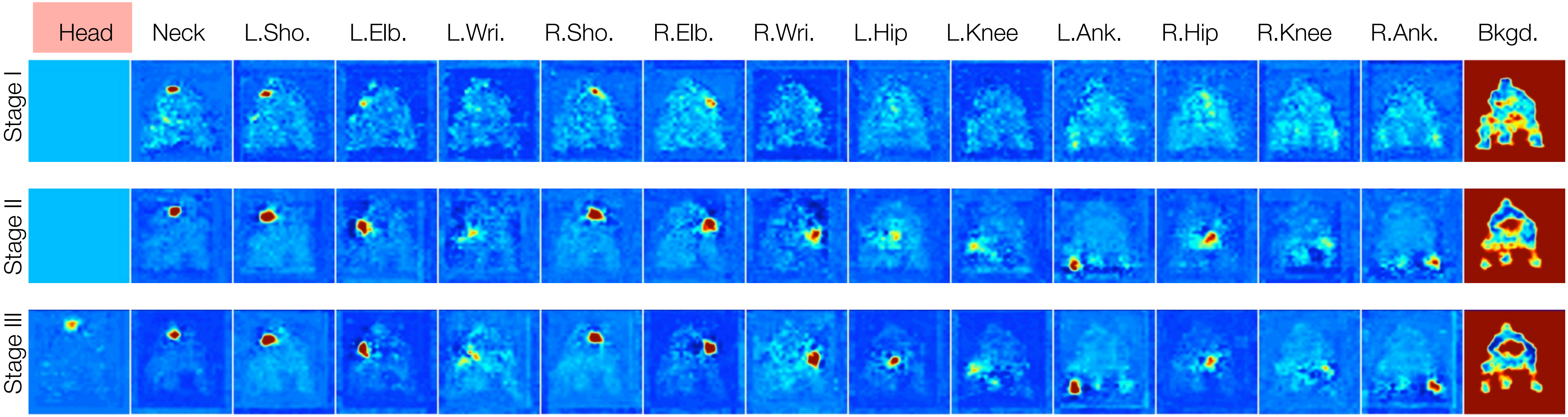


Context from the confidence map of *head* is removed

Ablative Spatial Analysis

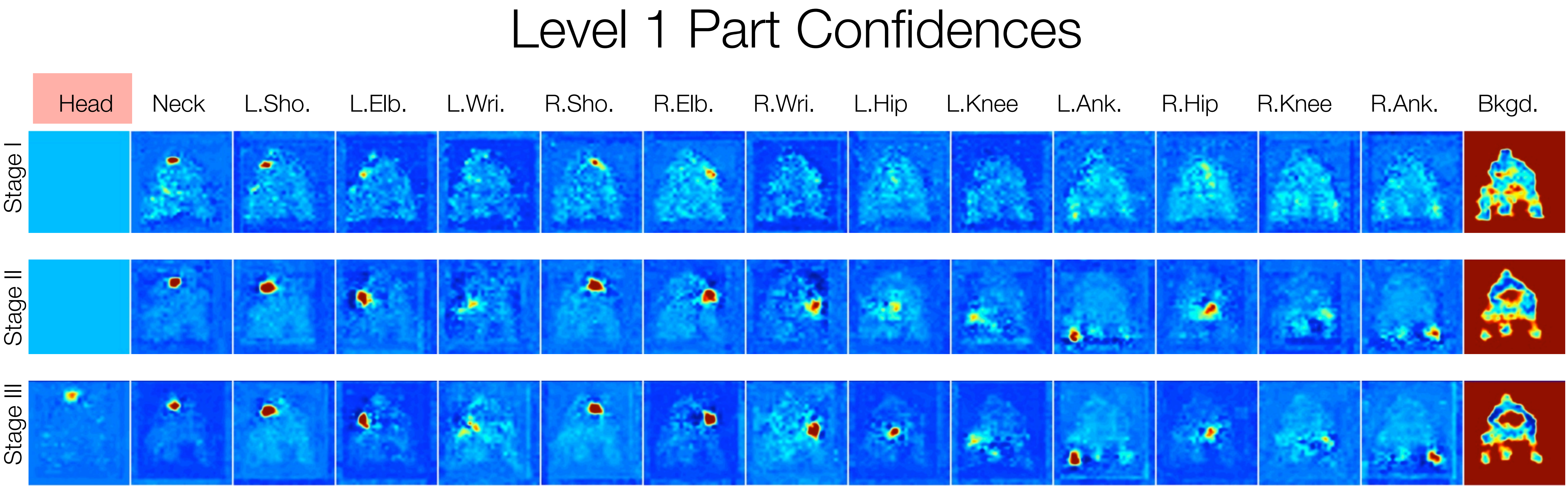
Level 1 Part Confidences

Predicted Pose



Ablative Spatial Analysis

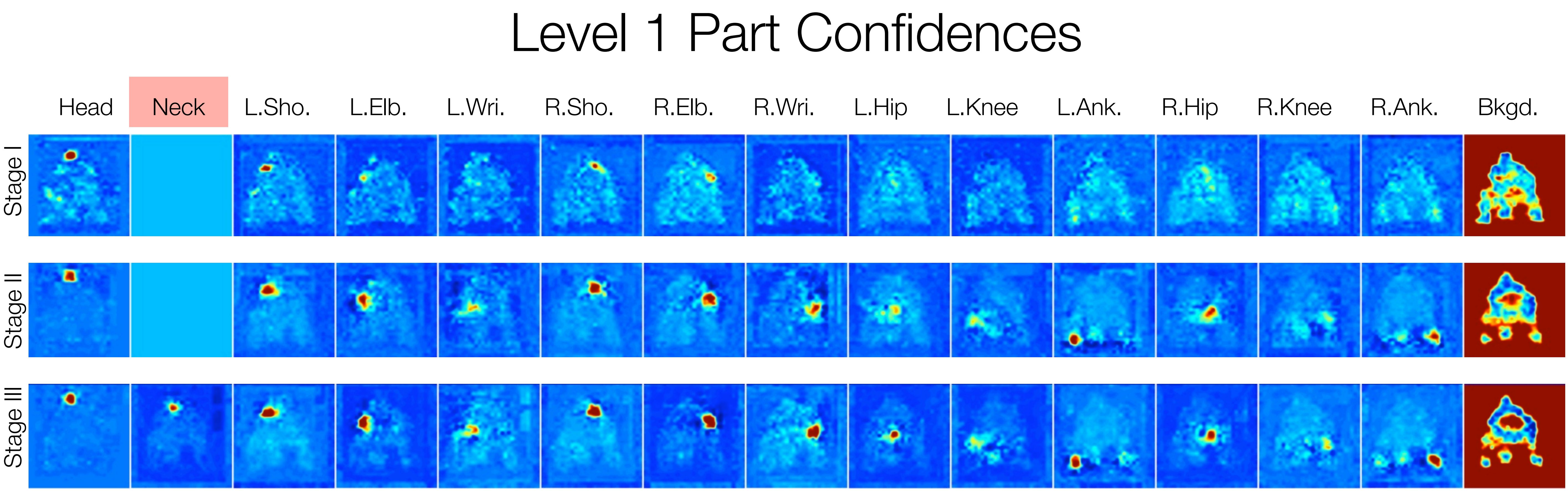
Predicted Pose



Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

Predicted Pose

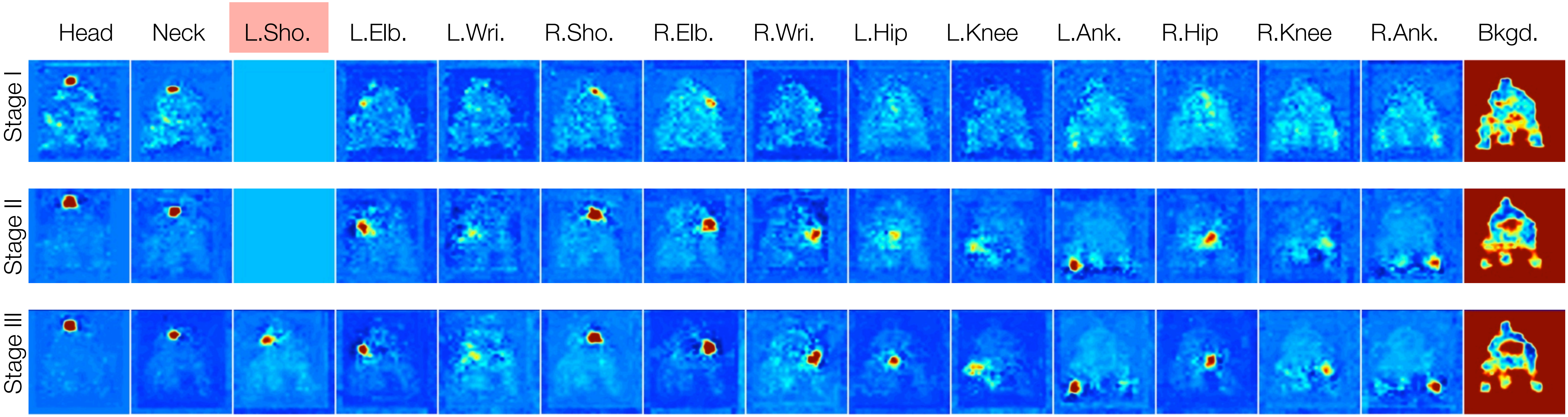


Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

Predicted Pose

