ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Context Aware Group Activity Recognition

Avijit Dasgupta[1]

C. V. Jawahar[1]

Karteek Alahari[2]

[1] CVIT, IIIT Hyderabad, India

[2] THOTH, Inria, France

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

CVIT

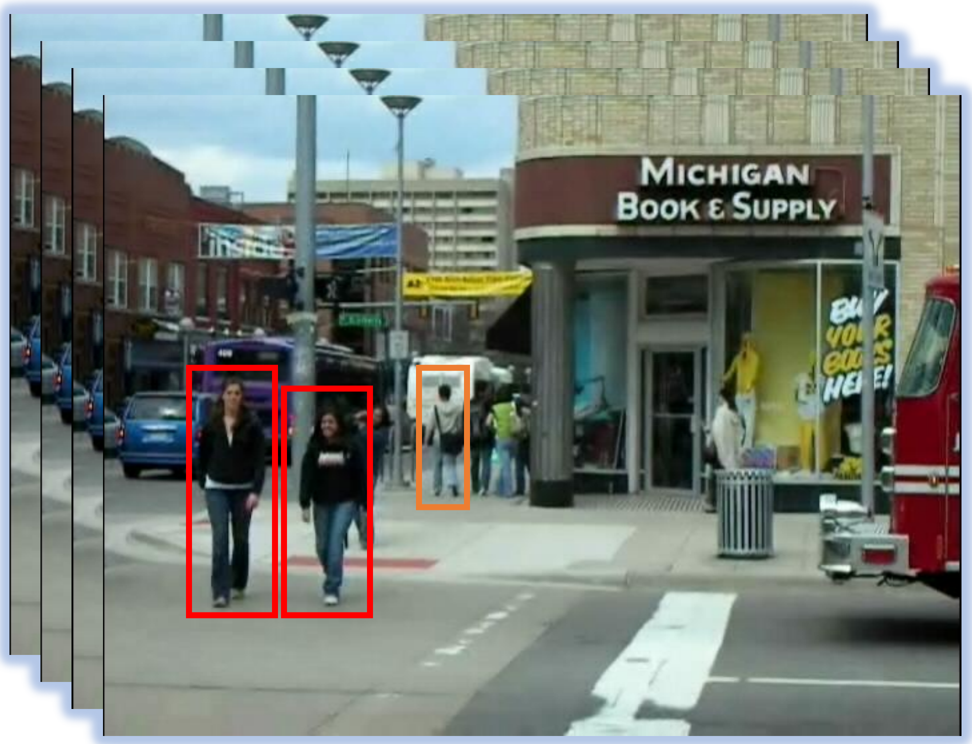
Inria

Task: Group Activity Recognition

Input Video

Task: Group Activity Recognition

- Predict individual activities and group activities



Input Video

Individual Activities

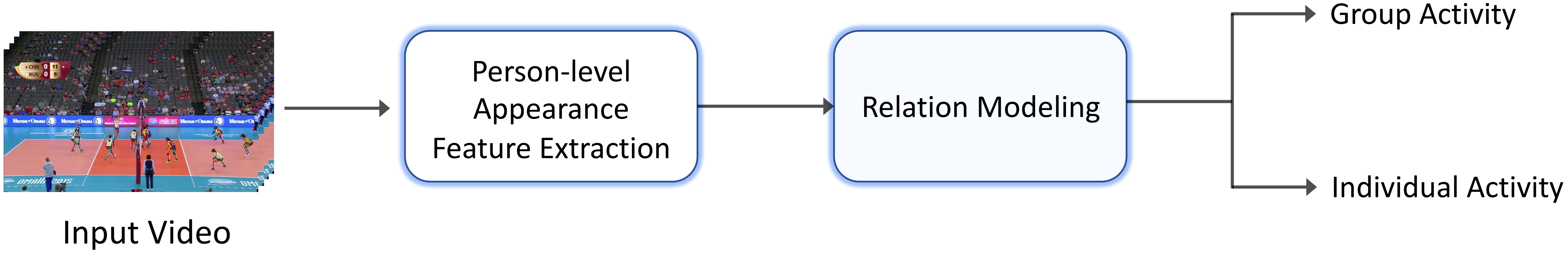
🟥 Crossing

🟧 Walking

Group Activity

Crossing

Typical Pipeline for Group Activity Recognition



Input Video

Person-level Appearance Feature Extraction

Relation Modeling

