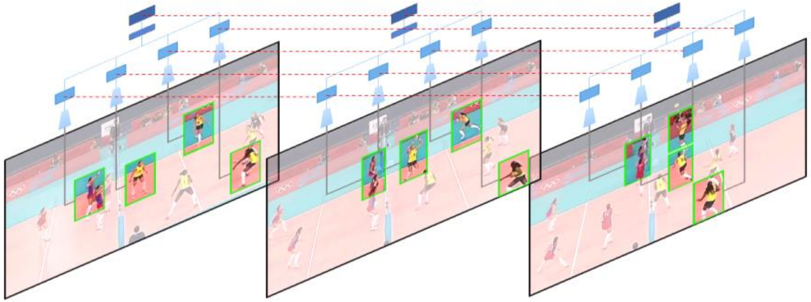
SFU

# Deep Structured Models for Group Activities and Label Hierarchies

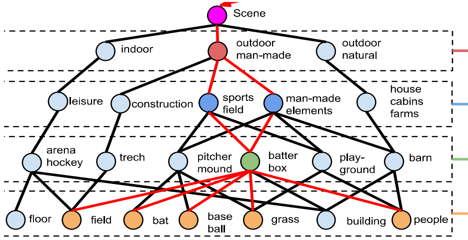
Greg Mori
Simon Fraser University

# Outline

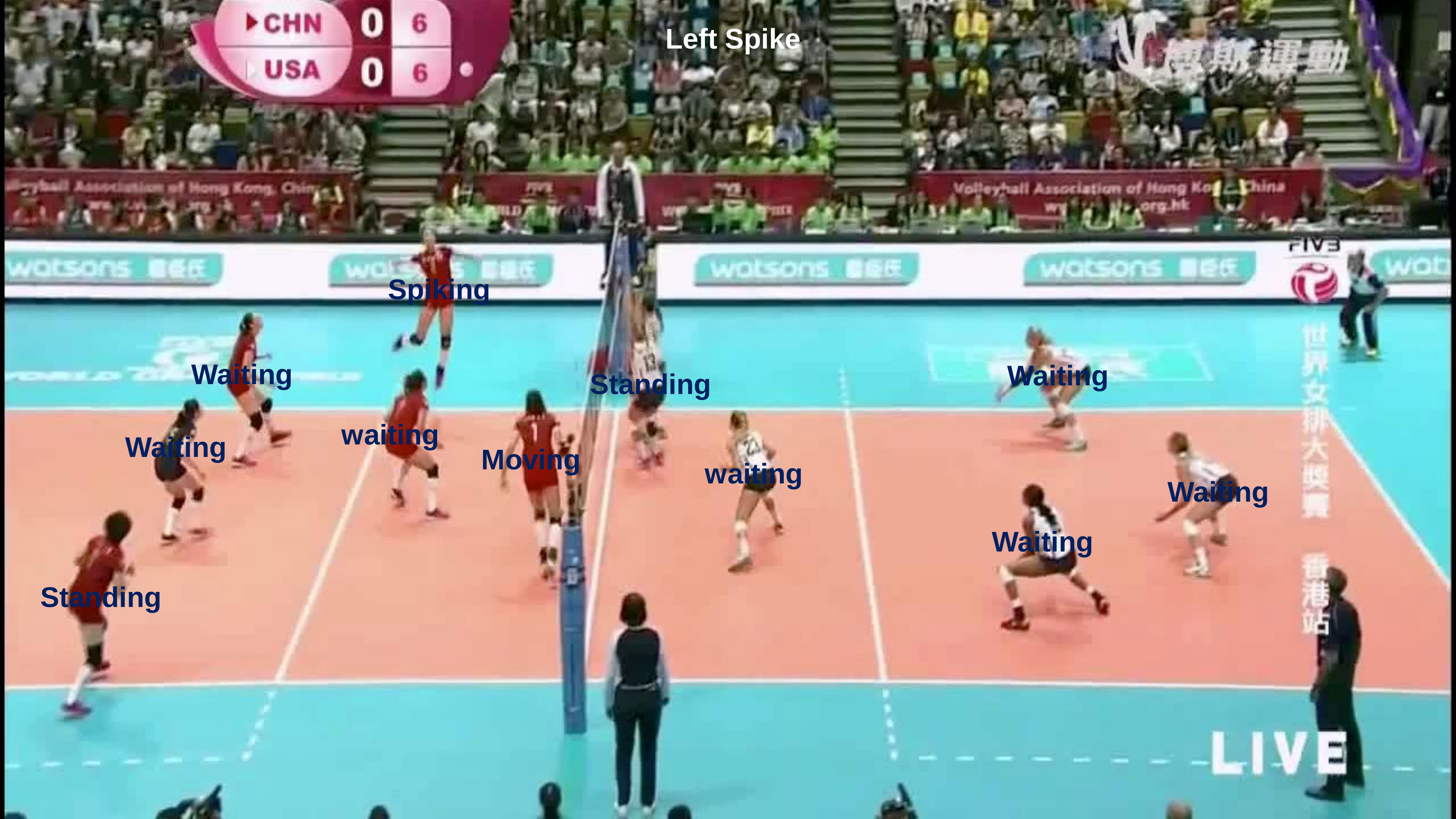
- Temporal structured models for group activities
  - Ibrahim et al. CVPR 2016



- Image annotation with label hierarchies
  - Hu et al. CVPR 2016



# Example: Rally in a Volleyball Game


Left Spike
CHN 0 6
USA 0 6
Spiking
Waiting
Waiting
waiting
Moving
Standing
Standing
waiting
Waiting
Waiting
Waiting
LIVE

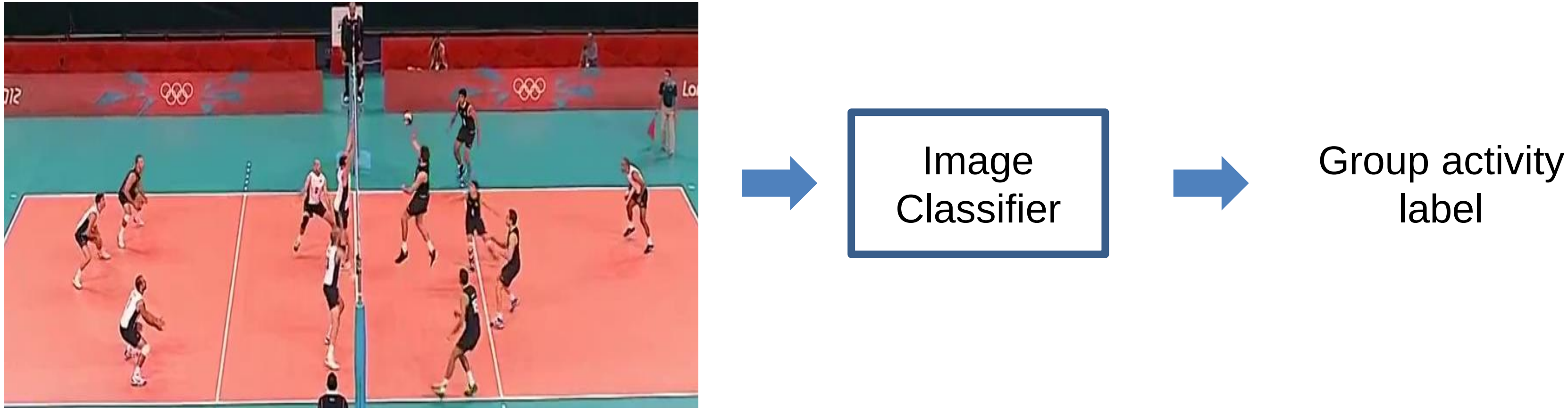


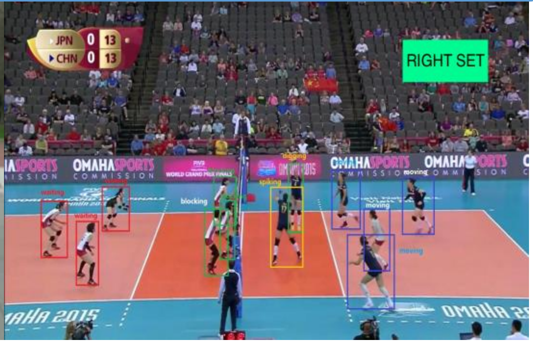
Image
Classifier
Group activity
label

Challenge:

- high level description
- aggregate information over whole scene
- focus on relevant people

