Mizzou
University of Missouri - Columbia

LEARNING HUMAN POSES IN NATURAL SCENES

Guanghan Ning

Advisor: Zhihai He

EECS Department

Email: gnxr9@mail.missouri.edu

Focus of This Talk

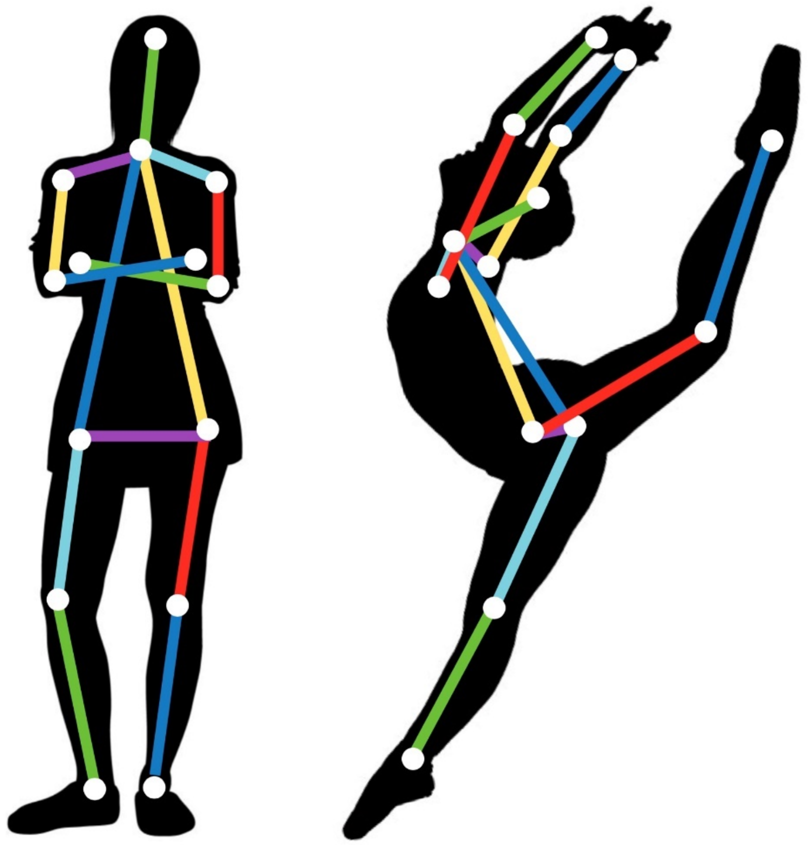
- ❑ Introduction
- ❑ Summarize Previous Work
  - Human Detection and Tracking
  - Single-person Human Pose Estimation
  - Human Pose Estimation with Adversarial Training
  - Multi-person Human Pose Estimation: PoseTrack Challenge