Group Activity

Individual Activity

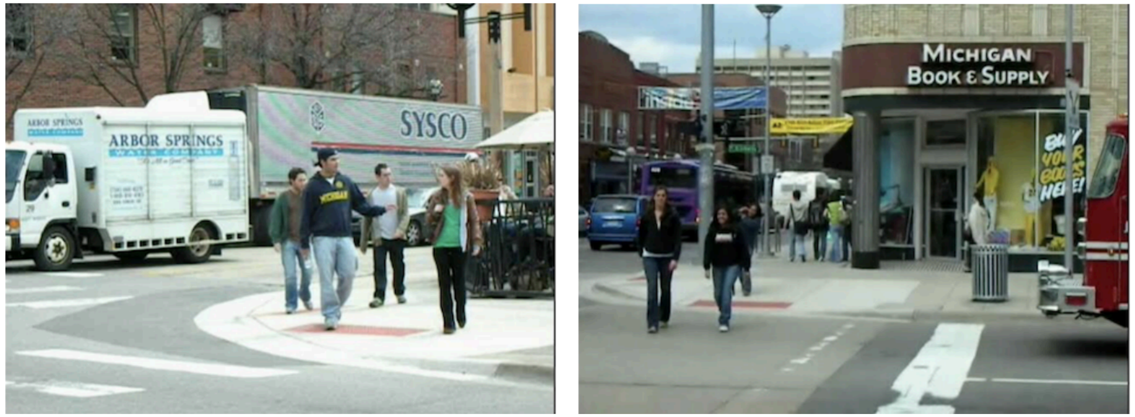
AlexNet – Ibrahim *et al.*, CVPR’16
VGG16 – Shu *et al.*, CVPR’17
Inception-V3 – Wu *et al.*, CVPR’19
I3D – Azar *et al.,* CVPR’19

Hierarchical LSTM – Ibrahim *et al.*, CVPR’16
GCN – Wu *et al.*, CVPR’19
Transformers – Gavrilyuk *et al.*, CVPR’20

Context Aware Group Activity Recognition

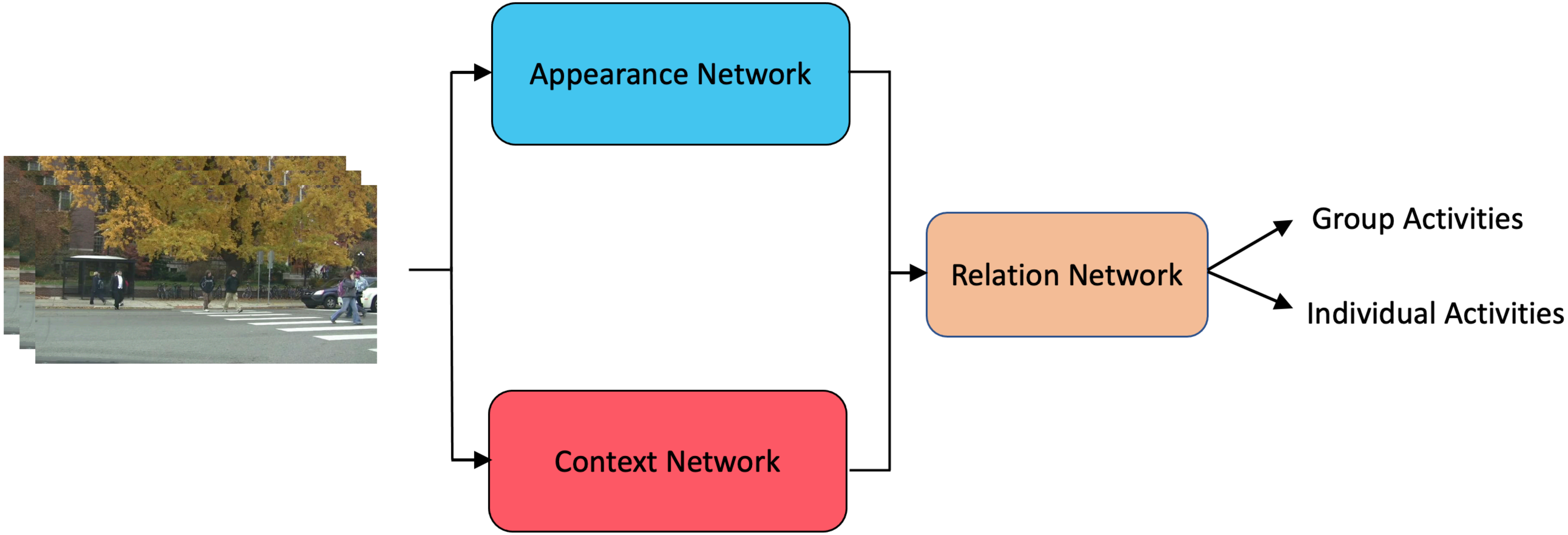
In this paper, we argue –

- Person-level appearance only features unable to distinguish between visually similar activities



- Context provides important cues about the environment (e.g. ***sidewalk*** vs. ***road***) to differentiate between visually similar (e.g. ***walking*** vs. ***crossing***) activities.