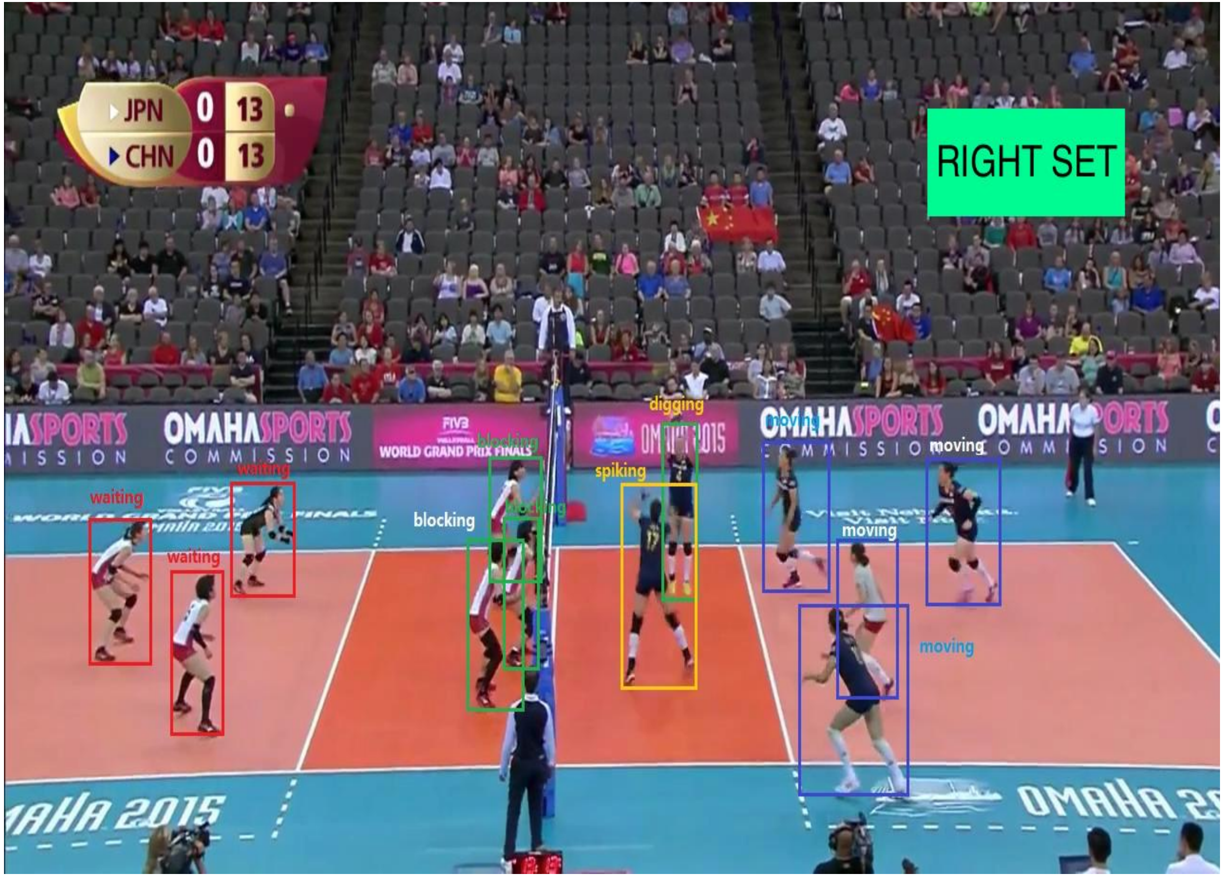
Group Activity = Majority's Activity


Group Activity = Key Player's Activity



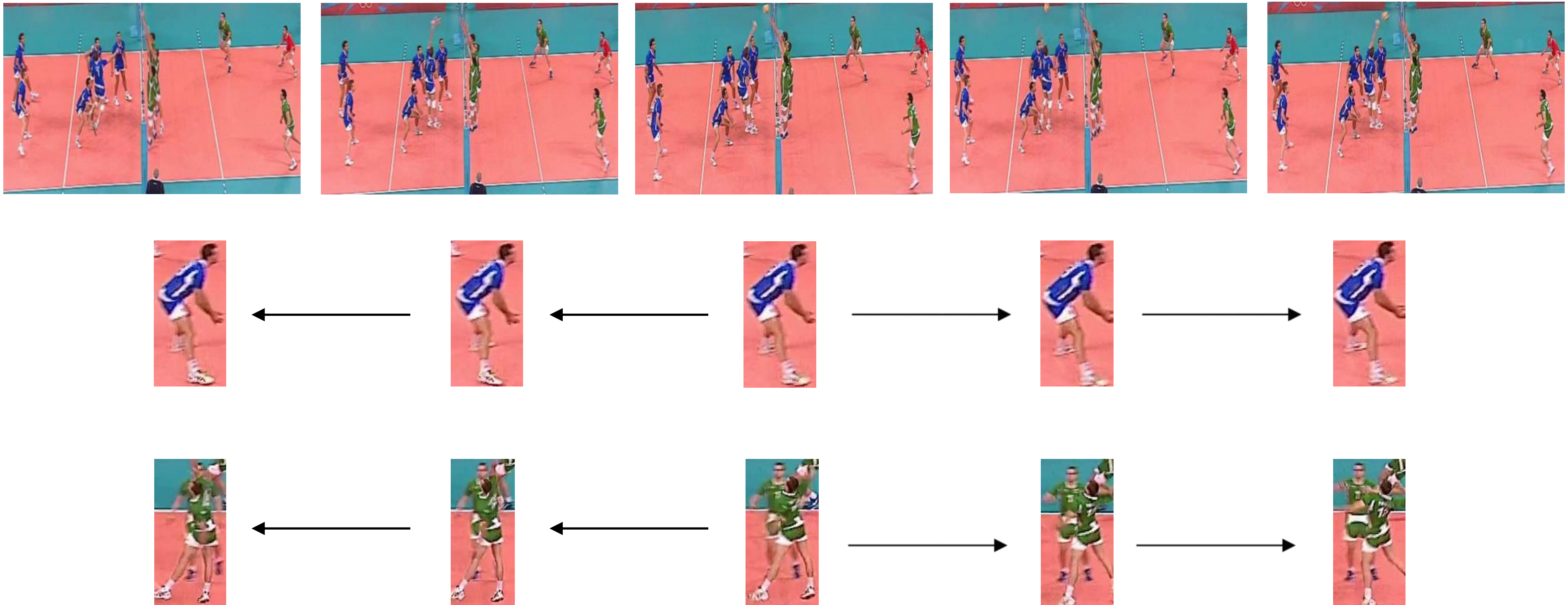

Group Activity – Right spike

Intuitive fix: use person-centric representation