Mizzou
University of Missouri - Columbia

Chapter. 1
Introduction

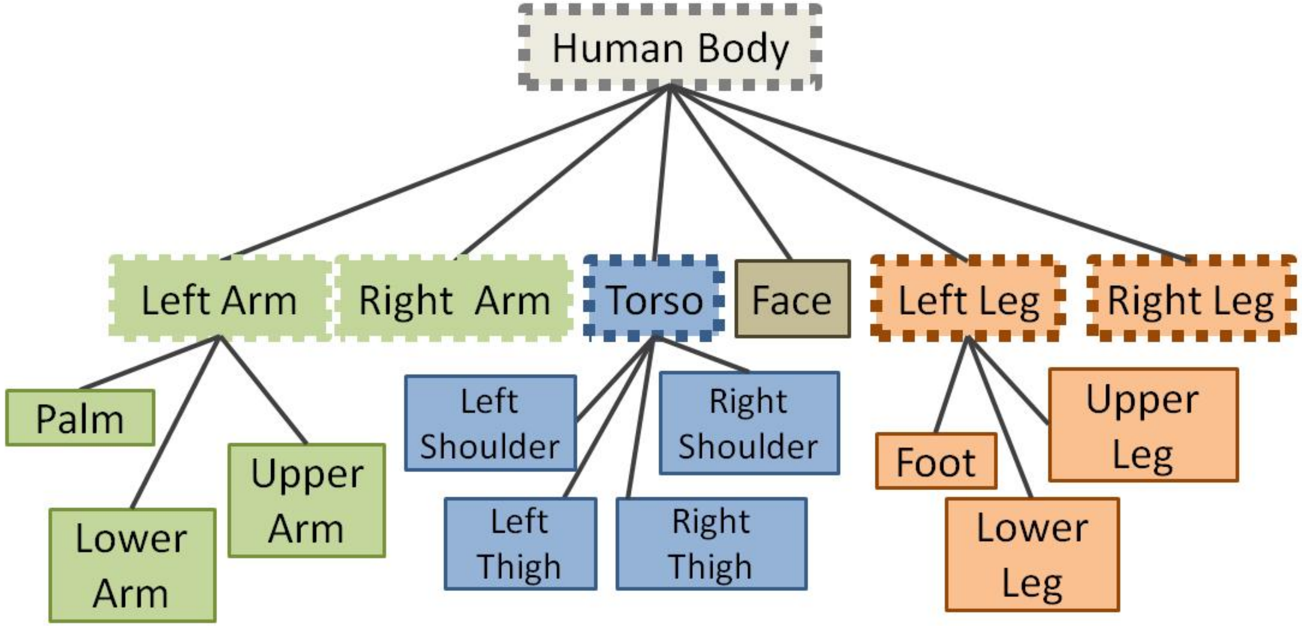
Introduction

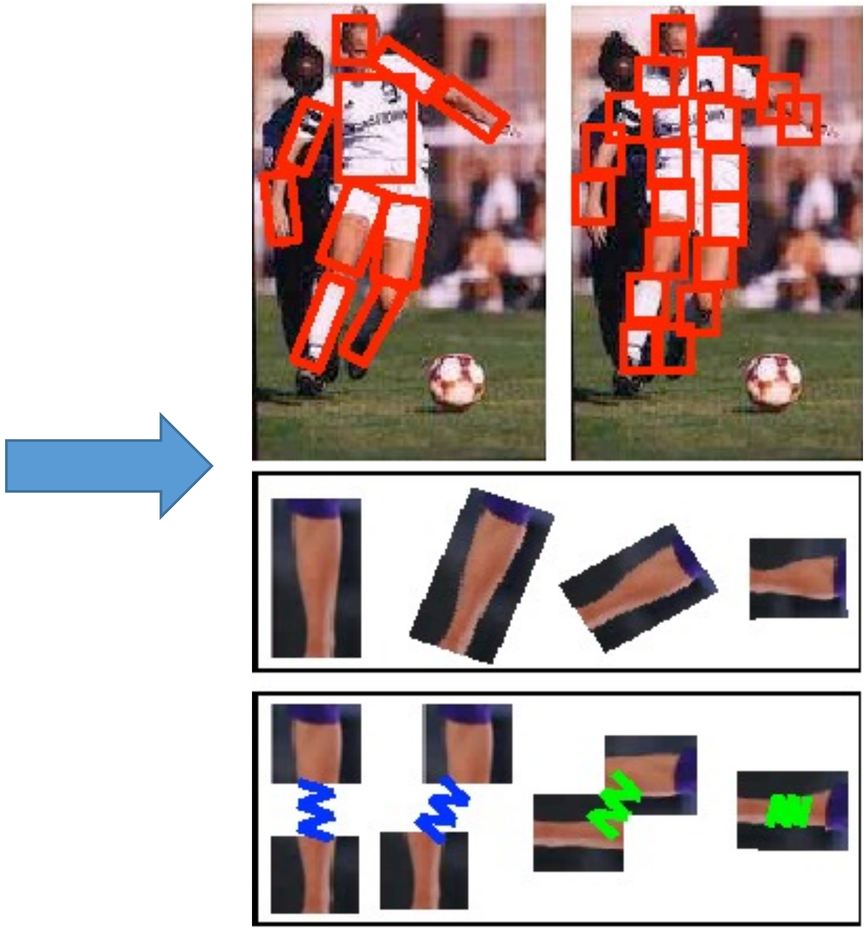
- ❑ Sports Video Analytics
- ❑ Video Surveillance
- ❑ Activity Recognition
- ❑ Human-Computer Interaction
- ❑ etc…