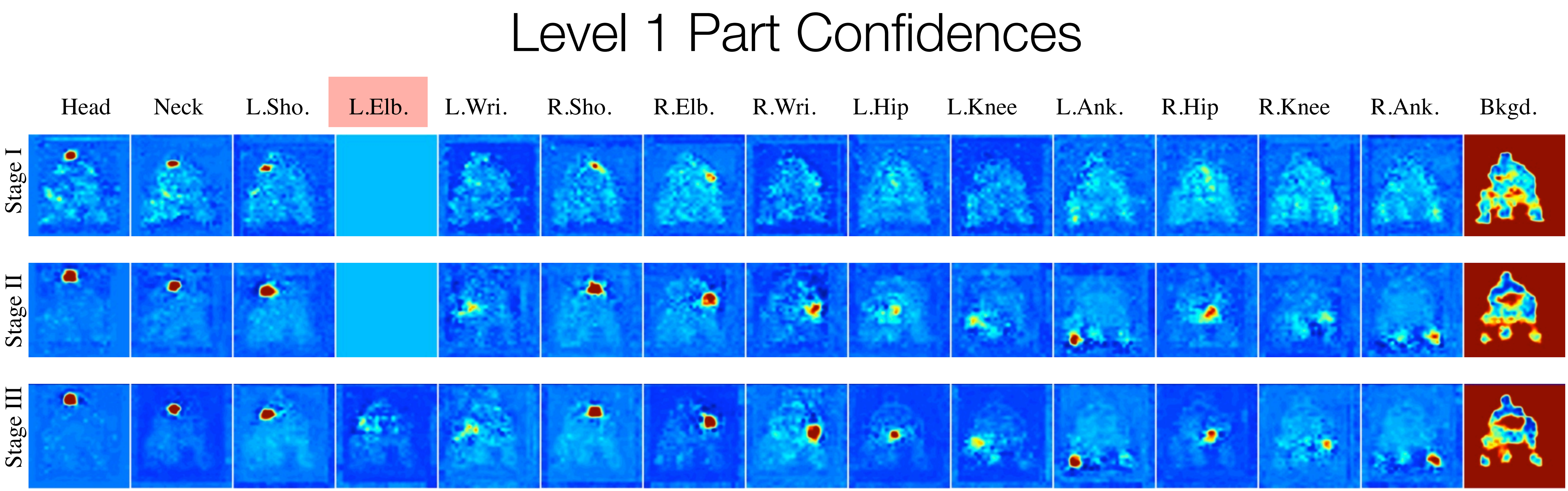
Level 1 Part Confidences



Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

Predicted Pose

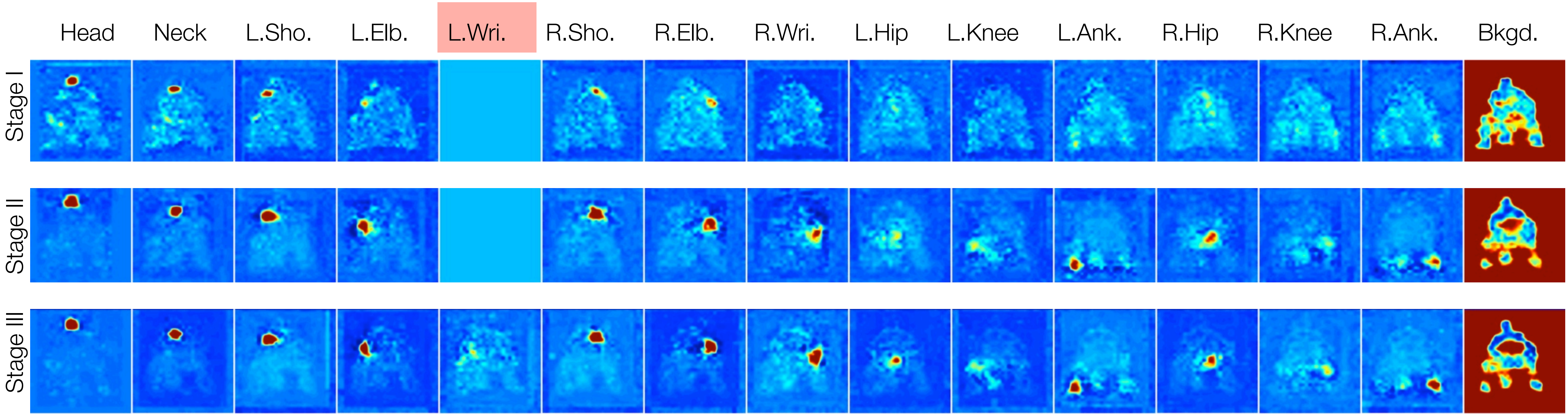


Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

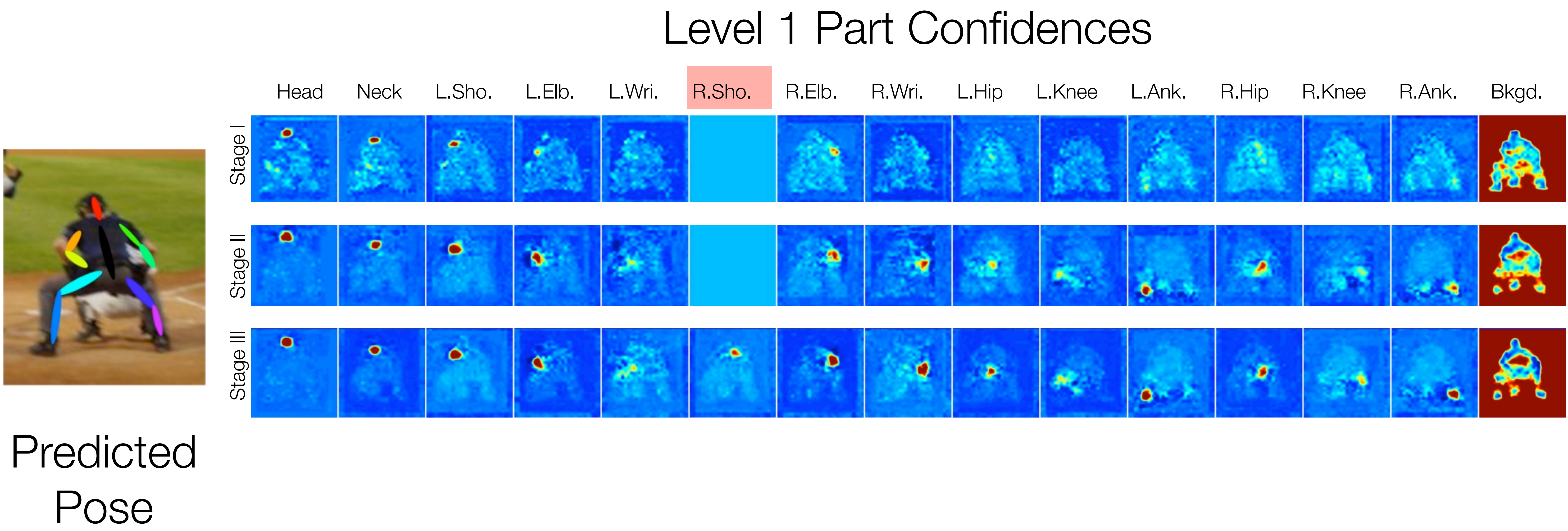
Predicted
Pose

Level 1 Part Confidences



Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

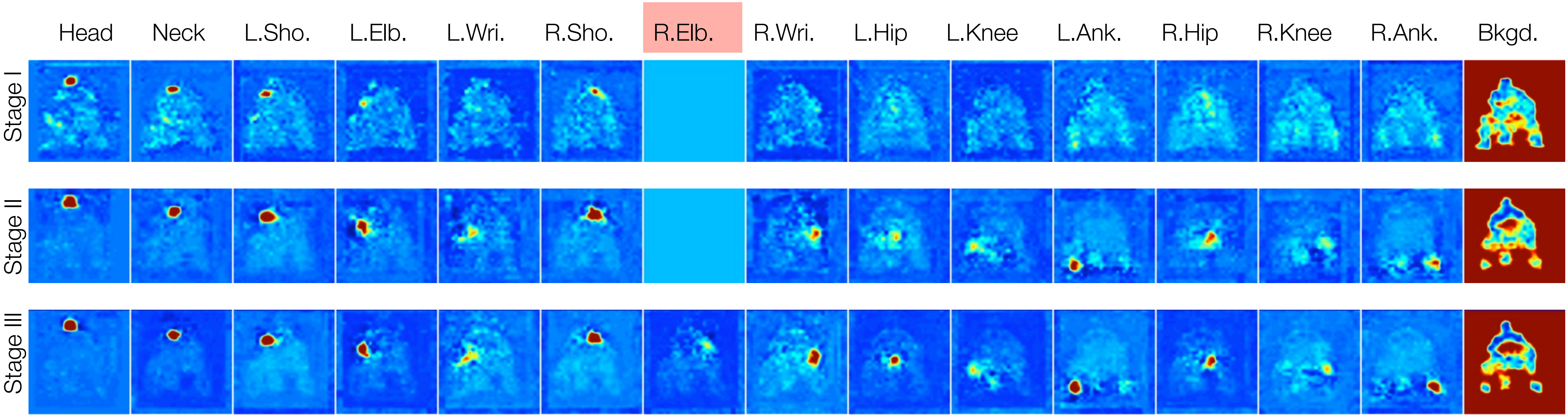


Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

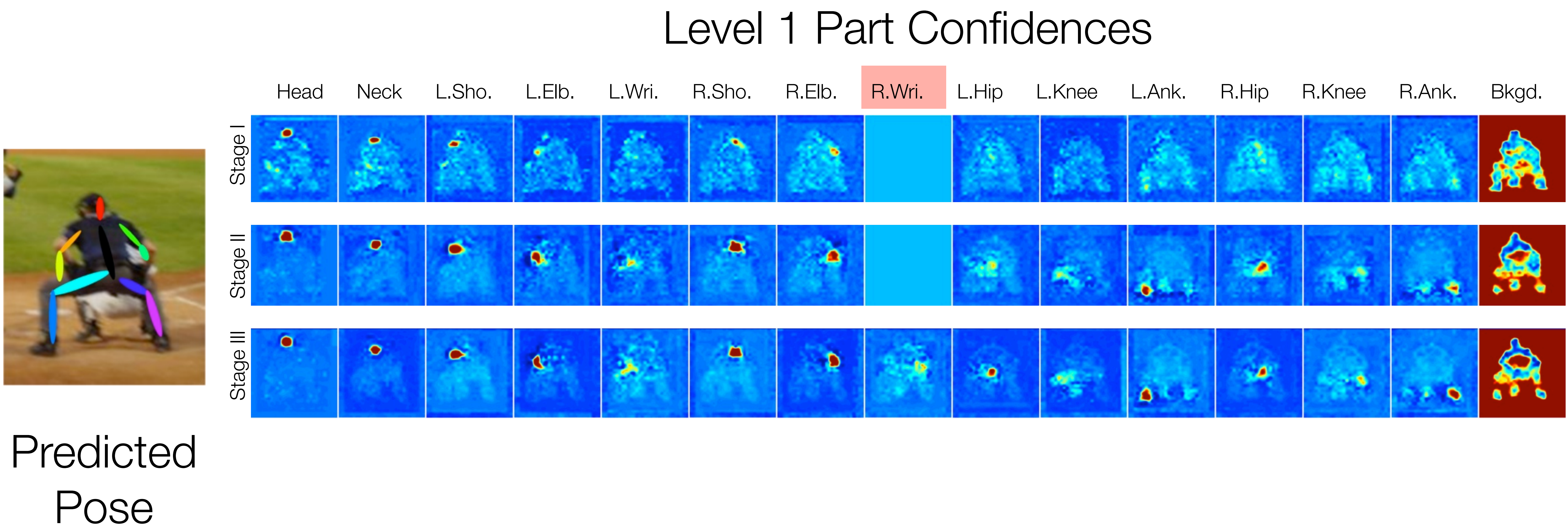