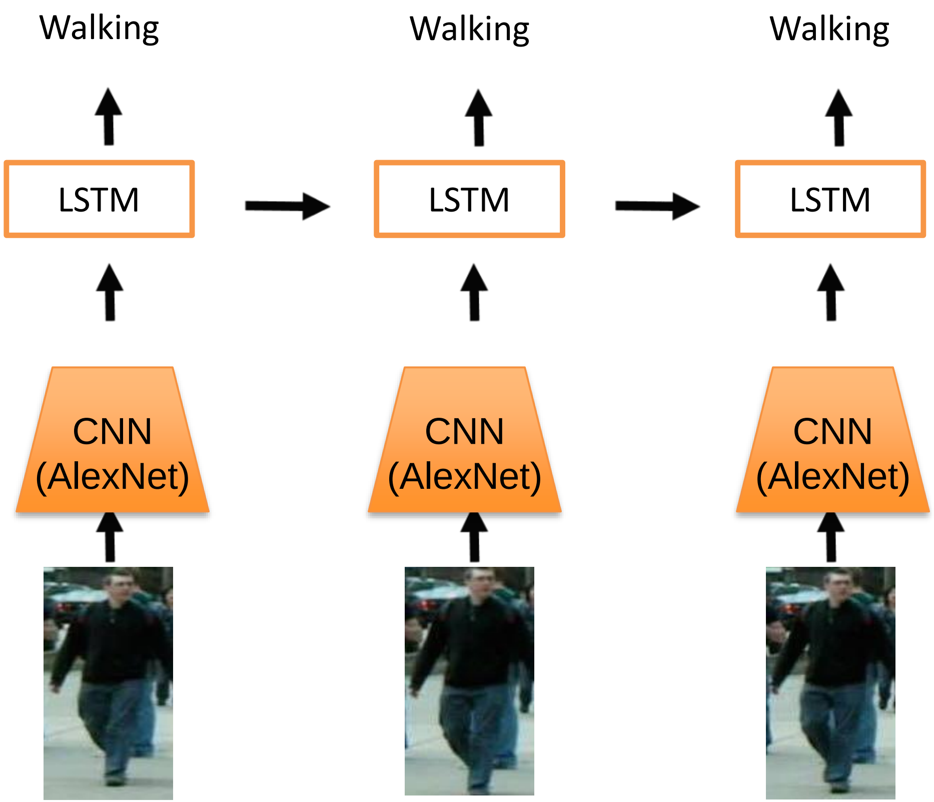


# Person Tracks

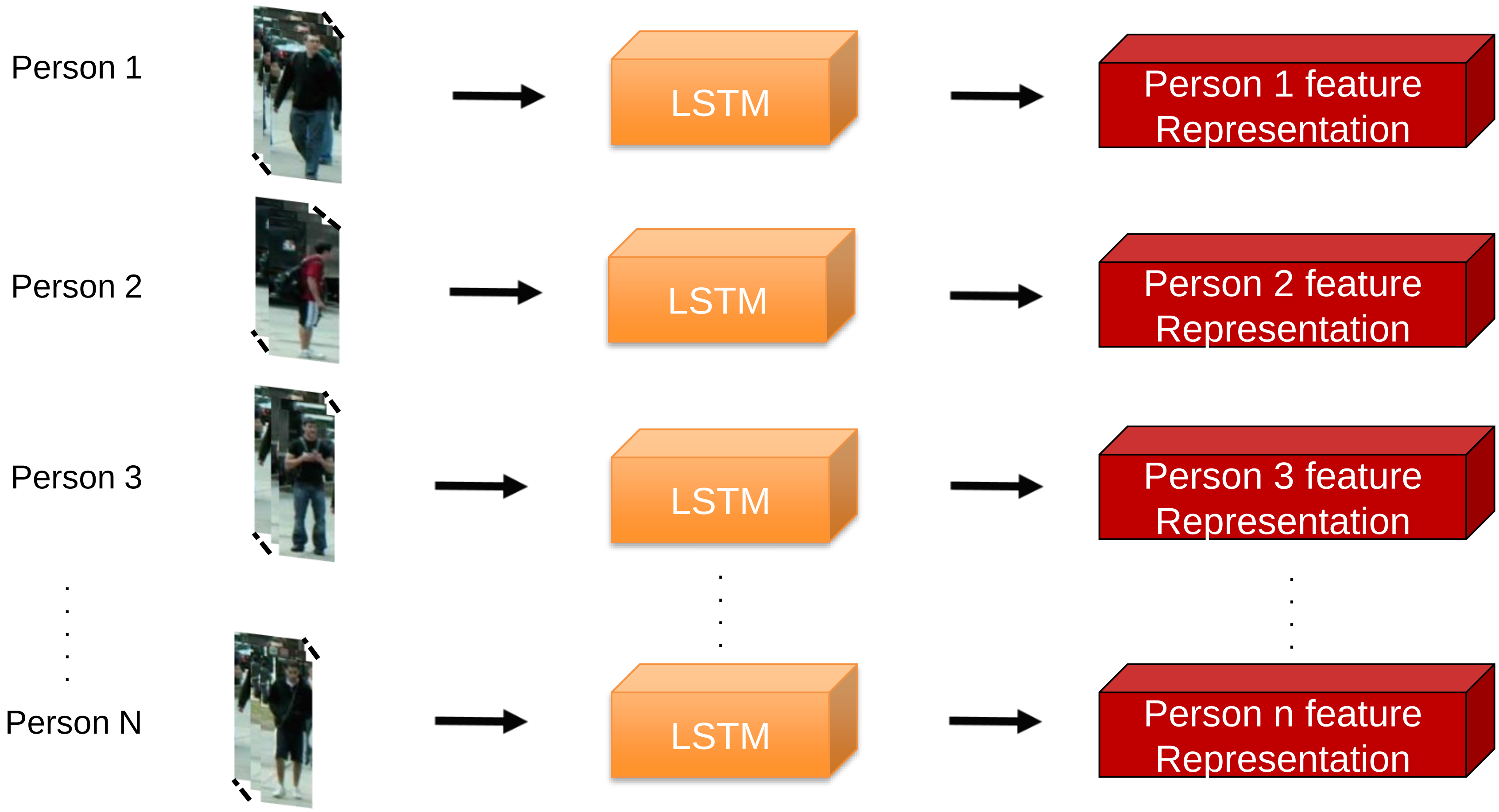
- Extract trajectories by tracking each person forward/backward in time



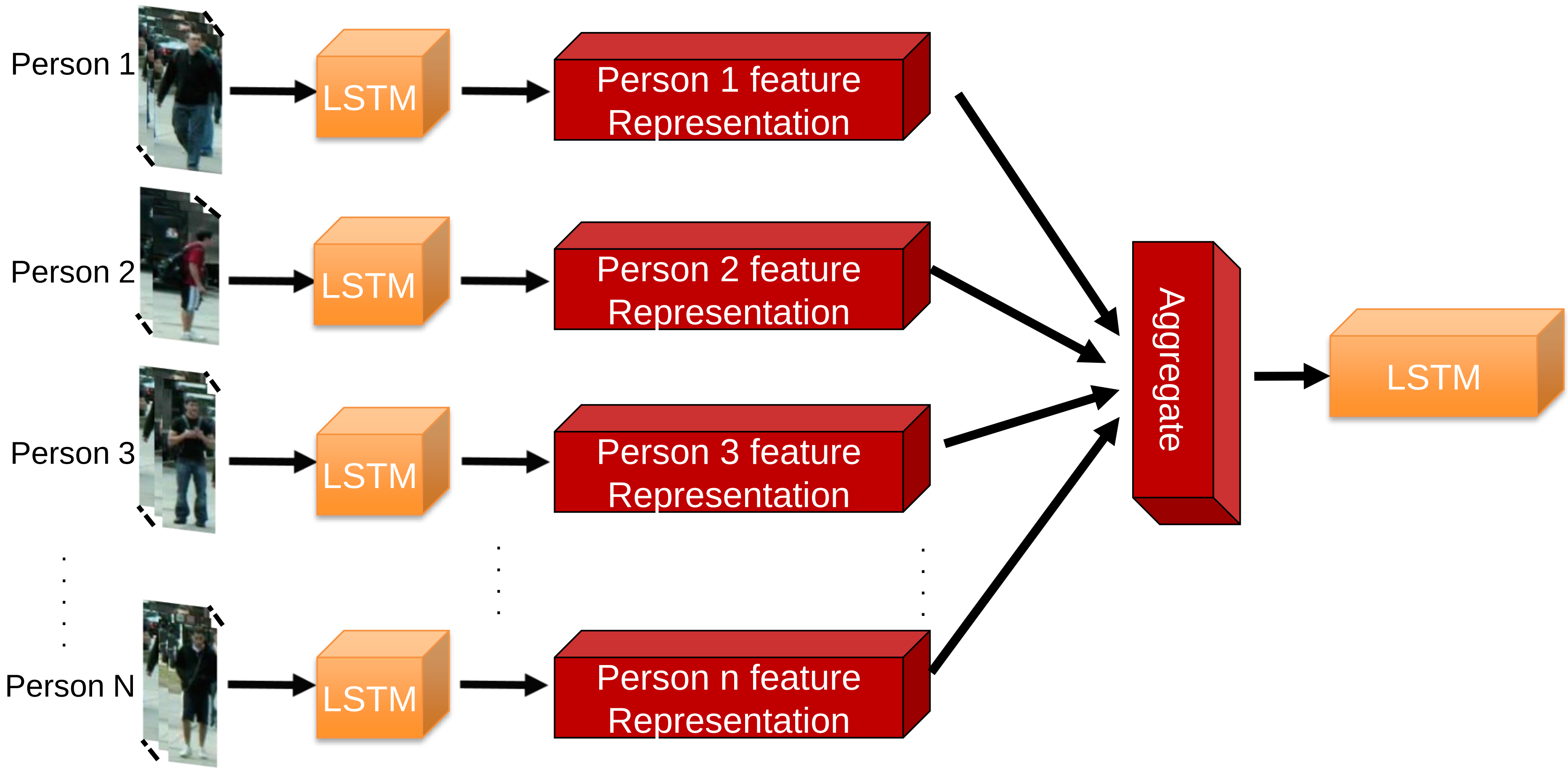
# Stage 1 : Learning Individual Action Features