Introduction

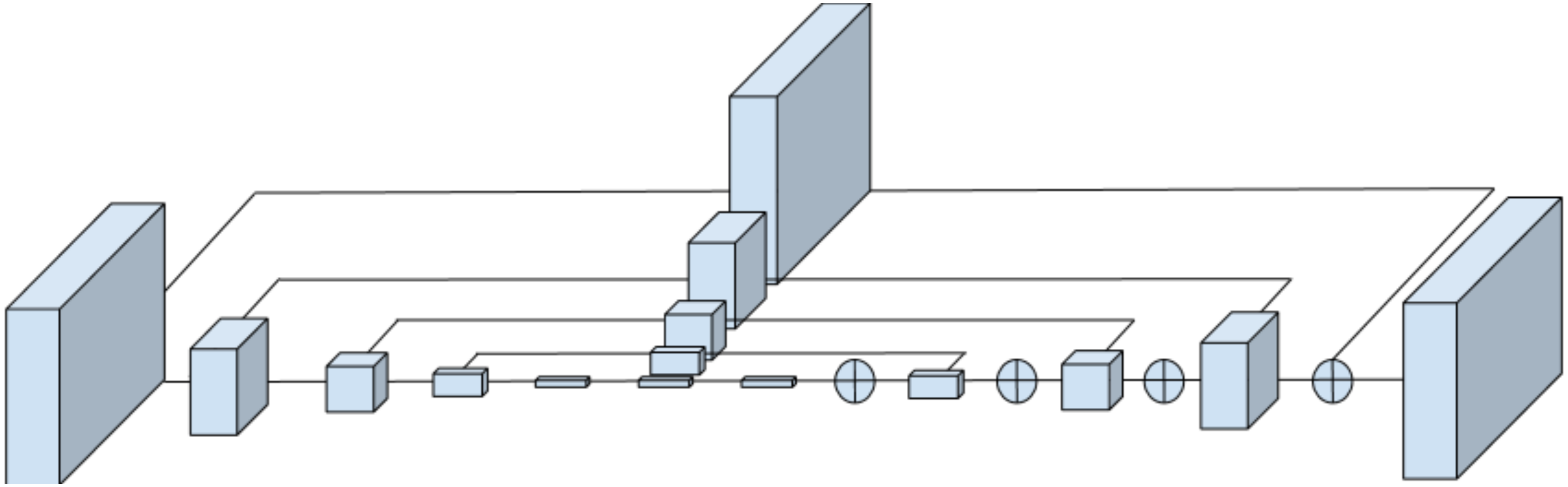
- Early Works: Pictorial Structure Models