Level 1 Part Confidences

Predicted Pose



Predicted confidences are resilient to missing context (of one part)

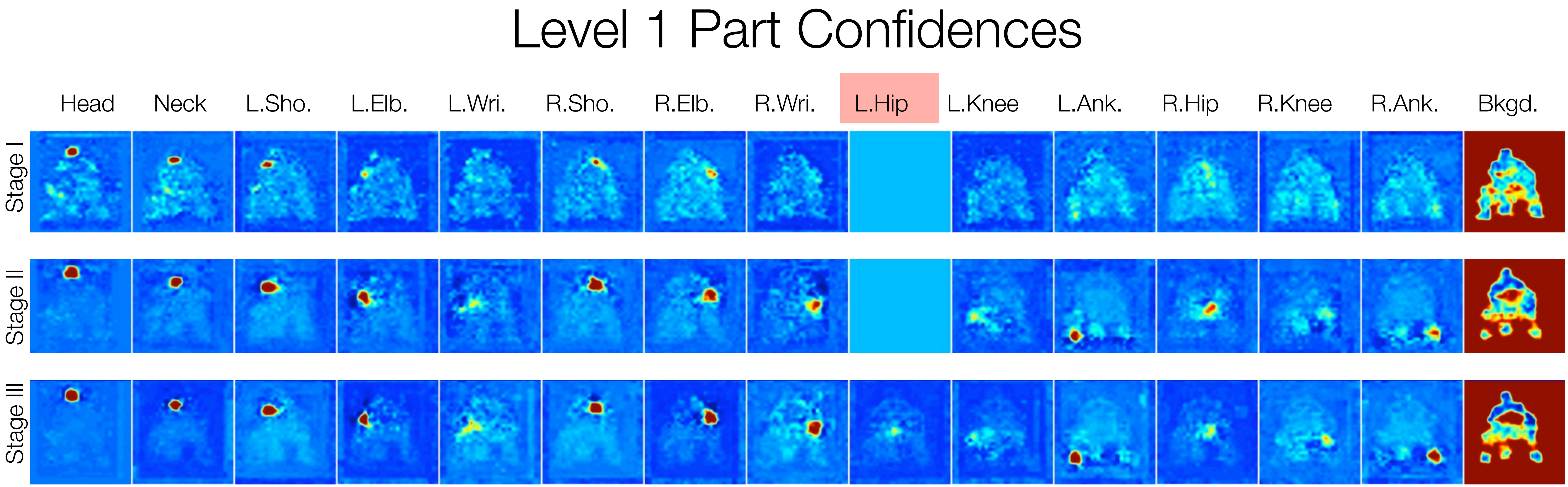
Ablative Spatial Analysis



Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

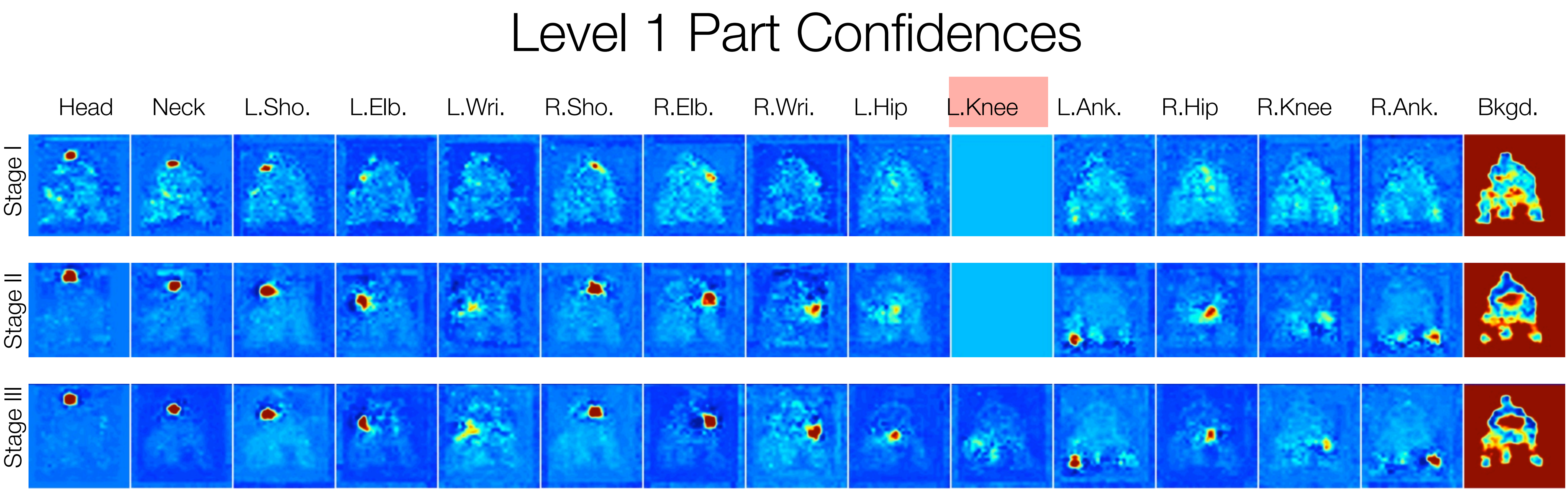
Predicted Pose



Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

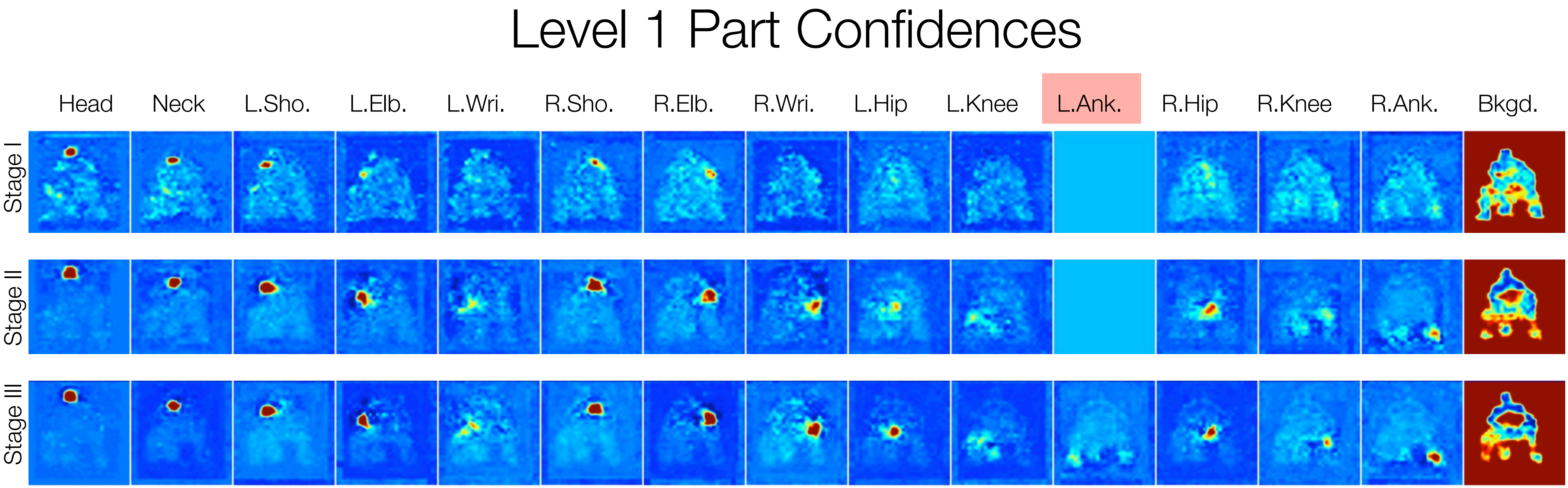
Predicted Pose



Predicted confidences are resilient to missing context (of one part)