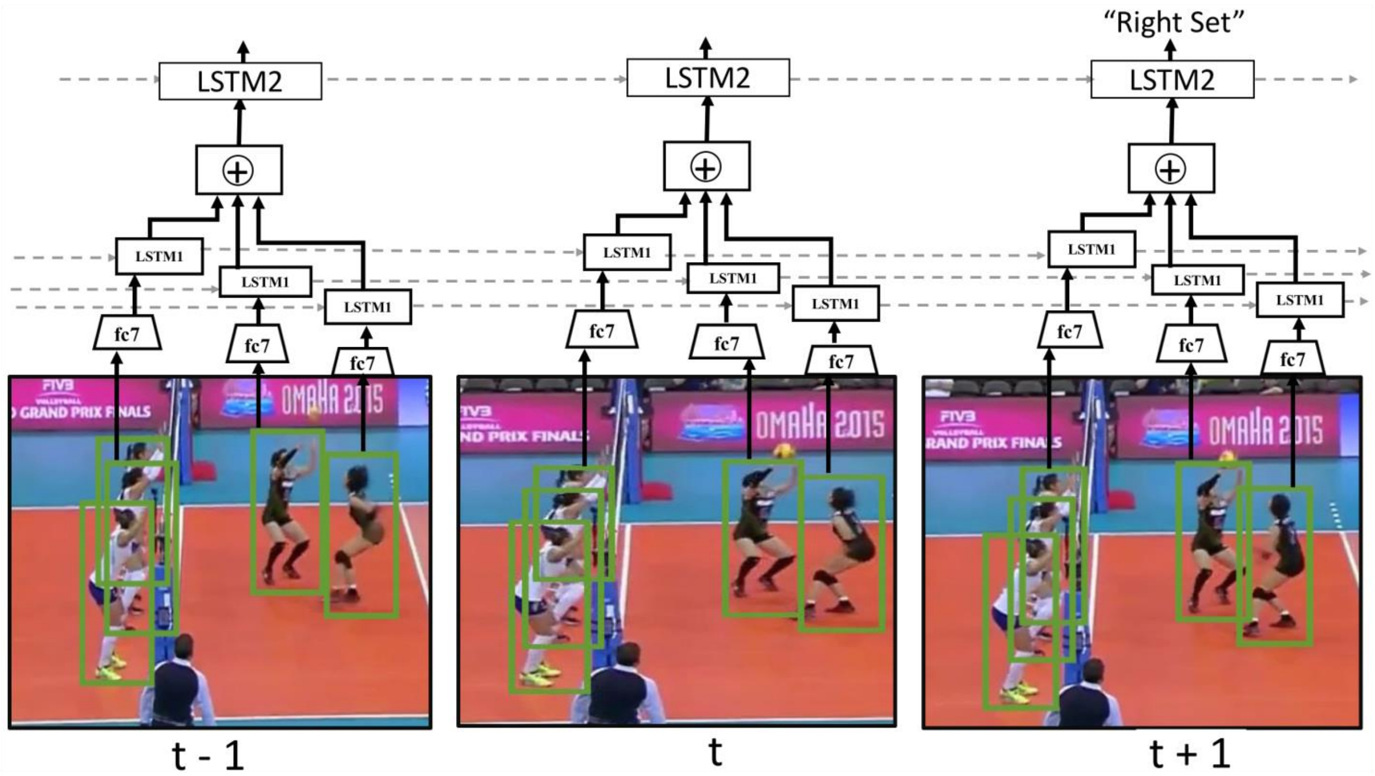
# Stage1 : Learning Individual Action Features



# Stage 2: Learning Frame Representations



# Summary

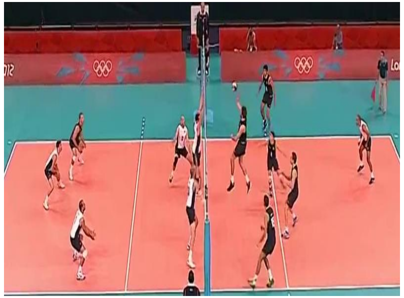


"Right Set"
LSTM2
LSTM2
LSTM2
LSTM1
LSTM1
LSTM1
LSTM1
LSTM1
LSTM1
LSTM1
LSTM1
LSTM1
fc7
fc7
fc7
fc7
fc7
fc7
fc7
fc7
fc7
OMAHA 2015
t - 1
t
t + 1

# Volleyball Dataset – Frame Labels

- 4830 frames annotated from 55 volleyball videos
- 2/3 videos for training, 1/3 testing
- 9 player action labels
- 4 scene labels