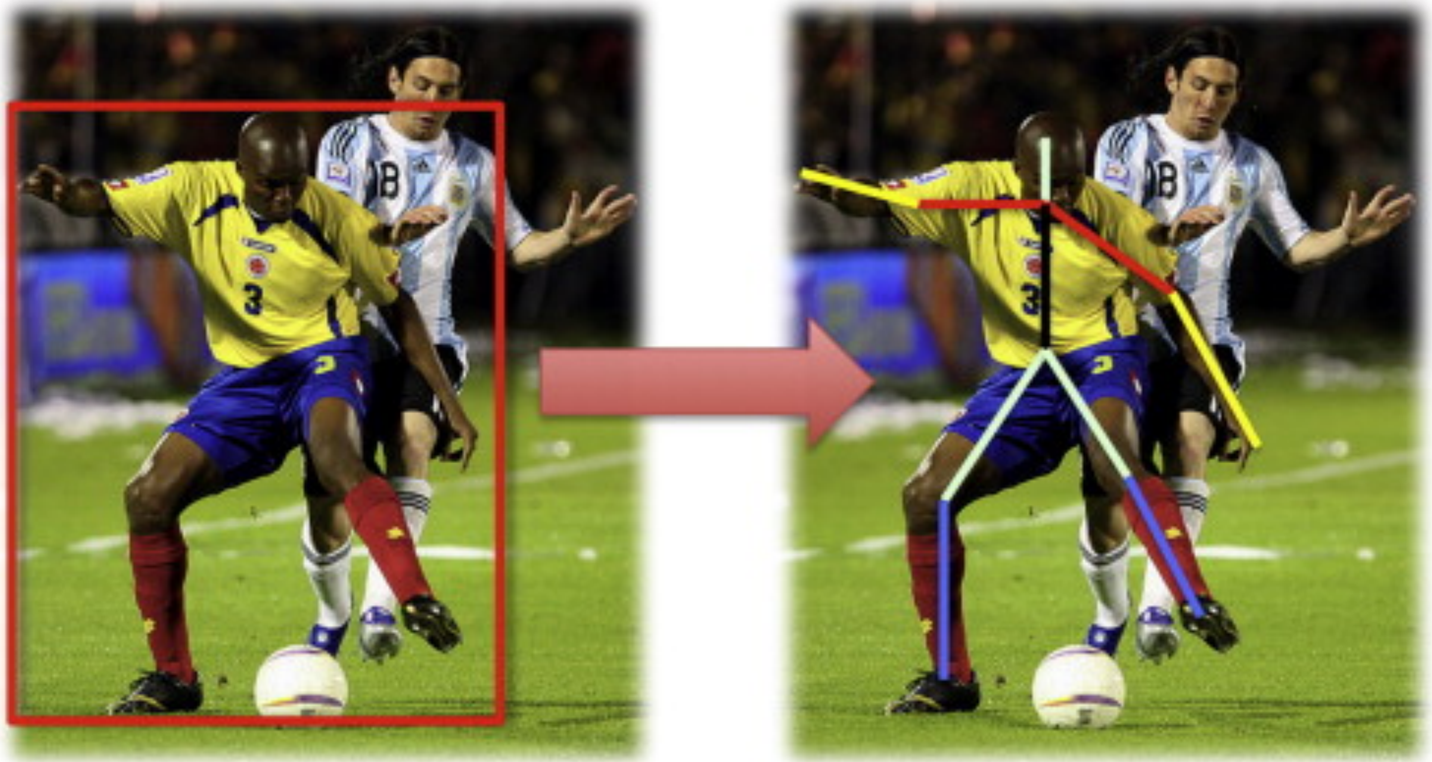
Introduction

- Recent Works: Convolutional Neural Networks



Introduction

❑ Top-down Approach



Detect persons　　　　Estimating poses

Introduction

- ❑ Convolutional Neural Networks
- ❑ Generic Object Detection
- ❑ From Proposal To Regression
- ❑ Human Pose Estimation

Introduction

❑ Convolutional Neural Networks

Motivations

- Do not need careful hand-design
- Allow a machine to automatically discover the representations needed for specific task
- Do not need domain expertise
- Great feature representation capacity