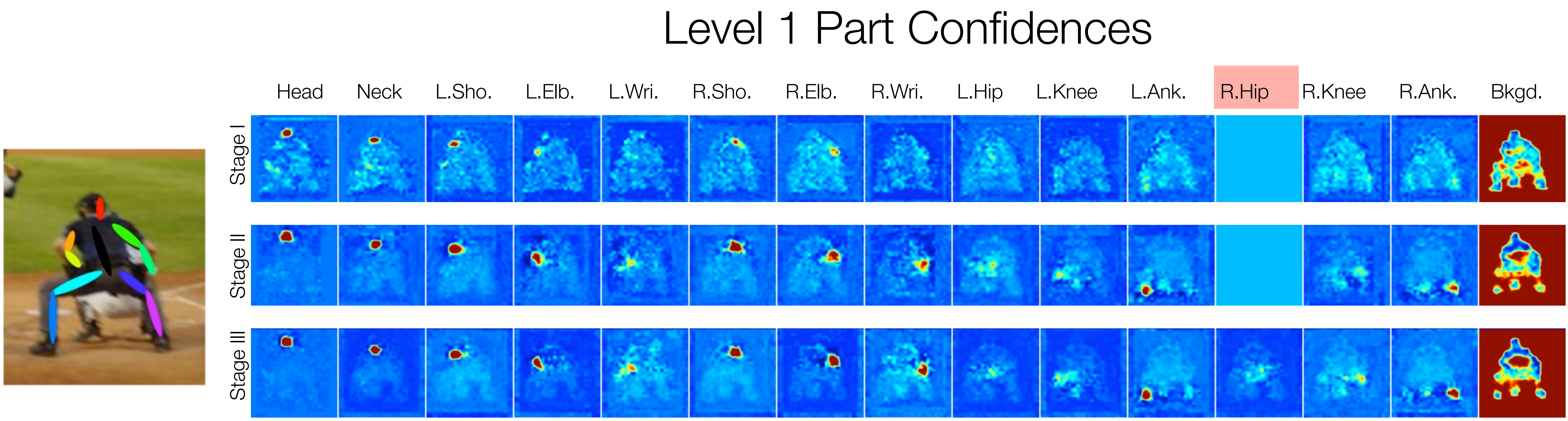
Ablative Spatial Analysis

Predicted Pose



Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis



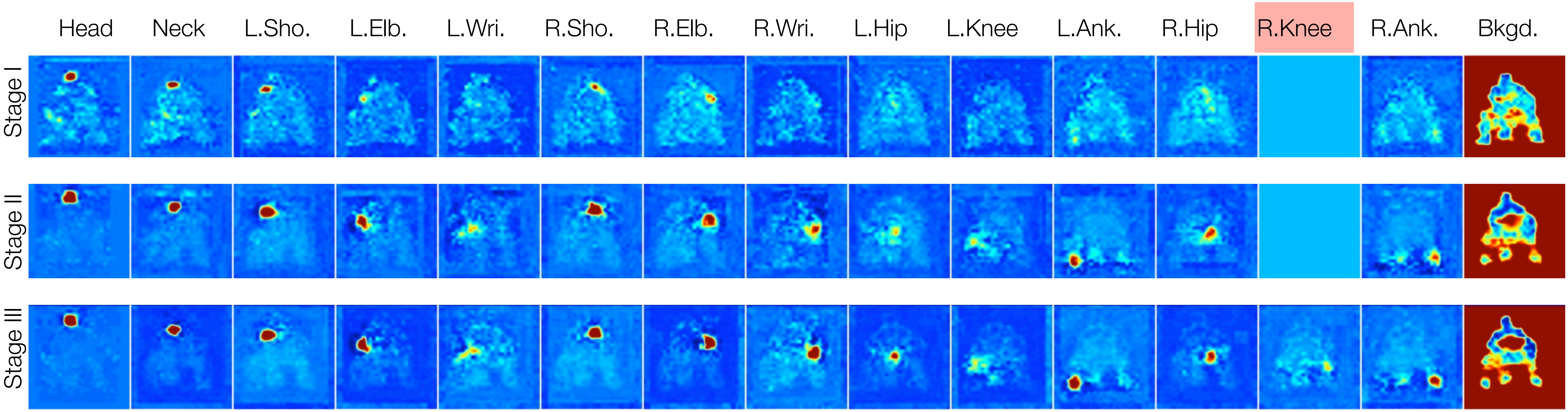
Predicted Pose

Predicted confidences are resilient to missing context (of one part)

Ablative Spatial Analysis

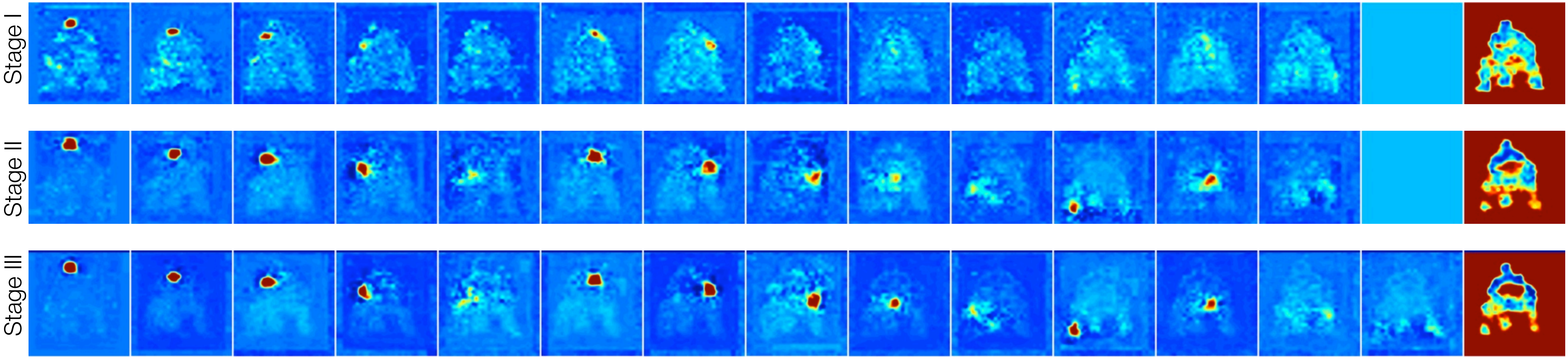
Predicted Pose

Level 1 Part Confidences



Predicted confidences are resilient to missing context (of one part)