**Spiking**



**Setting**

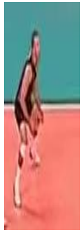
**Passing**

**Win point**

Left/right team variants

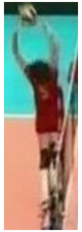
# Volleyball Dataset – People Labels

**Waiting**

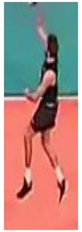


**Digging**

**Setting**



**Spiking**



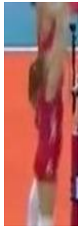
**Falling**

**Jumping**

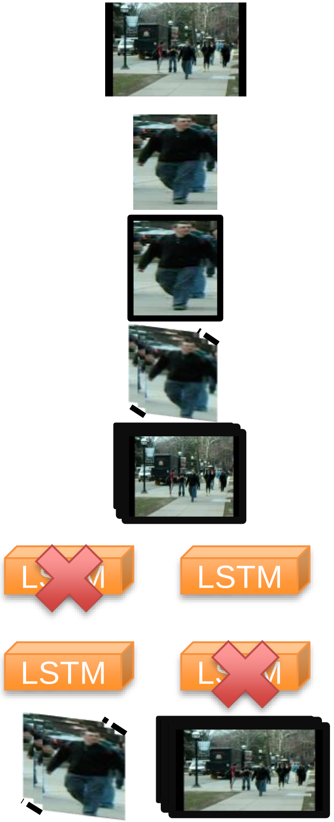
**Moving**

**Standing**

**Blocking**



# Experimental results on Volleyball Dataset

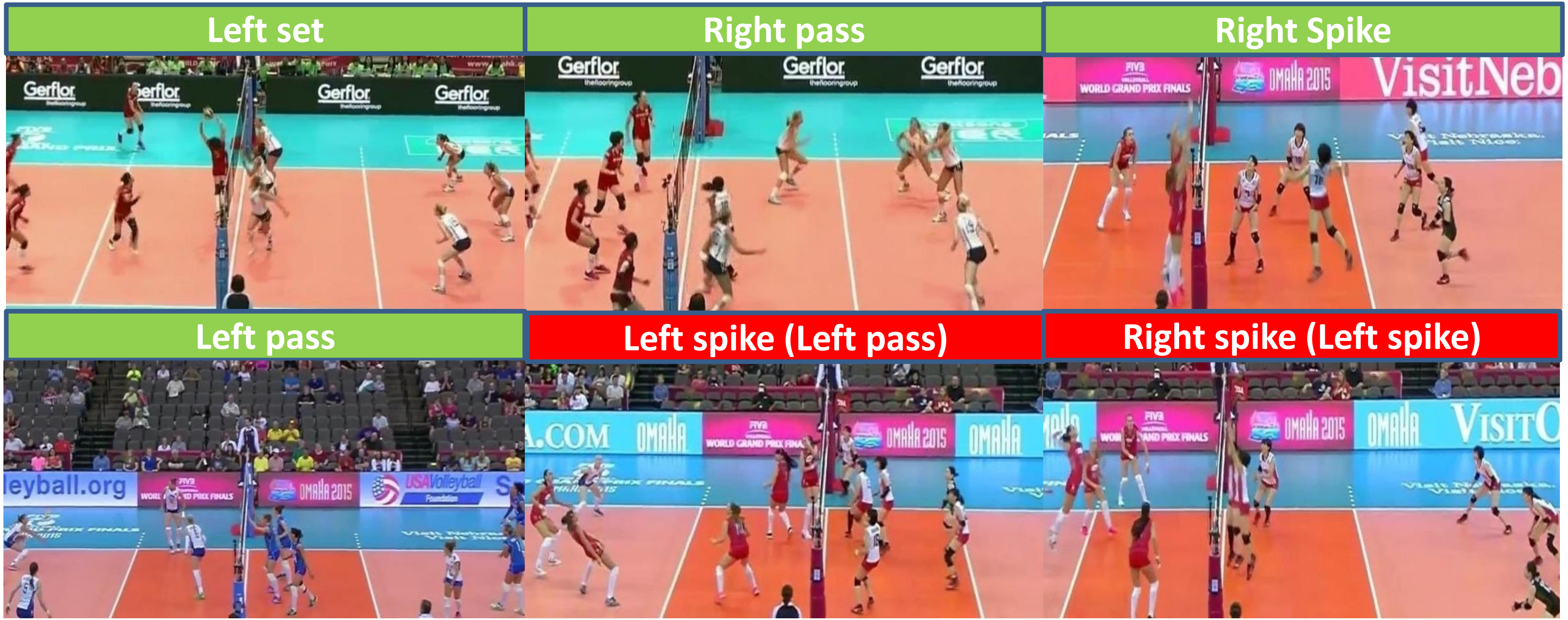



| Method | Accuracy |
|---|---|
| Image Classification | 66.7 |
| Person Classification | 64.5 |
| Person - Fine tuned | 66.8 |
| Temp Model - Person | 67.5 |
| Temp Model - Image | 63.1 |
| Our Model w/o LSTM1 | 73.3 |
| Our Model w/o LSTM2 | 80.9 |
| Our Model | 81.6 |

| | lpass | rpass | lset | rset | lspike | rspike | lwin | rwin |
|---|---|---|---|---|---|---|---|---|
| lpass | 79.65 | 3.98 | 9.73 | 0.00 | 3.10 | 2.65 | 0.44 | 0.44 |
| rpass | 4.29 | 80.00 | 0.00 | 9.52 | 2.86 | 1.90 | 0.95 | 0.48 |
| lset | 8.33 | 1.79 | 85.12 | 0.60 | 2.38 | 1.19 | 0.60 | 0.00 |
| rset | 5.21 | 19.27 | 1.04 | 68.23 | 0.00 | 4.69 | 1.56 | 0.00 |
| lspike | 3.35 | 1.12 | 5.03 | 0.00 | 89.94 | 0.56 | 0.00 | 0.00 |
| rspike | 2.31 | 5.20 | 2.31 | 4.62 | 1.16 | 83.24 | 1.16 | 0.00 |
| lwin | 2.94 | 3.92 | 0.00 | 0.00 | 0.00 | 0.00 | 88.24 | 4.90 |
| rwin | 1.15 | 1.15 | 0.00 | 0.00 | 0.00 | 0.00 | 12.64 | 85.06 |

Dense trajectories: 73.4-78.7

# Visualization of results



# Summary

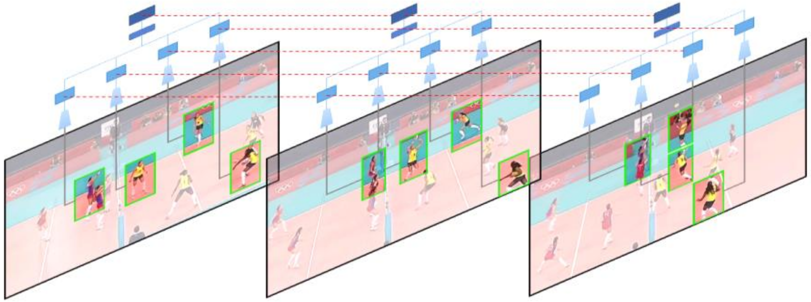
- A two stage hierarchical model for group activity recognition

- LSTMs as a highly effective temporal model and temporal feature source

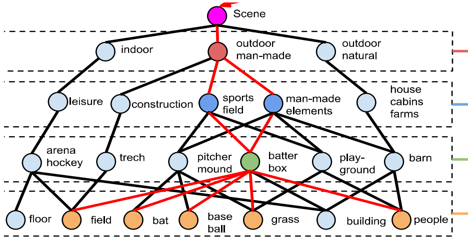
- People-relation modeling with simple pooling

# Outline

- Temporal structured models for group activities
  - Ibrahim et al. CVPR 2016



- Image annotation with label hierarchies
  - Hu et al. CVPR 2016



# Image Classification

- A natural image can be categorized with labels at different concept layers

Indoor | outdoor man-made | outdoor natural

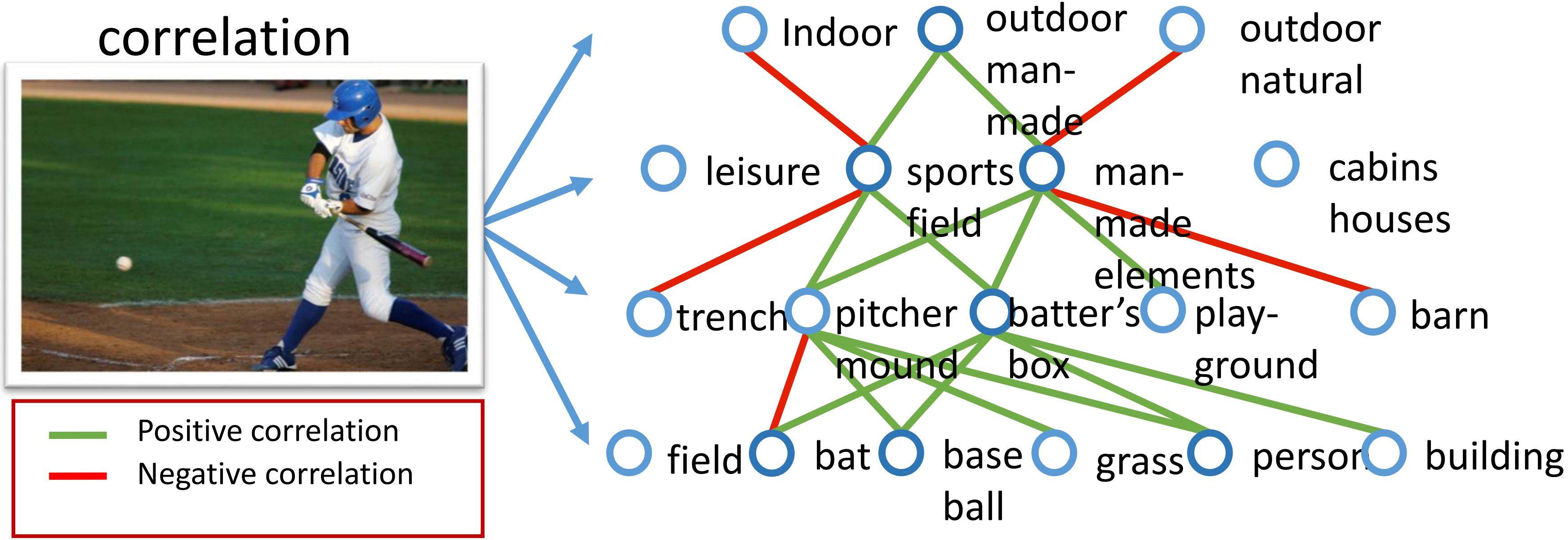
leisure | sports field | man-made elements | cabins houses