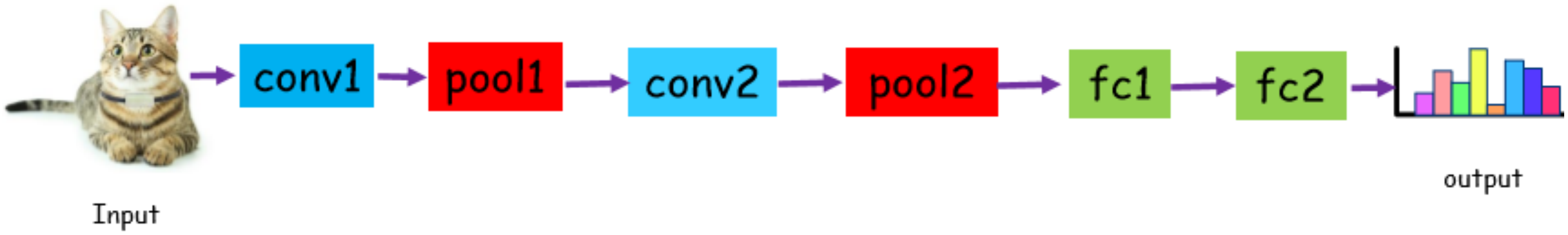
Introduction

❑ Convolutional Neural Networks

Basic Building Blocks

- Convolutional Layer
- Pooling Layer
- Fully-connected Layer