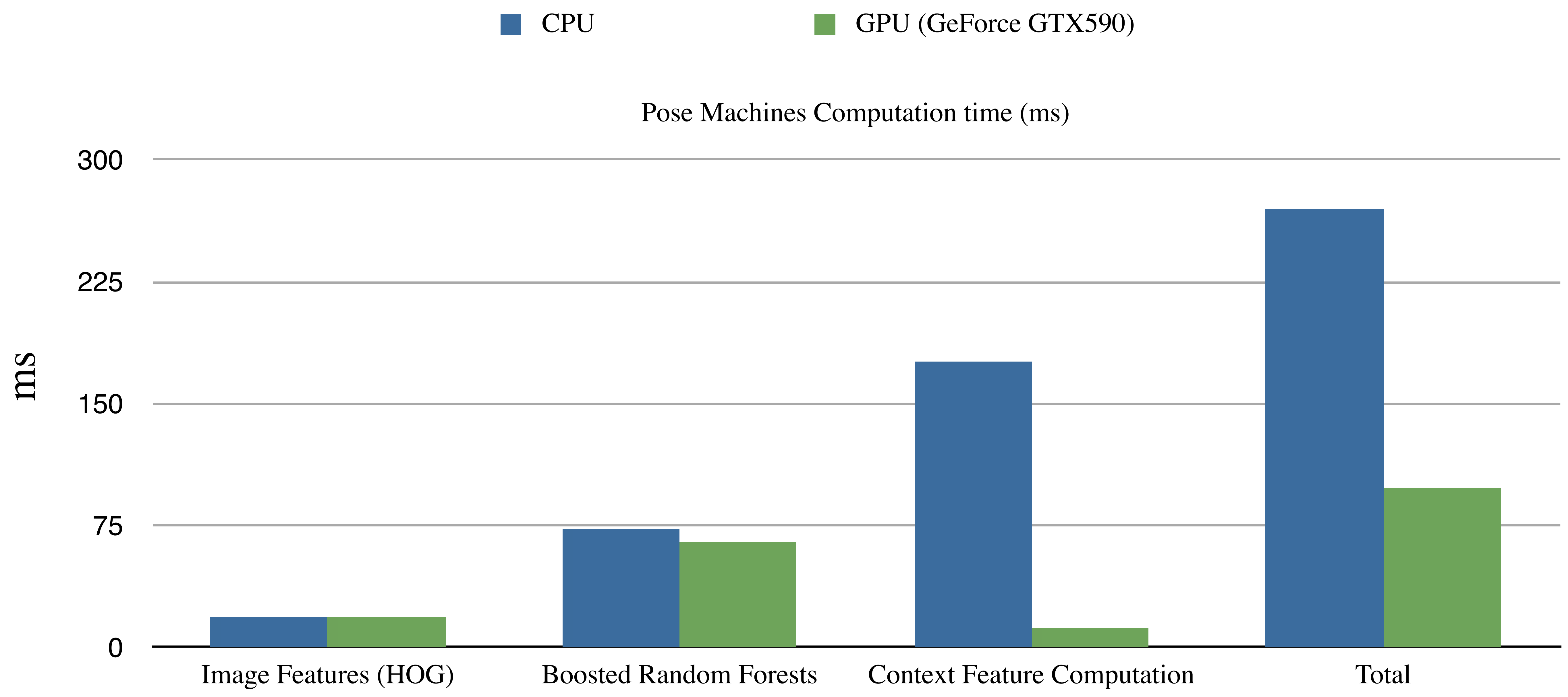
Ablative Spatial Analysis



Predicted confidences are resilient to missing context (of one part)

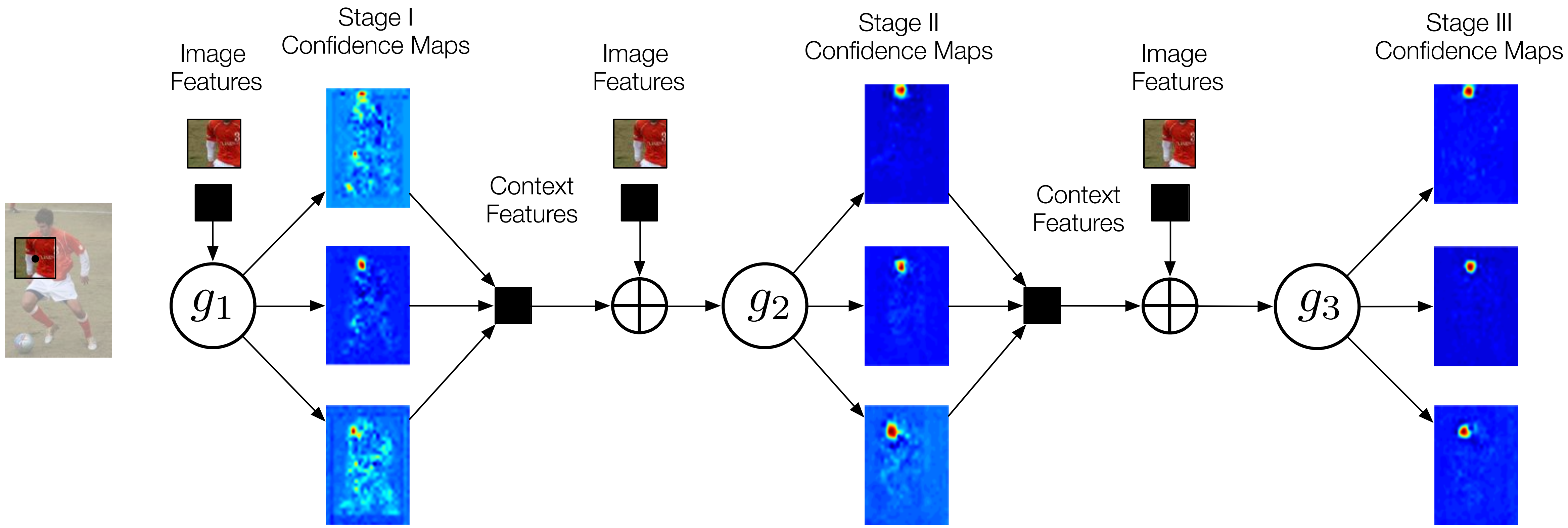
Efficient Prediction (~10 fps)

Fast and Parallelizable Inference



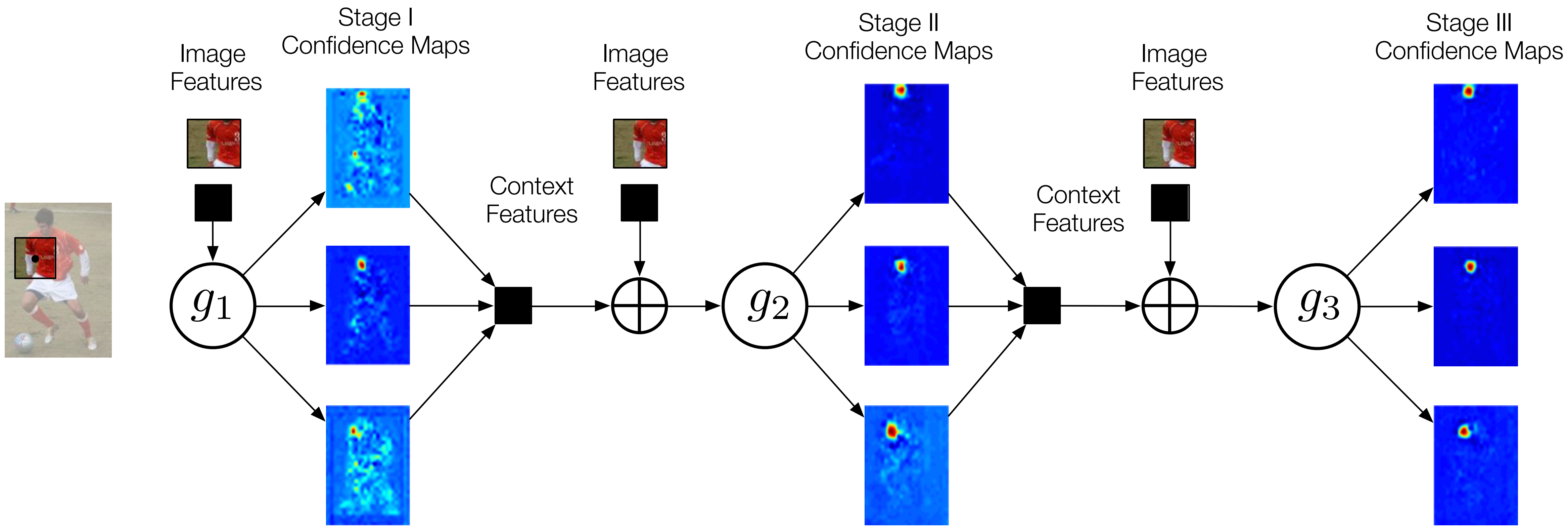
Conclusions

Pose Machines: Articulated Pose Estimation via Inference Machines



Conclusions

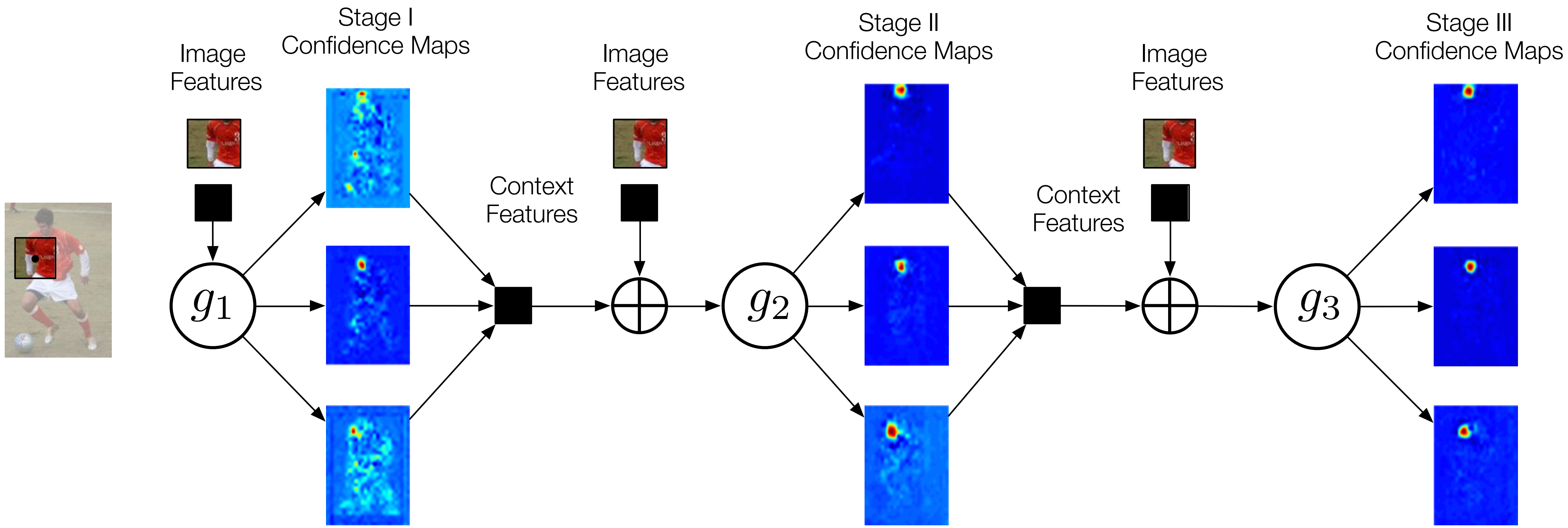
Pose Machines: Articulated Pose Estimation via Inference Machines