trench | pitcher mound | batter's box | play-ground | barn

field | bat | base ball | grass | person | building

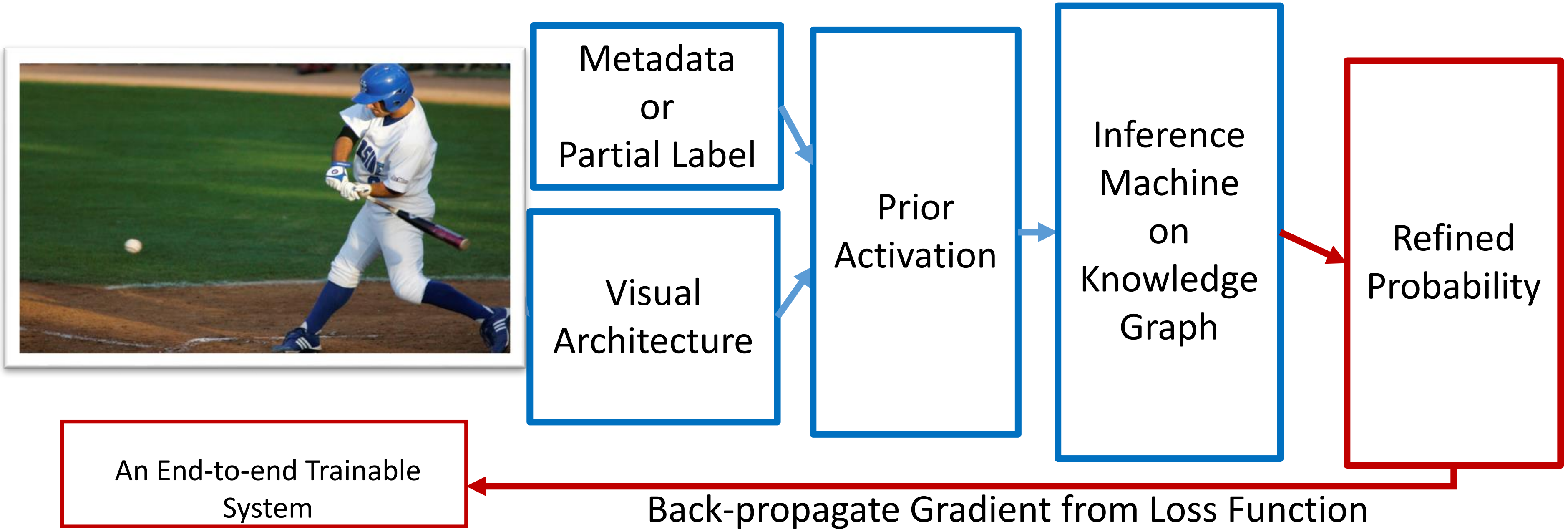
# Label Correlation Helps

- Such categorization at different concept layers can be modeled with label graphs
- It is natural and straightforward to leverage label correlation



# Goal: A generic label relation model

- Infer the entire label space from visual input
- Infer missing labels given a few fixed provided labels

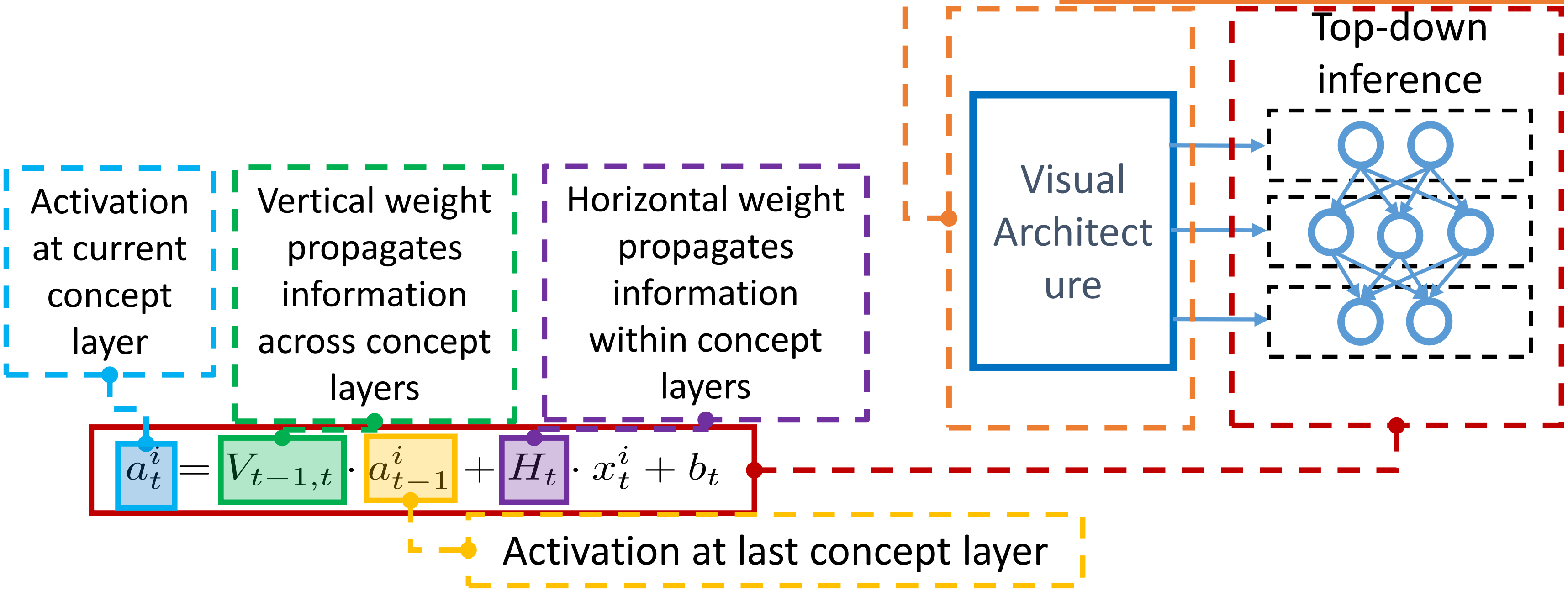


# Top-down Inference Neural Network

- Refine activations for each label
- Pass messages top-down and within each layer of label graph

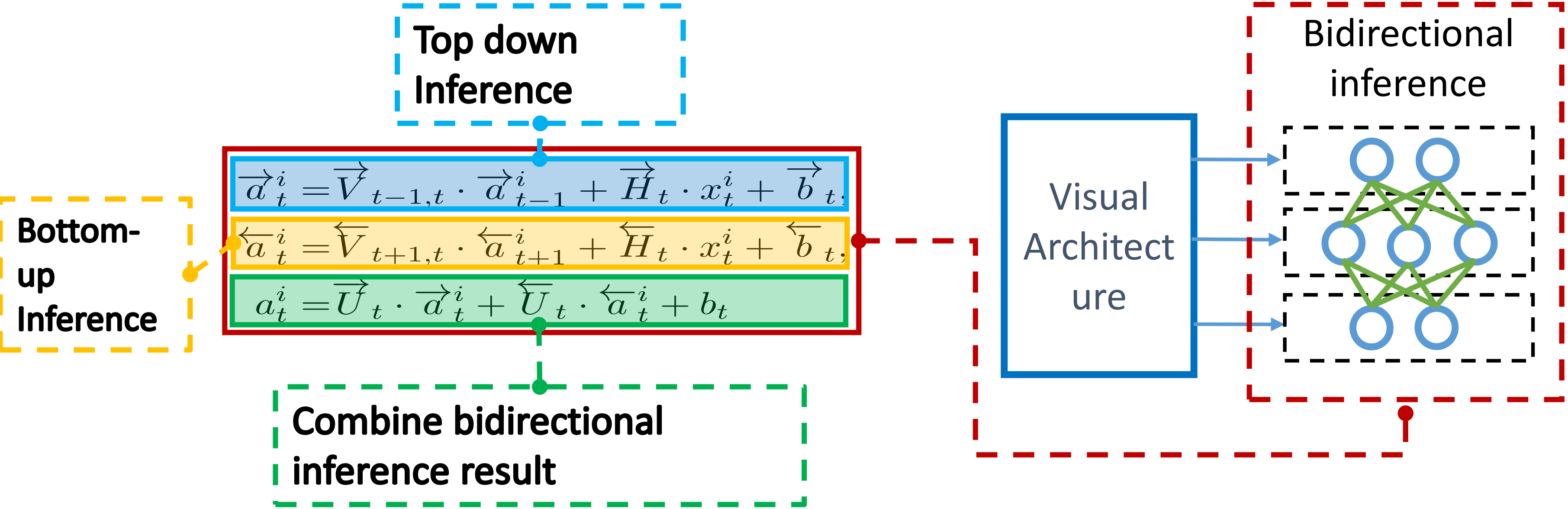
Produce Visual Prior activation from CNN