The Proposed Solution



Appearance Network

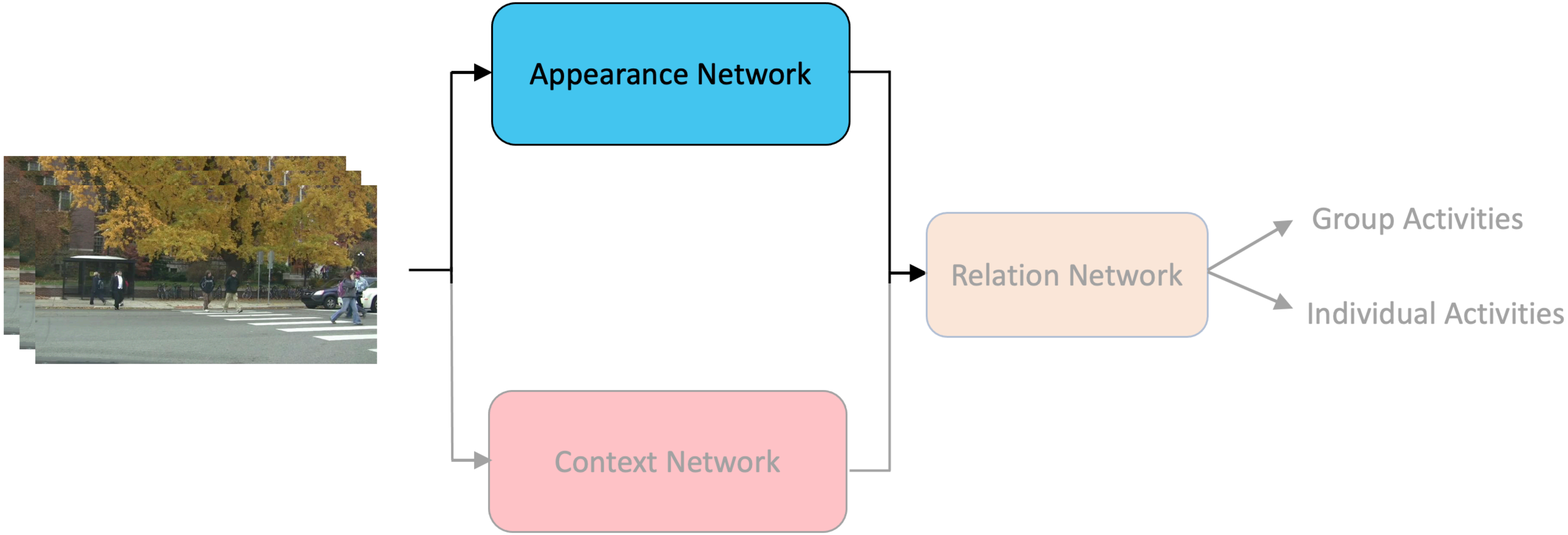
Context Network

Relation Network

Group Activities

Individual Activities

The Proposed Solution



Appearance Network

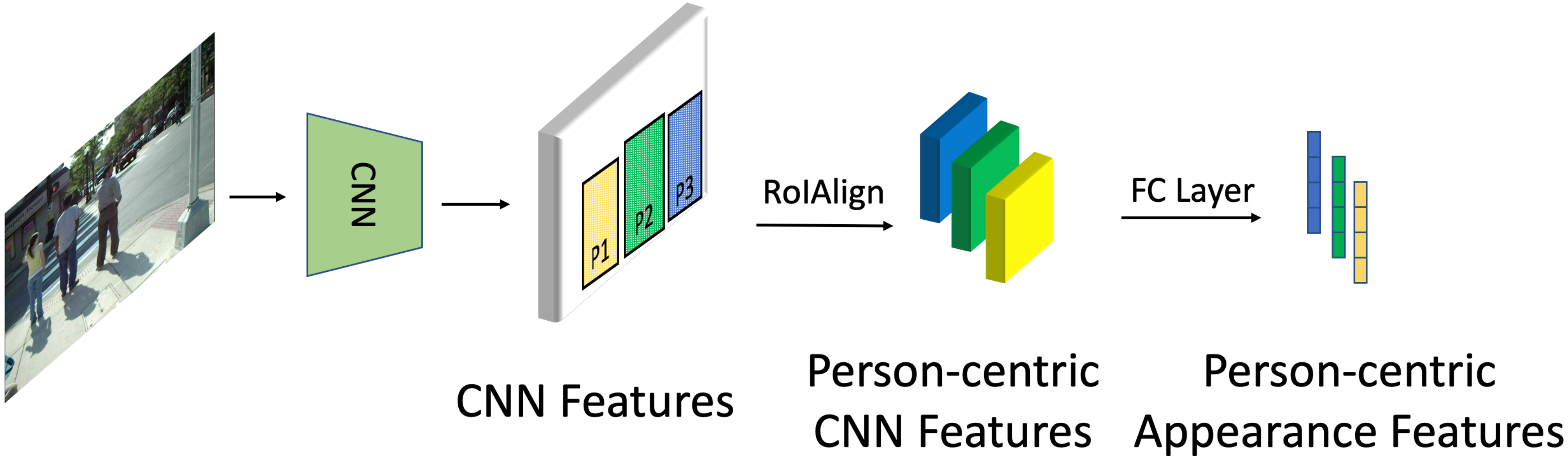
Context Network