Local image evidence is weak

Conclusions

Pose Machines: Articulated Pose Estimation via Inference Machines

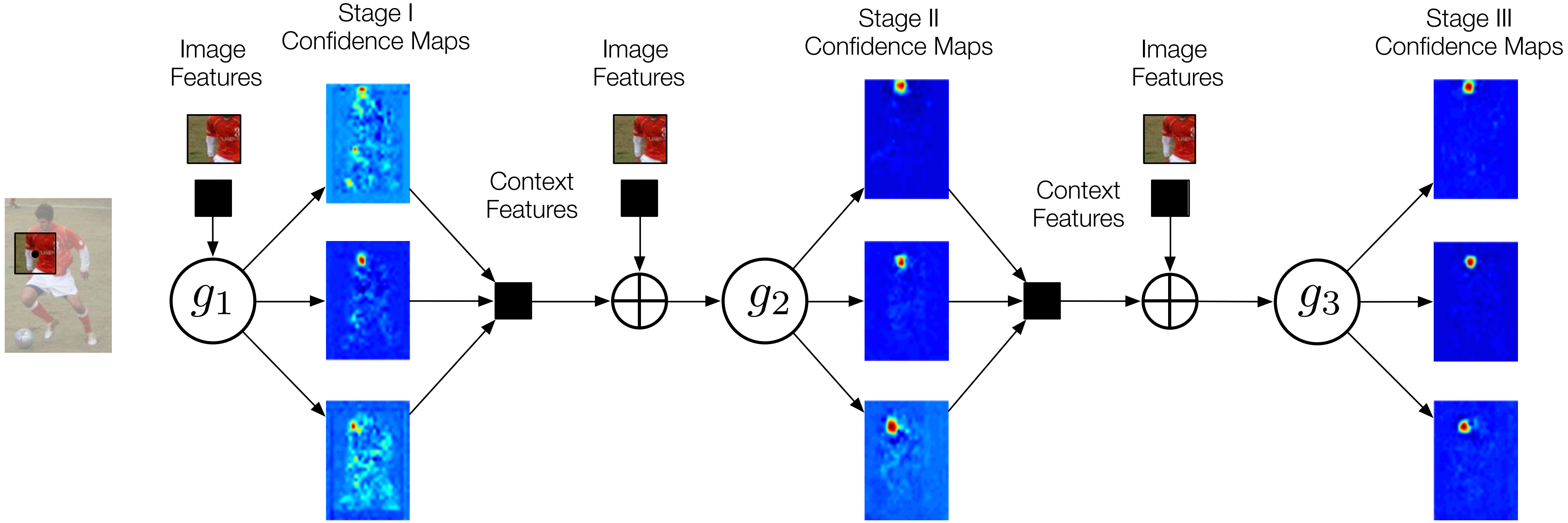


Local image evidence is weak

Sequential classification with modular architecture

Conclusions

Pose Machines: Articulated Pose Estimation via Inference Machines



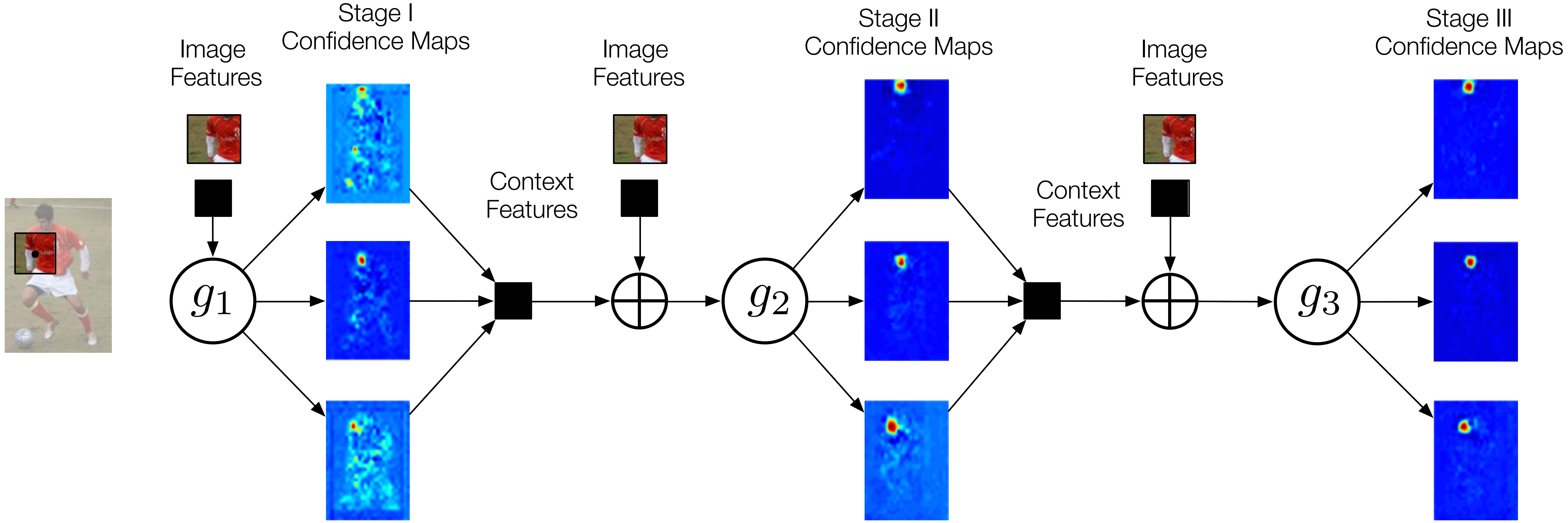
Local image evidence is weak

Sequential classification with modular architecture

Part context is a strong cue

Conclusions

Pose Machines: Articulated Pose Estimation via Inference Machines



Local image evidence is weak

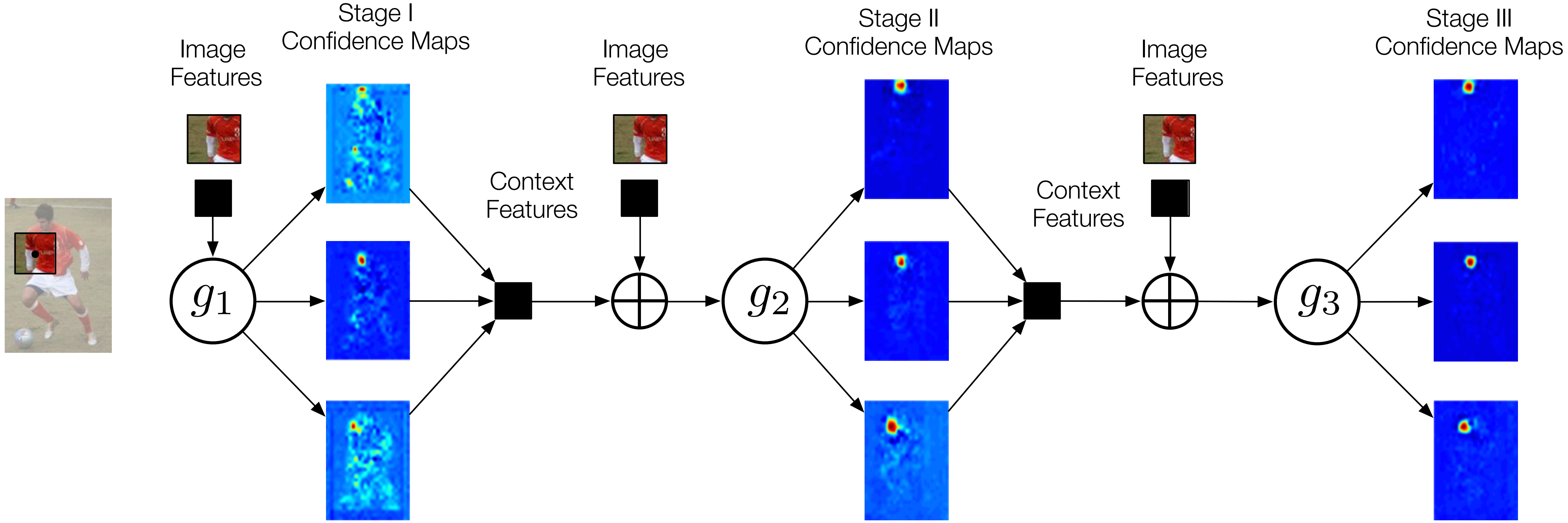
Sequential classification with modular architecture