$$x_t^i = W_t \cdot CNN(I^i) + b_t$$



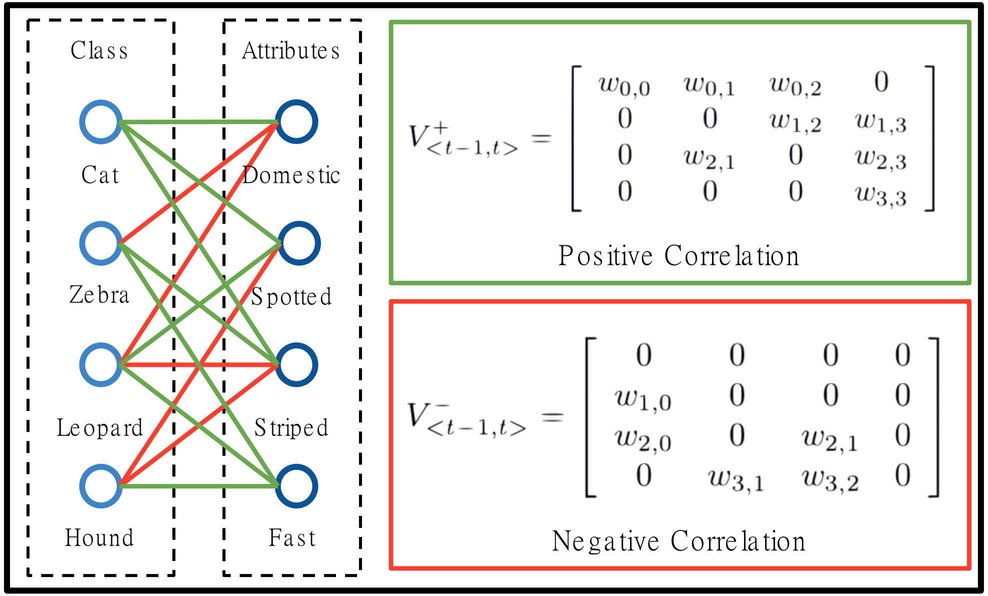
# Bidirectional Inference Neural Network (BINN)

- Bidirectional inference to make information propagate across entire label structure
- Inference in each direction independently and blend results



# Structured Inference Neural Network (SINN)

- BINN is **hard** to train well
- **Regularize** connections with prior knowledge about label correlations
- Decompose connections into **Positive correlation** + **Negative correlation**



# Structured Inference Neural Network (SINN)

- Evolve BINN formulation with regularization in connections

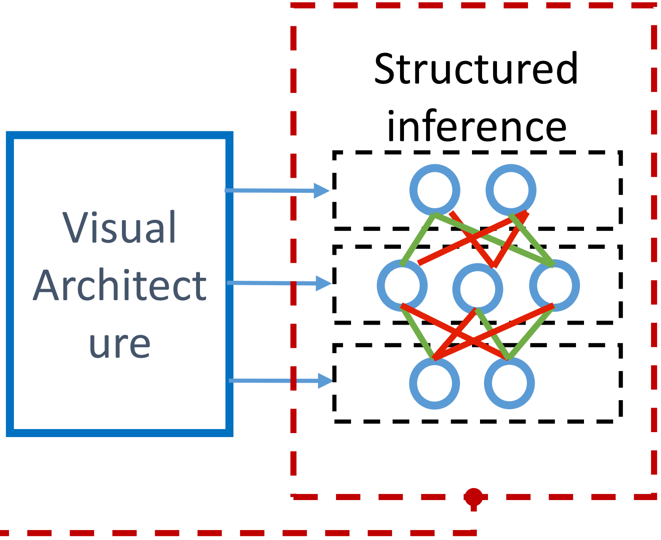
**Negative Component**

**Positive Component**

$$\overrightarrow{a}^i_t = \gamma(\overrightarrow{V}^+_{t-1,t} \cdot \overrightarrow{a}^i_{t-1}) + \gamma(\overrightarrow{H}^+_t \cdot x^i_t)$$
$$- \gamma(\overrightarrow{V}^-_{t-1,t} \cdot \overrightarrow{a}^i_{t-1}) - \gamma(\overrightarrow{H}^-_t \cdot x^i_t) + \overrightarrow{b}_t,$$
$$\overleftarrow{a}^i_t = \gamma(\overleftarrow{V}^+_{t+1,t} \cdot \overleftarrow{a}^i_{t+1}) + \gamma(\overleftarrow{H}^+_t \cdot x^i_t)$$
$$- \gamma(\overleftarrow{V}^-_{t+1,t} \cdot \overleftarrow{a}^i_{t+1}) - \gamma(\overleftarrow{H}^-_t \cdot x^i_t) + \overleftarrow{b}_t,$$
$$a^i_t = \overrightarrow{U}_t \cdot \overrightarrow{a}^i_t + \overleftarrow{U}_t \cdot \overleftarrow{a}^i_t + b_t$$

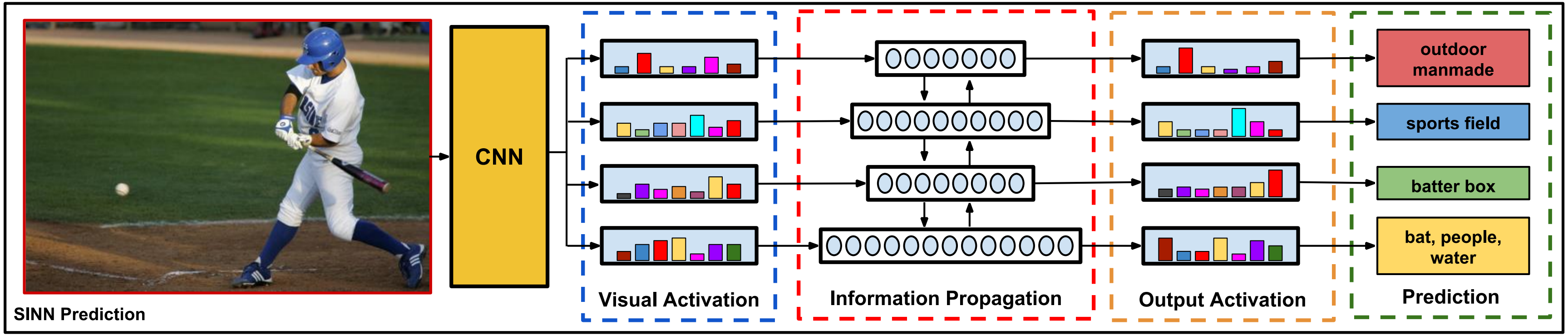
**ReLU** neuron is **essential** to keep positive/negative contribution

$\gamma(x) = ReLU(x)$



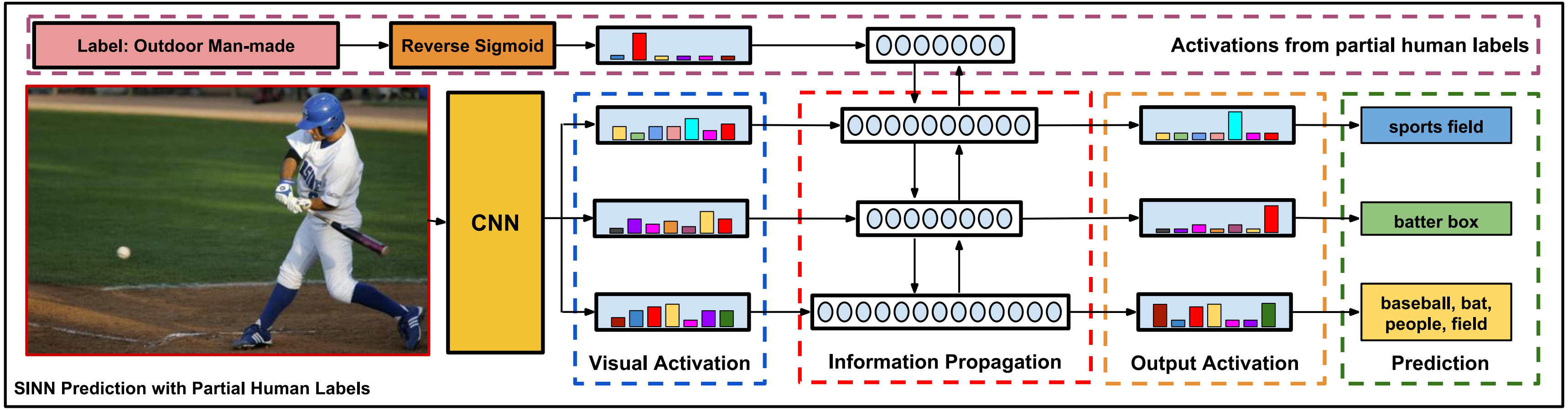
# Prediction from Purely Visual Input

- Visual architecture (e.g. Convolutional Neural Networks ) produces **visual activation**
- SINN implements **Information propagation** bidirectionally and produces refined **output activation**



# Prediction with Partially Observed Labels

- **Reverse Sigmoid** neuron produces activation from Partial labels
- **SINN** adapts both **visual activation** and **activation from partial labels** to infer the remaining labels