Relation Network

Group Activities

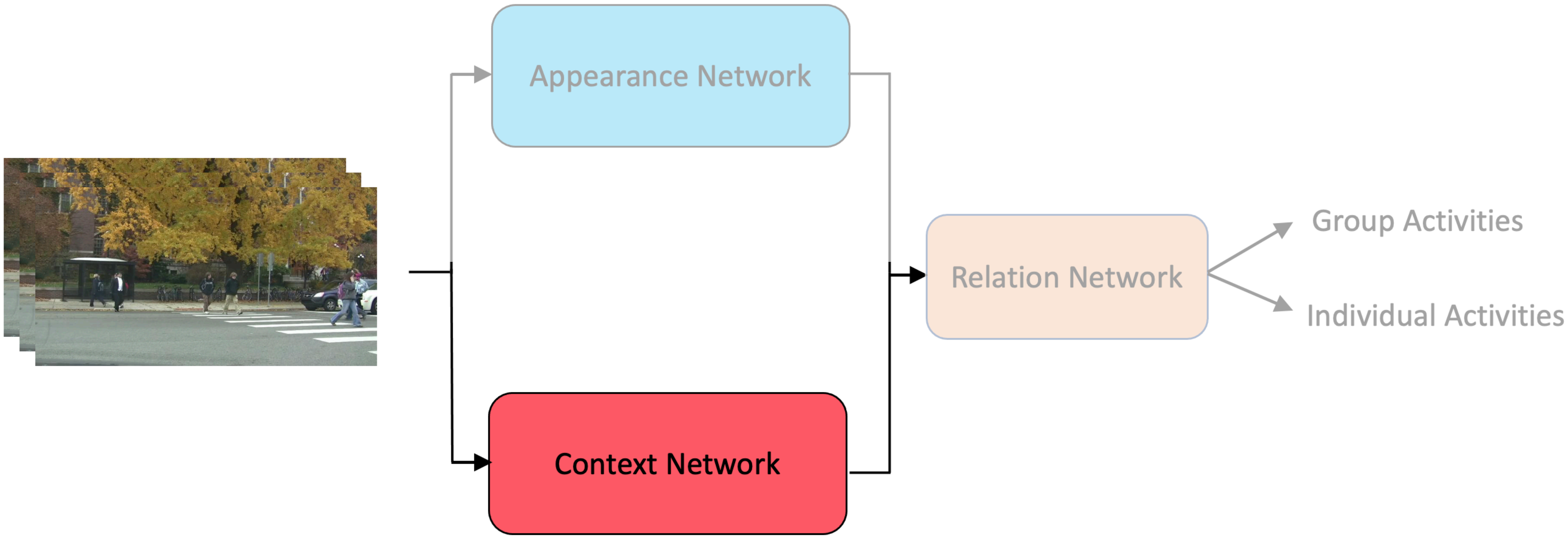
Individual Activities

The Appearance Network

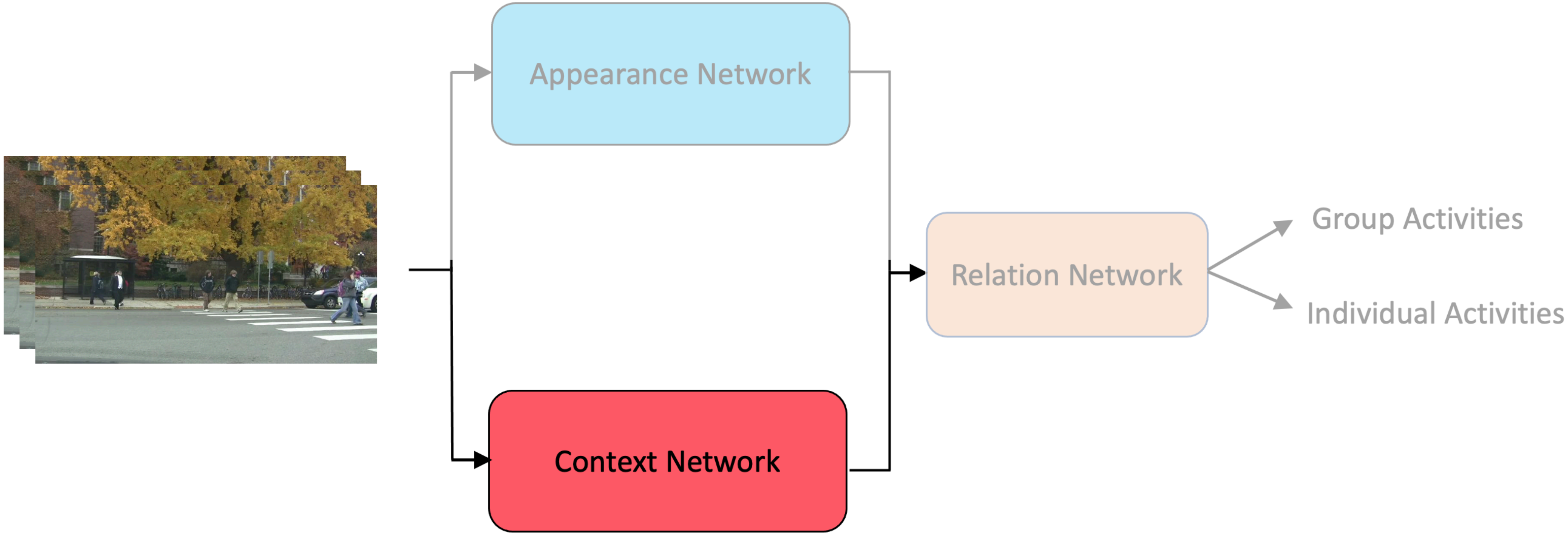


CNN
P1
P2
P3
RoIAlign
FC Layer
CNN Features
Person-centric CNN Features
Person-centric Appearance Features