Basic Architecture



Introduction

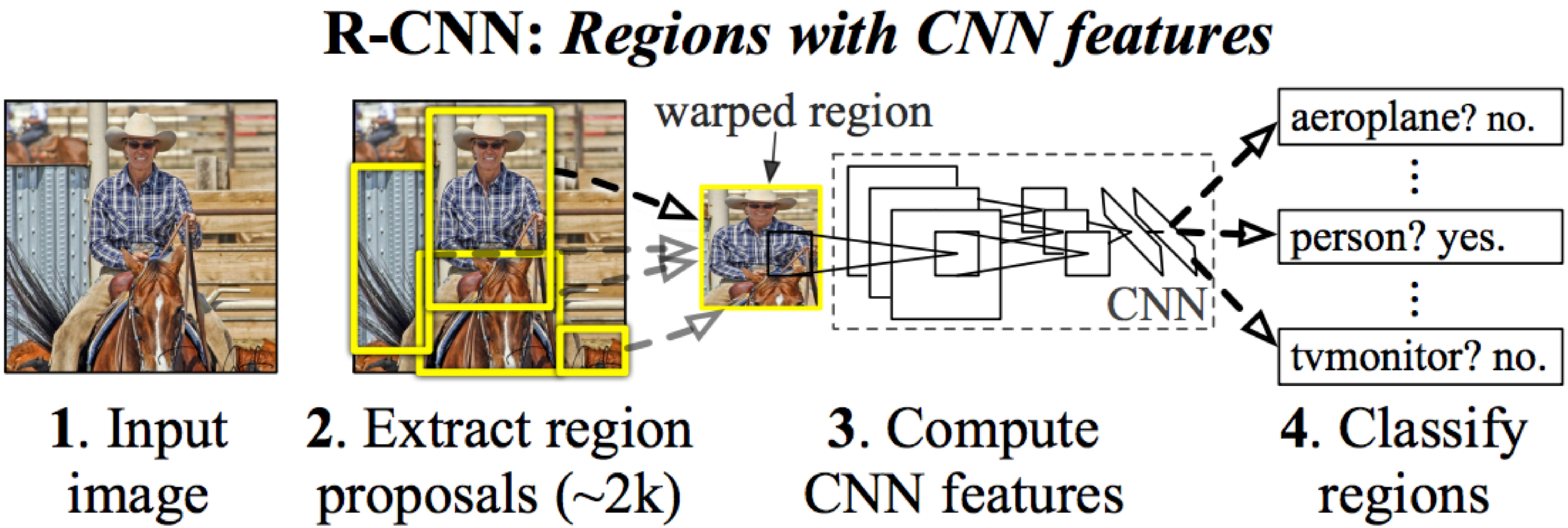
❑ Convolutional Neural Networks

- Inference / Forward

- Learning / Training

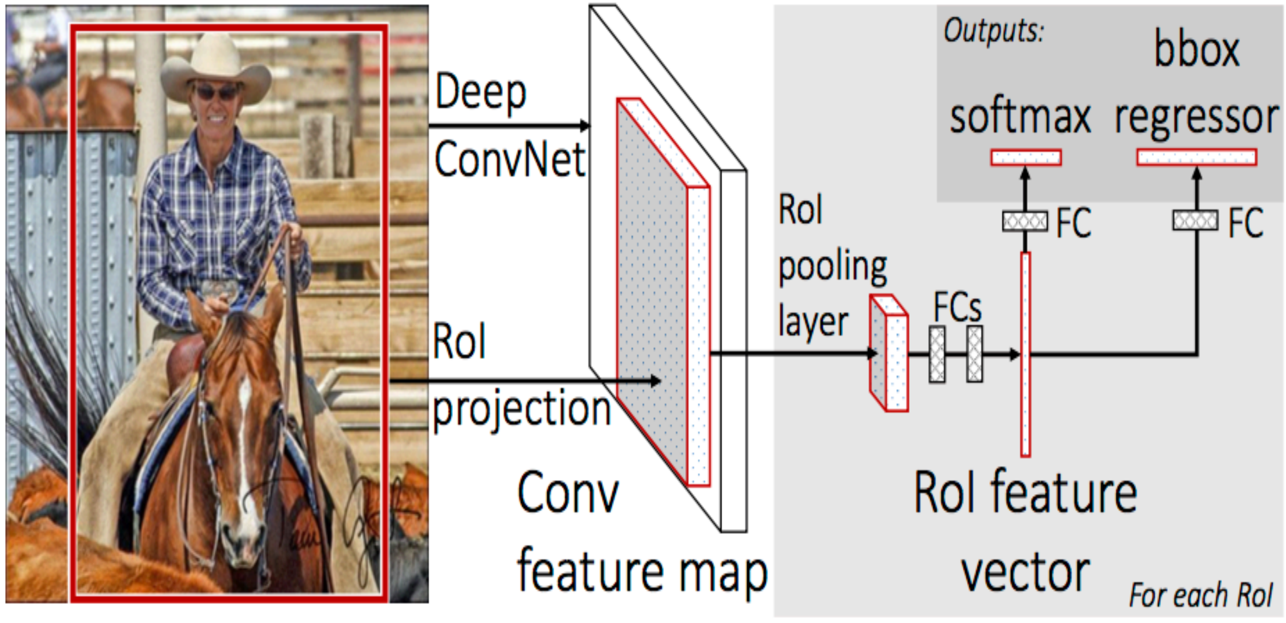