Part context is a strong cue

Large composite parts are easier to detect

Conclusions

Pose Machines: Articulated Pose Estimation via Inference Machines



Local image evidence is weak

Sequential classification with modular architecture

Part context is a strong cue

Large composite parts are easier to detect

Implicitly learn rich spatial and hierarchical relationships

Thank You

www.cs.cmu.edu/~vramakri/poseMachines.html

Varun Ramakrishna, Daniel Munoz, Martial Hebert,
J. Andrew Bagnell, Yaser Sheikh

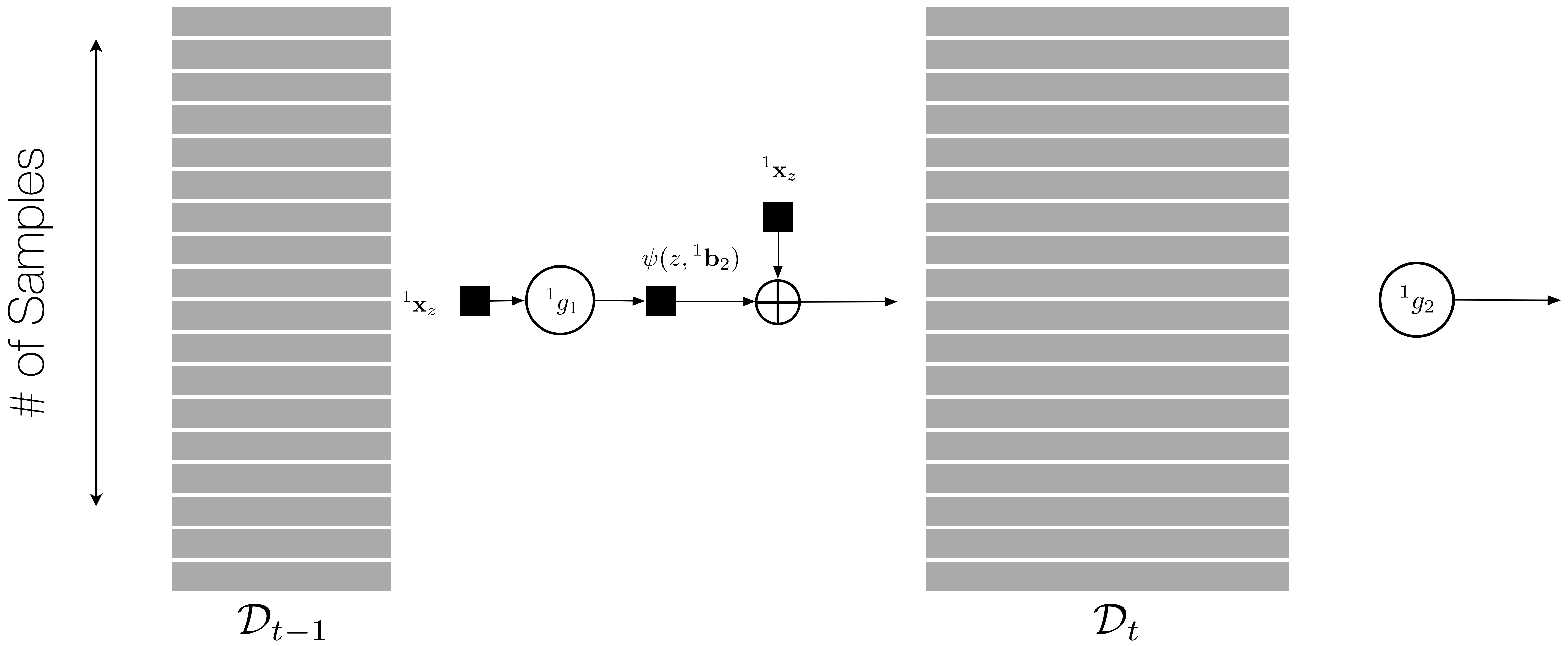
{vramakri, dmunoz, hebert, dbagnell, yaser}@cs.cmu.edu