The Proposed Solution



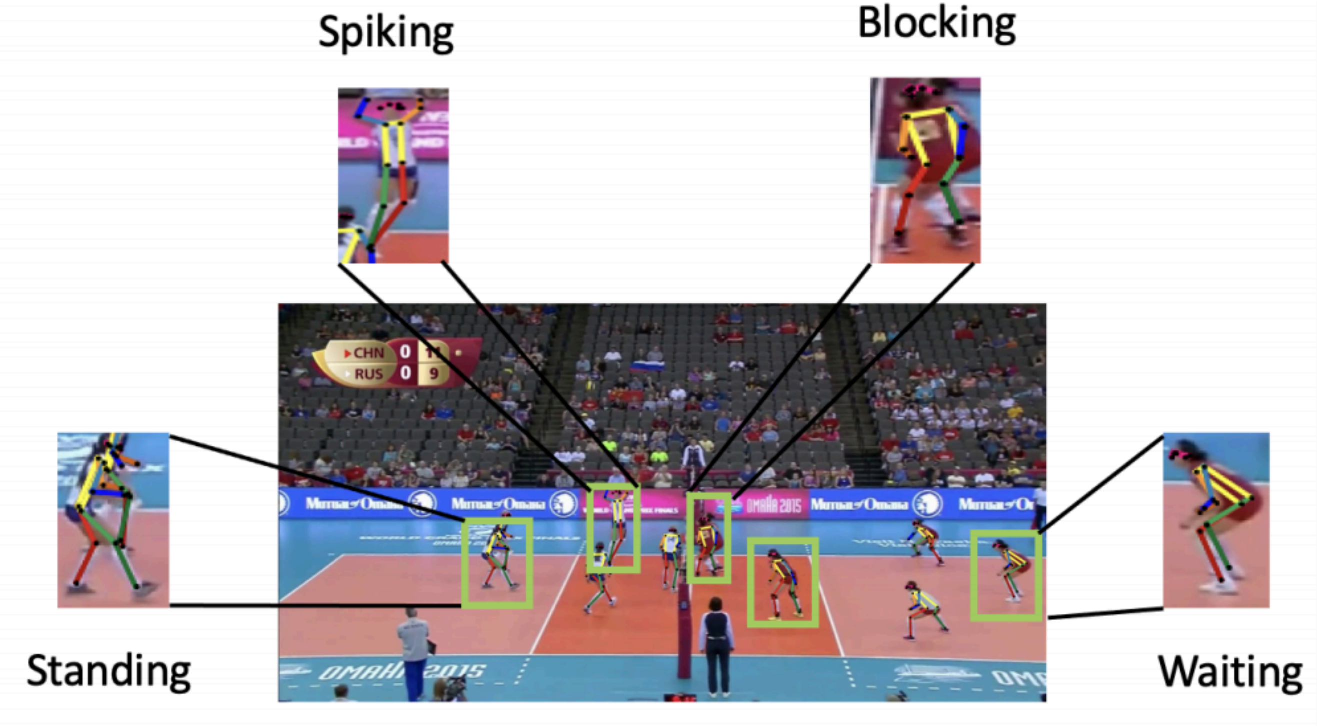
The Proposed Solution



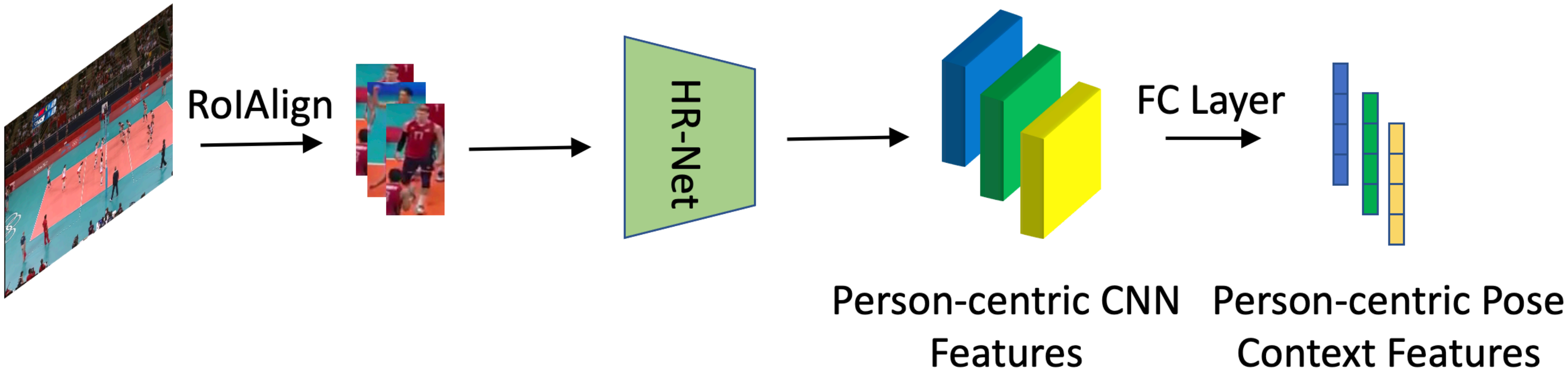
- Pose
- Scene Labels

The Pose Contextual Cues

Posture provide important cues about different activities



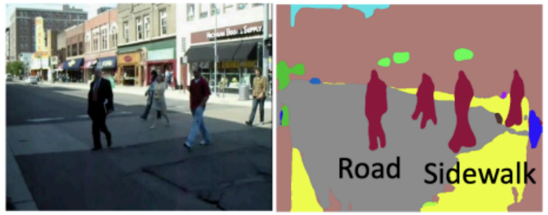
The Pose Context Network


RoIAlign
HR-Net
FC Layer
Person-centric CNN Features
Person-centric Pose Context Features