# Datasets

- Evaluate method with two types of experiments on three datasets

## Animals with Attributes

[Lampert et al. 2009]

<u>Labels</u>

28 taxonomy terms

50 animal classes

85 attributes

**Task**: predict entire label set

- Taxonomy terms are constructed from Word Net as [Hwang et al. 2012]
- Knowledge graph constructed by combining class-attributes graph with taxonomy graph

## NUS-WIDE

[Chua et al. 2009]

<u>Labels</u>

698 image groups

81 concepts

1000 tags

**Task**: predict 81 concepts with observing tags/image groups

- Knowledge graph produced by Word Net using <u>semantic similarity</u>
- 698 image groups constructed from image meta data

## SUN 397

[Xiao et al. 2012]


<u>Labels</u>

3 coarse

16 general

397 fine-grained

**Task 1**: predict entire label set

**Task 2**: predict fine-grained scene given coarse scene category

- Knowledge graph provided by dataset

# Ex2: Inference from partial labels (NUS-WIDE)

- Produce predictions given partial 1k tags and 698 image groups

**Ground Truth**: railroad
**CNN + Logistic**: statue buildings person
**Our Predictions**: railroad person sky

**Ground Truth**: animal grass water dog
**CNN + Logistic**: grass person animal
**Our Predictions**: water animal dog

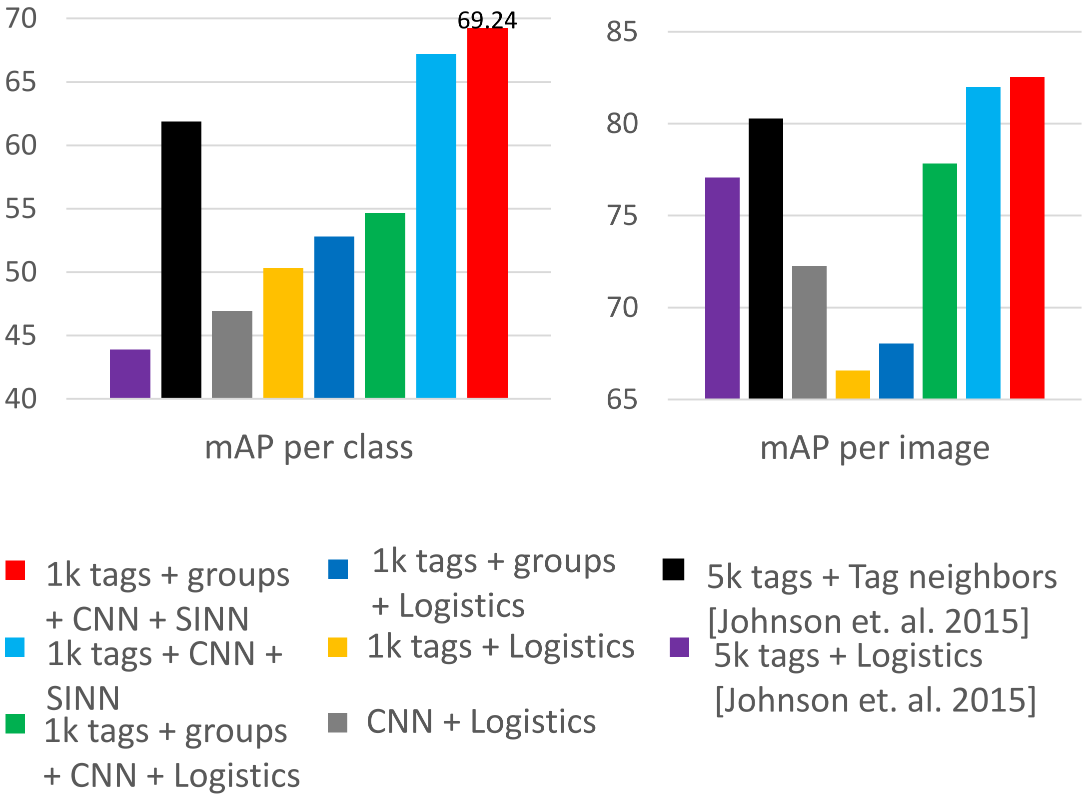
**Ground Truth**: rainbow clouds sky
**CNN + Logistic**: clouds water sky
**Our Predictions**: rainbow clouds sky

**Ground Truth**: food water
**CNN + Logistic**: food plants flower
**Our Predictions**: food plants water

Correct predictions are marked in **blue** while incorrect are marked in **red**

# Ex2: Inference from partial labels (NUS-WIDE)

- Evaluate on standard 81 ground truth classes of NUSWIDE
- **Outperform all baselines by large margin**

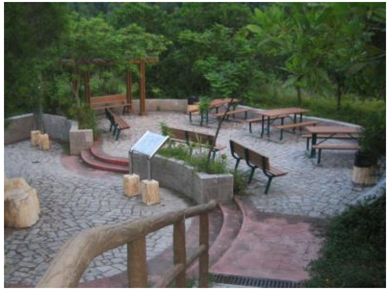


# Ex2: Inference with partial labels (SUN397)

- Produce predictions given coarse-level labels (3 coarse categories)

**CNN + Logistic**: campus
**Observed Label**: outdoor/man-made
**Our Predictions**: abbey
**Ground Truth**: abbey

**CNN + Logistic**: building facade
**Observed Label:** outdoor/man-made
**Our Predictions**: library/outdoor
**Ground Truth**: library/outdoor



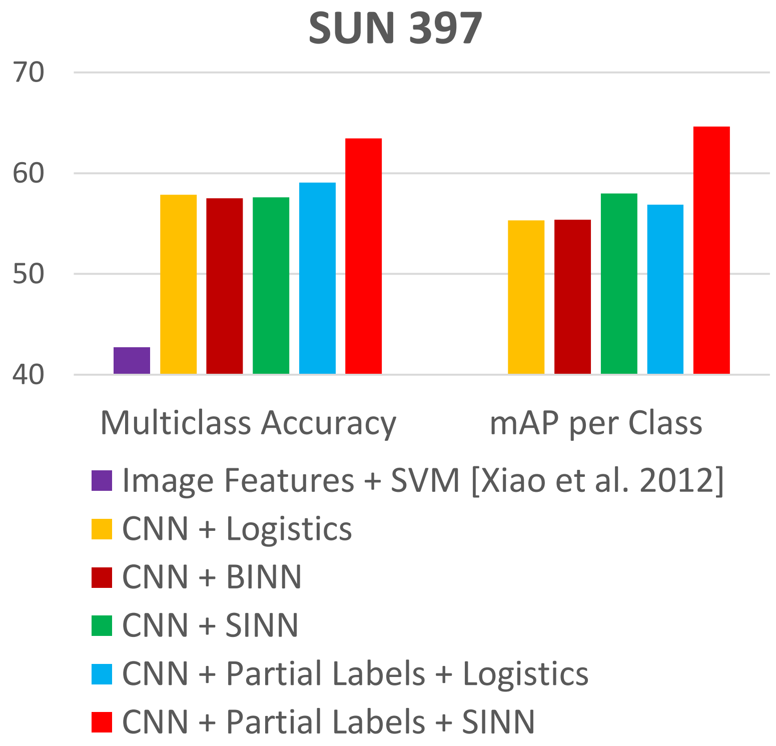
**CNN + Logistic**: patio
**Observed Label:** outdoor/natural; outdoor/man-made
**Our Predictions**: picnic area
**Ground Truth**: picnic area

**CNN + Logistic**: operating room
**Observed Label:** indoor
**Our Predictions**: dentists office
**Ground Truth**: dentists office

Correct predictions are marked in **blue** while incorrect are marked in **red**

# Ex2: Inference with partial labels (SUN397)

- Evaluate on 397 fine-grained scene categories
- **Significantly improved performance**



# Summary

- Inference in structured label space
- Relations within and across levels of a label space
- Model positive and negative correlations between labels in end-to-end trainable model

# Conclusion

- Methods for handling *structures* in deep networks
  - Spatial structure: learning gating functions to connect people for group activity recognition [Deng, Vahdat, Hu, Mori CVPR 2016]
  - Temporal structure: hierarchies of long short-term memory models for group activities [Ibrahim, Muralidharan, Deng, Vahdat, Mori CVPR 2016]
  - Label structure: message passing algorithms for multi-level image labeling; purely from image data or with partial labels [Hu, Zhou, Deng, Liao, Mori CVPR 2016]

# Acknowledgement