Carnegie Mellon University

THE ROBOTICS INSTITUTE

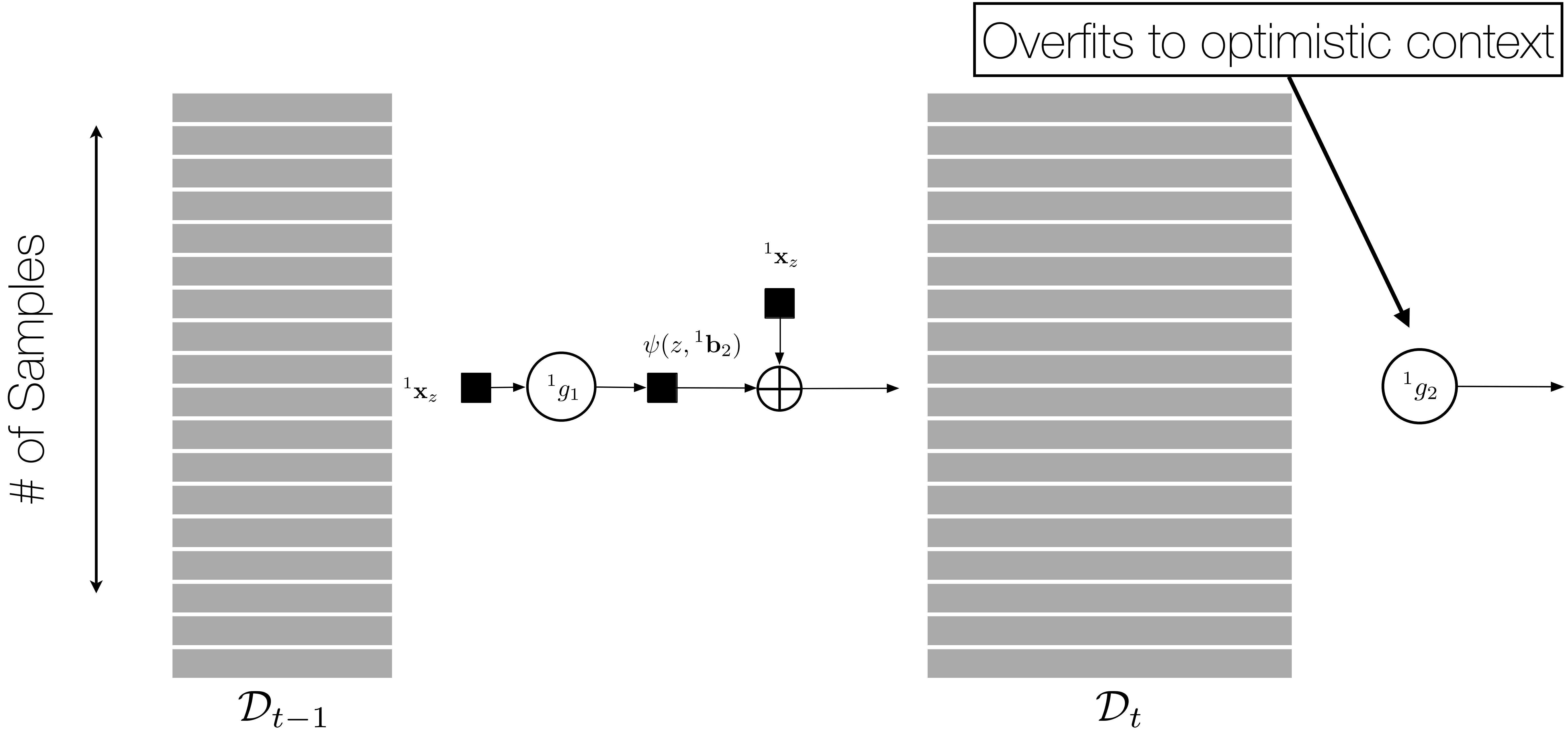
Backup Slides

Stacking



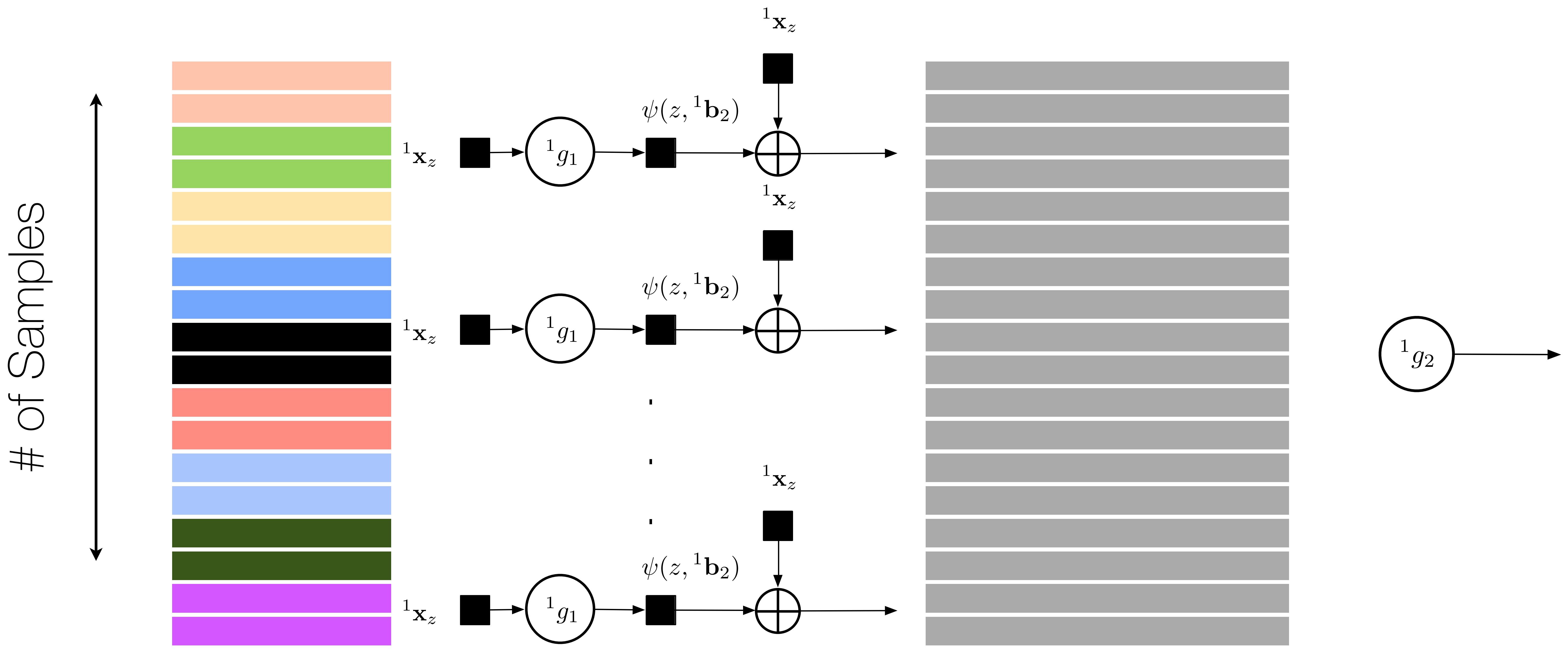
Stacked Generalization, Wolpert et. al

Stacking



Stacked Generalization, Wolpert et. al

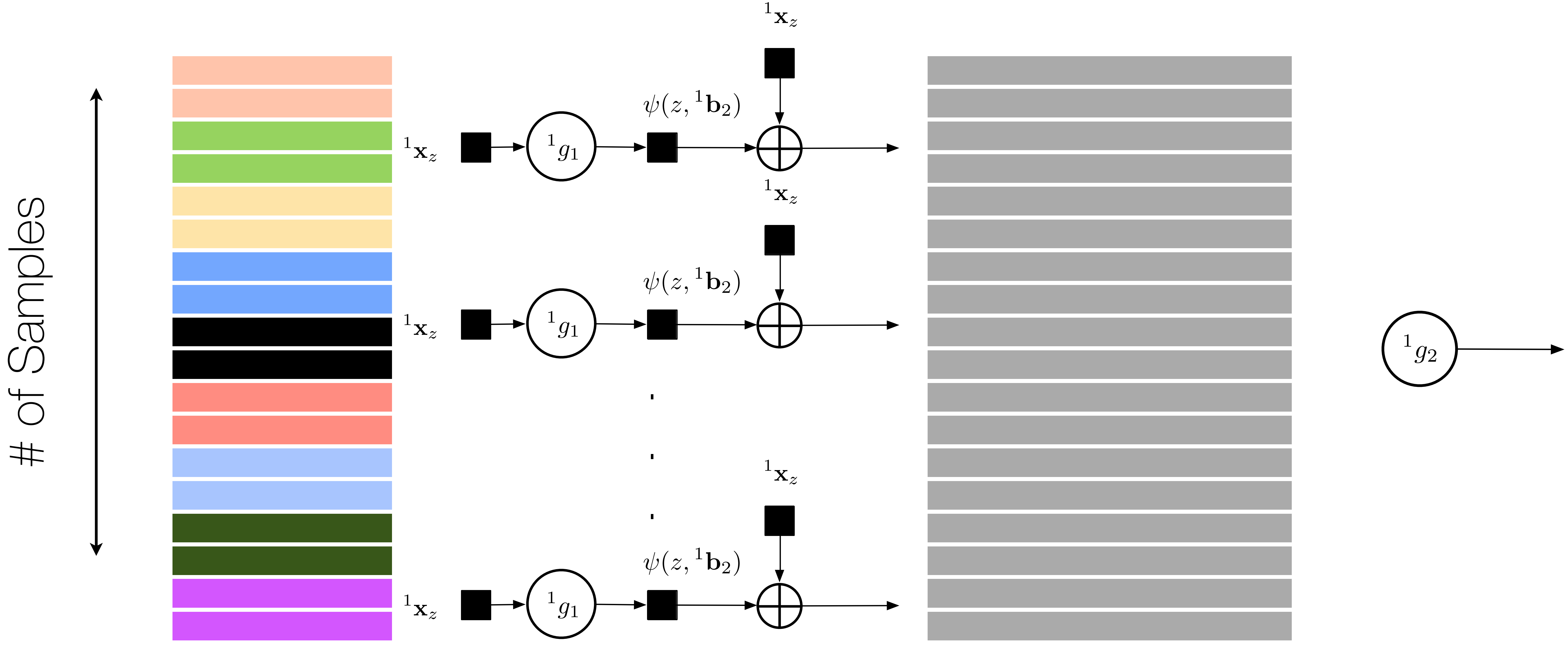
Stacking



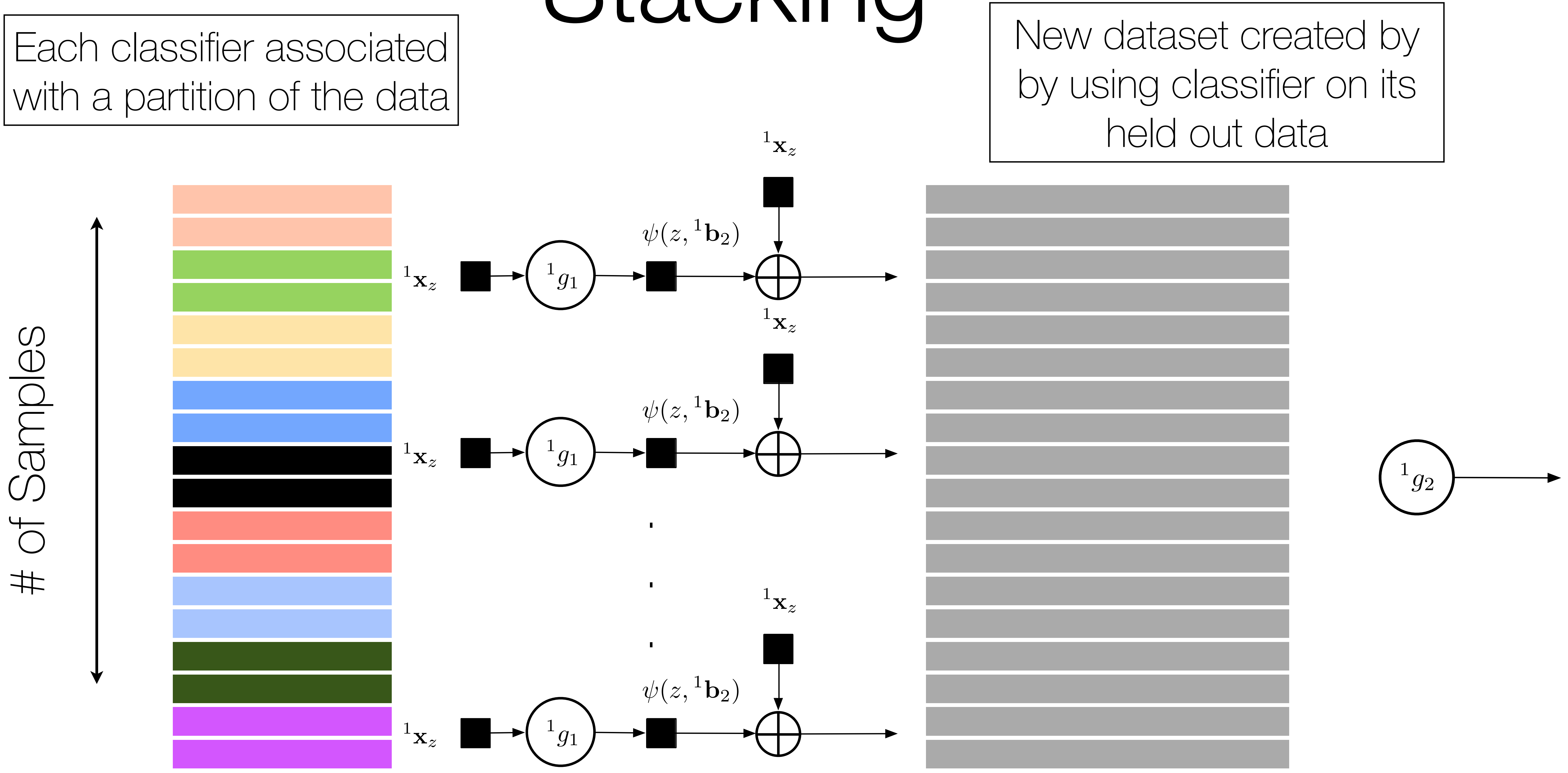
# of Samples
$^1\mathbf{x}_z$
$^1g_1$
$\psi(z, {}^1\mathbf{b}_2)$
$^1\mathbf{x}_z$
$^1g_1$
$\psi(z, {}^1\mathbf{b}_2)$
$^1\mathbf{x}_z$
$^1g_1$
$\psi(z, {}^1\mathbf{b}_2)$
$^1g_2$

Stacking

Each classifier associated with a partition of the data



Stacking



Each classifier associated with a partition of the data
New dataset created by by using classifier on its held out data
# of Samples
$^1\mathbf{x}_z$
$^1g_1$
$\psi(z, {}^1\mathbf{b}_2)$
$^1g_2$

Choice of Classifier

Boosted Random Forest with a Max-Margin Loss Functional

$$\mathcal{L}(f) = \frac{\lambda}{2}\|f\|^2 + \sum_i max\left(0, 1 - f(x_i, y_i) + f(x_i, y)\right)$$

Functional Sub-gradient Descent == Boosting

Choice of Classifier