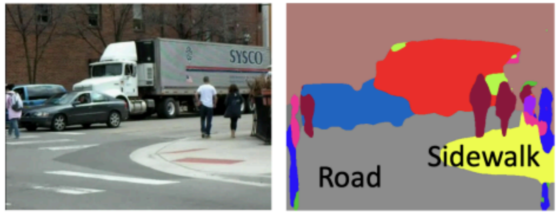
Ke *et al.*, “Deep High-Resolution Representation Learning for Human Pose Estimation”, CVPR’19

The Scene Contextual Cues

Scene labels important cues about the environment

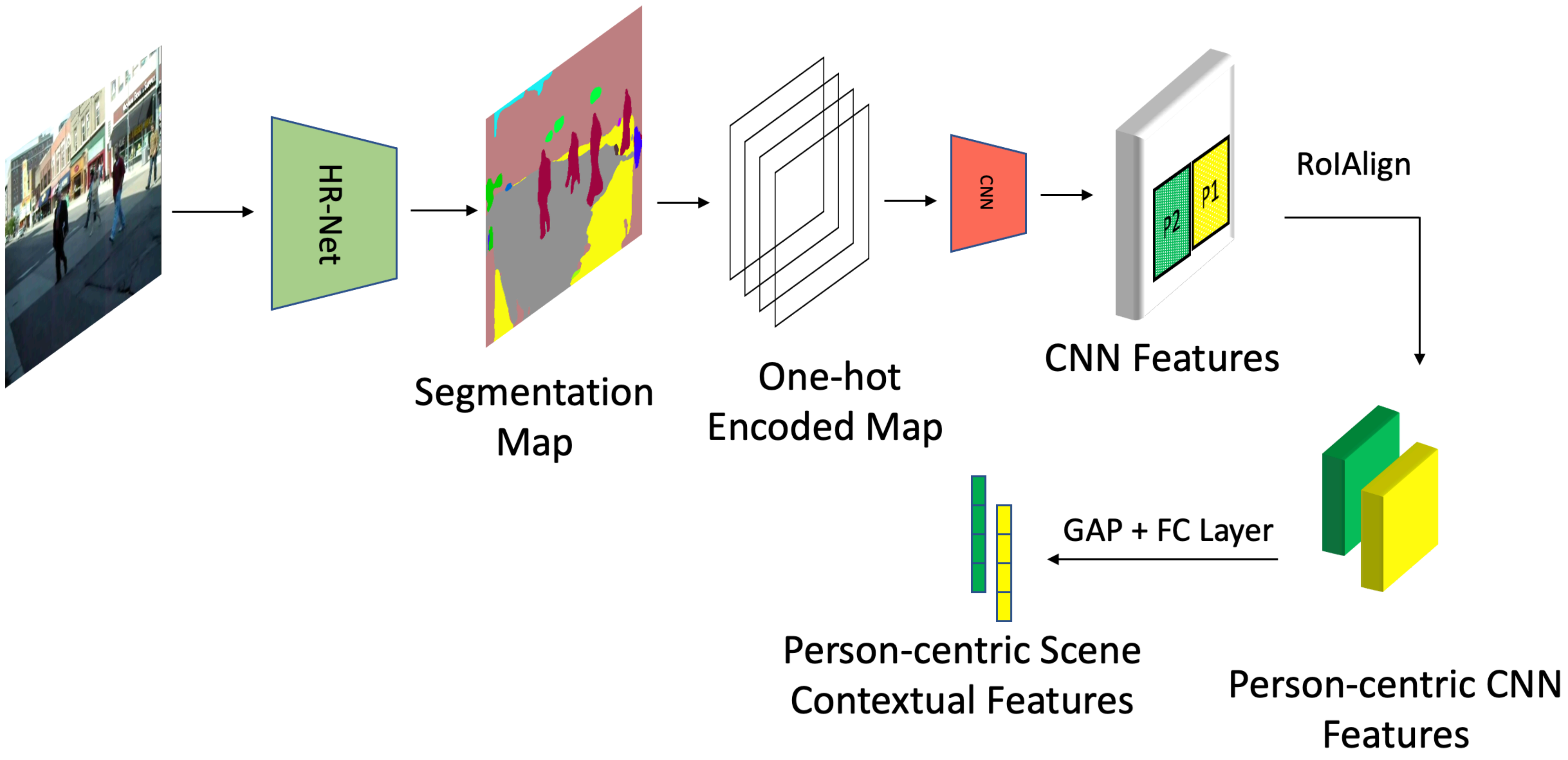


(a) Crossing activity



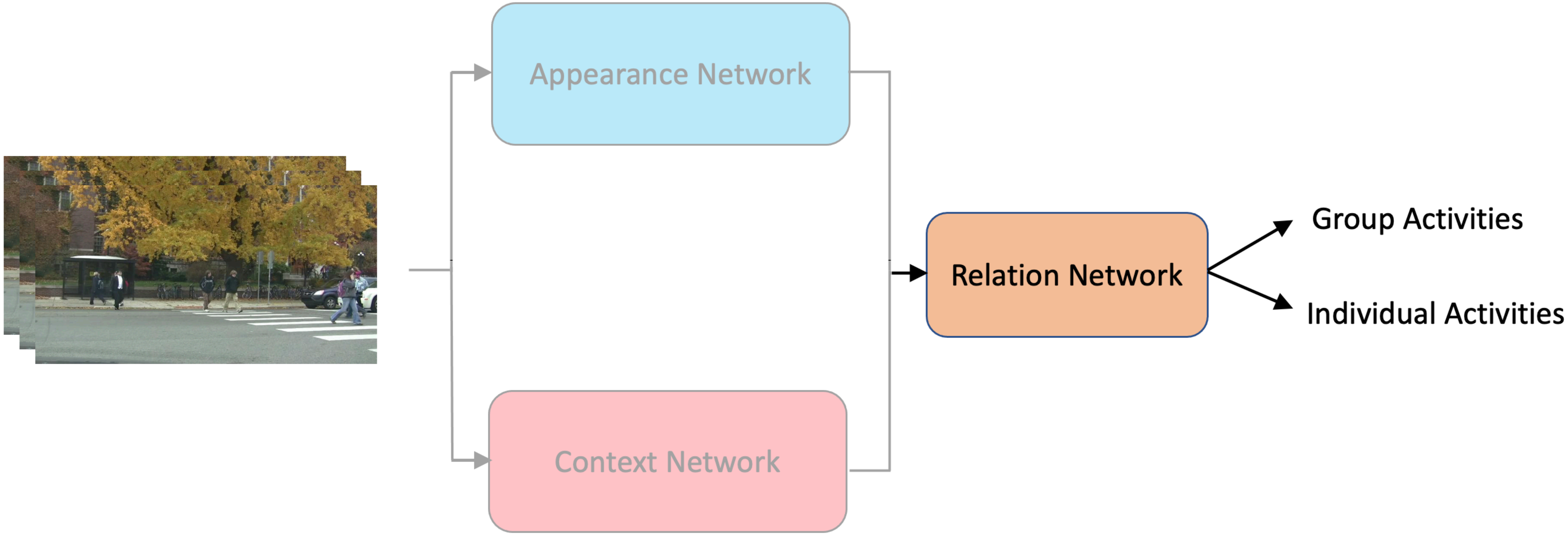
(b) Walking activity

The Scene Context Network



Wang et al., “Deep High-Resolution Representation Learning for Visual Recognition”, PAMI’19

The Proposed Solution



Appearance Network

Context Network

Relation Network

Group Activities

Individual Activities

Dataset

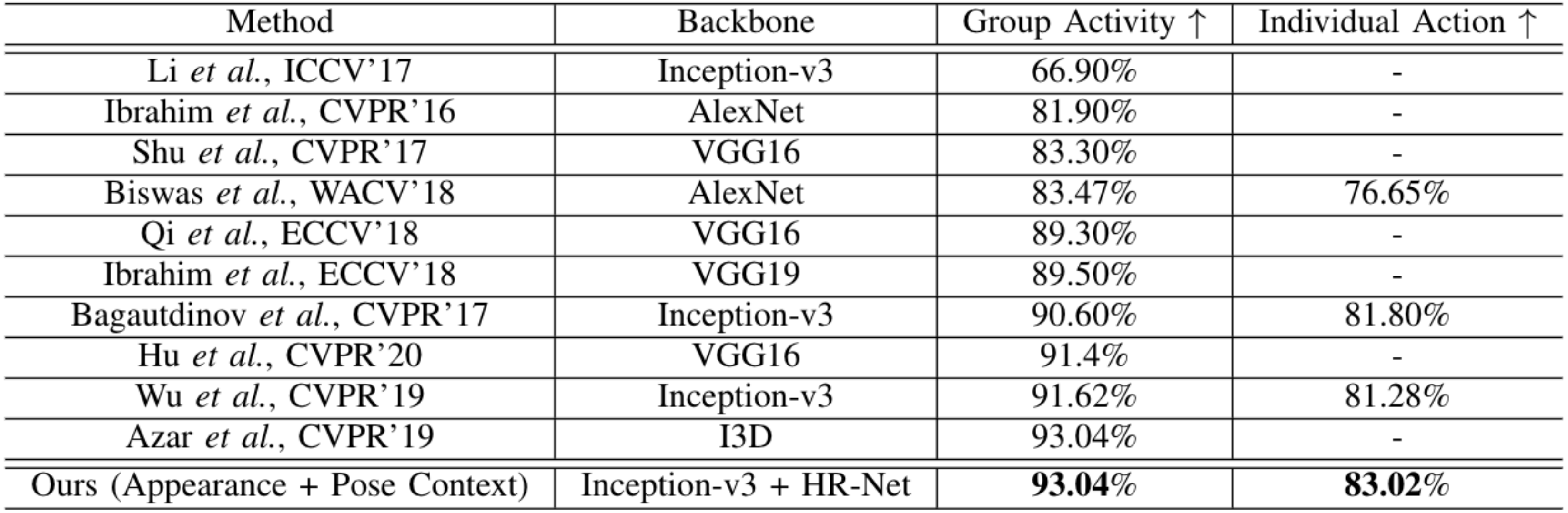
We use two publicly available datasets for experimental analysis –

- Volleyball Dataset
  - contains 4830 clips of 55 volleyball sports videos
  - 9 individual actions and 8 group activities

- Collective Activity Dataset
  - clips from 44 videos
  - 6 individual actions and 5 group activities

Experimental Results

Comparison with State-of-the-arts on Volleyball Dataset -

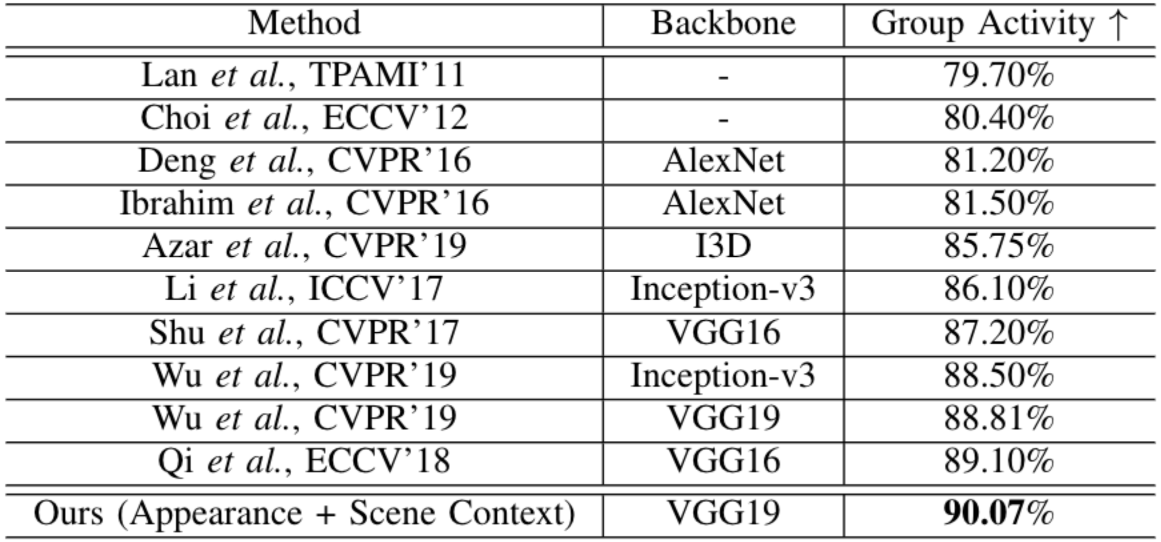


Method	Backbone	Group Activity ↑	Individual Action ↑
Li *et al.*, ICCV'17	Inception-v3	66.90%	-
Ibrahim *et al.*, CVPR'16	AlexNet	81.90%	-
Shu *et al.*, CVPR'17	VGG16	83.30%	-
Biswas *et al.*, WACV'18	AlexNet	83.47%	76.65%
Qi *et al.*, ECCV'18	VGG16	89.30%	-
Ibrahim *et al.*, ECCV'18	VGG19	89.50%	-
Bagautdinov *et al.*, CVPR'17	Inception-v3	90.60%	81.80%
Hu *et al.*, CVPR'20	VGG16	91.4%	-
Wu *et al.*, CVPR'19	Inception-v3	91.62%	81.28%
Azar *et al.*, CVPR'19	I3D	93.04%	-
Ours (Appearance + Pose Context)	Inception-v3 + HR-Net	**93.04**%	**83.02**%

Ibrahim et al., "A Hierarchical Deep Temporal Model for Group Activity Recognition", CVPR'16

Experimental Results

Comparison with State-of-the-arts on Collective Dataset -



Method	Backbone	Group Activity ↑
Lan *et al.*, TPAMI'11	-	79.70%
Choi *et al.*, ECCV'12	-	80.40%
Deng *et al.*, CVPR'16	AlexNet	81.20%
Ibrahim *et al.*, CVPR'16	AlexNet	81.50%
Azar *et al.*, CVPR'19	I3D	85.75%
Li *et al.*, ICCV'17	Inception-v3	86.10%
Shu *et al.*, CVPR'17	VGG16	87.20%
Wu *et al.*, CVPR'19	Inception-v3	88.50%
Wu *et al.*, CVPR'19	VGG19	88.81%
Qi *et al.*, ECCV'18	VGG16	89.10%
Ours (Appearance + Scene Context)	VGG19	**90.07**%

Choi et al., "What are they doing? : Collective Activity Classification Using Spatio-Temporal Relationship Among People", ICCV'09

Summary

- Context is important for group activity recognition

- Two types of contextual cues are proposed –
  - Pose
  - Scene labels

- The effectiveness of context is validated on two datasets showing improvements over appearance only features

Thank You!

Avijit Dasgupta[1]

C. V. Jawahar[1]

Karteek Alahari[2]

[1] CVIT, IIIT Hyderabad, India

[2] THOTH, Inria, France