Introduction

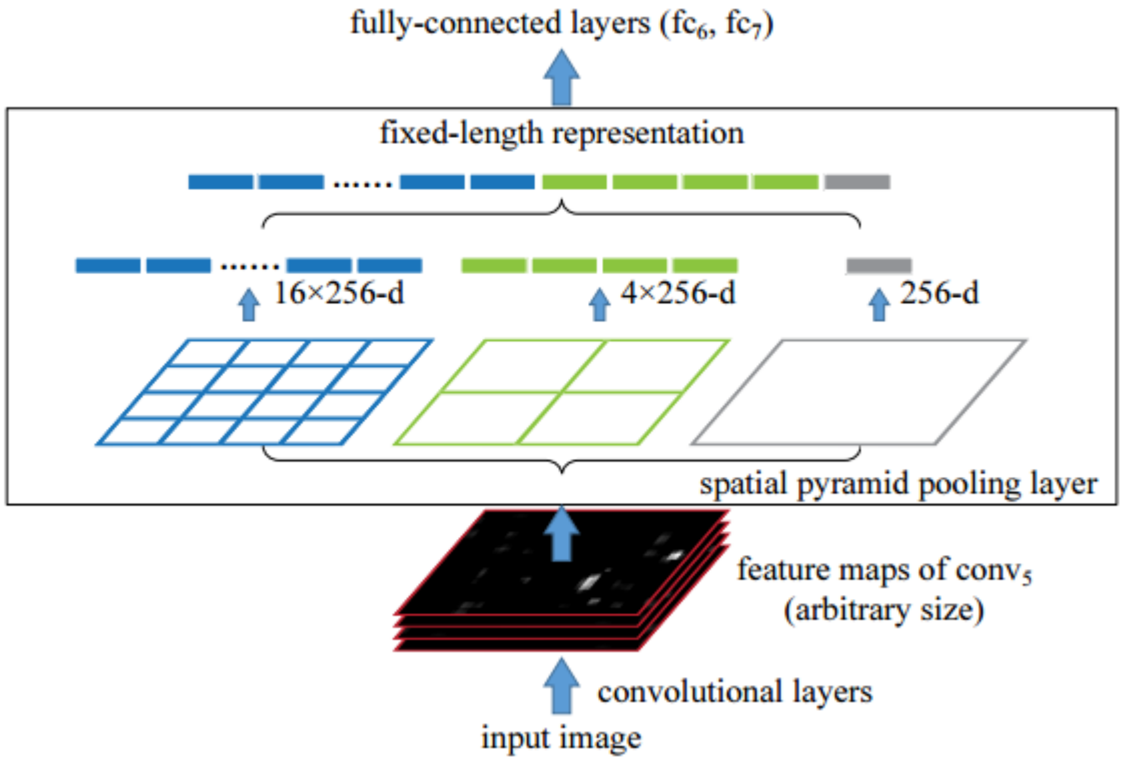
❑ Generic Object Detection



Introduction

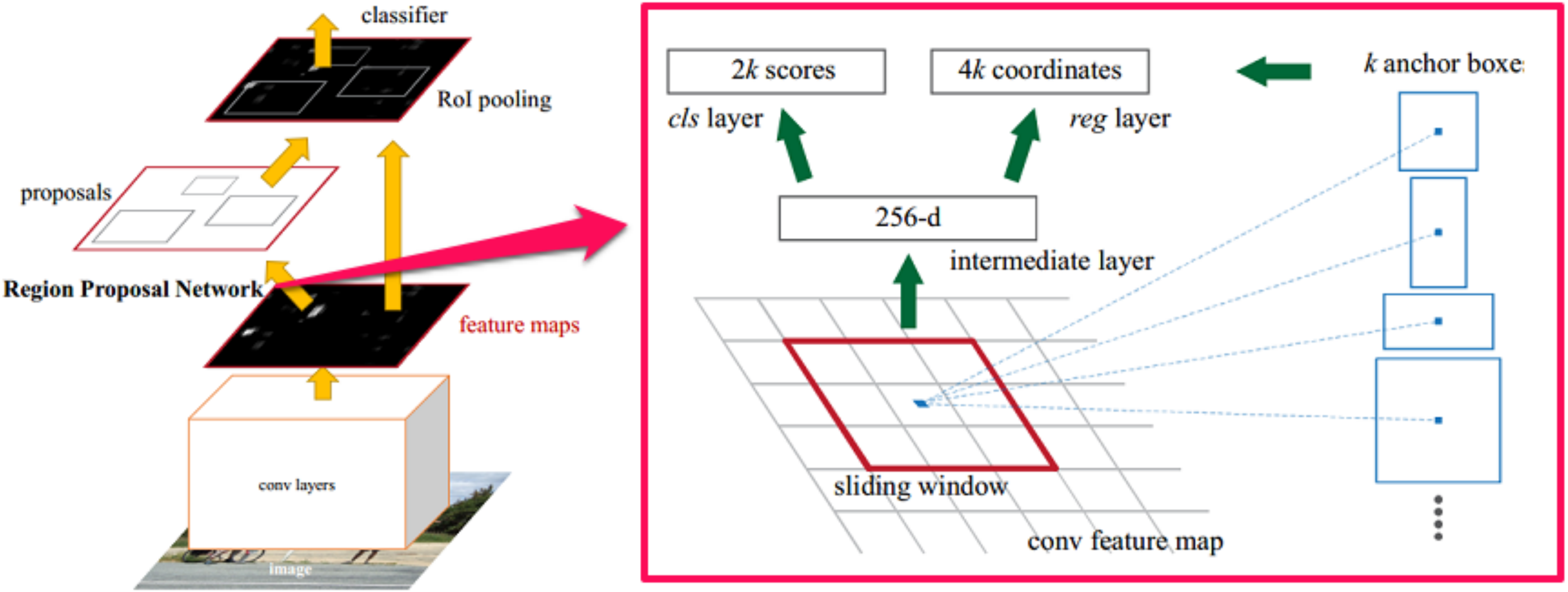
❑ Generic Object Detection





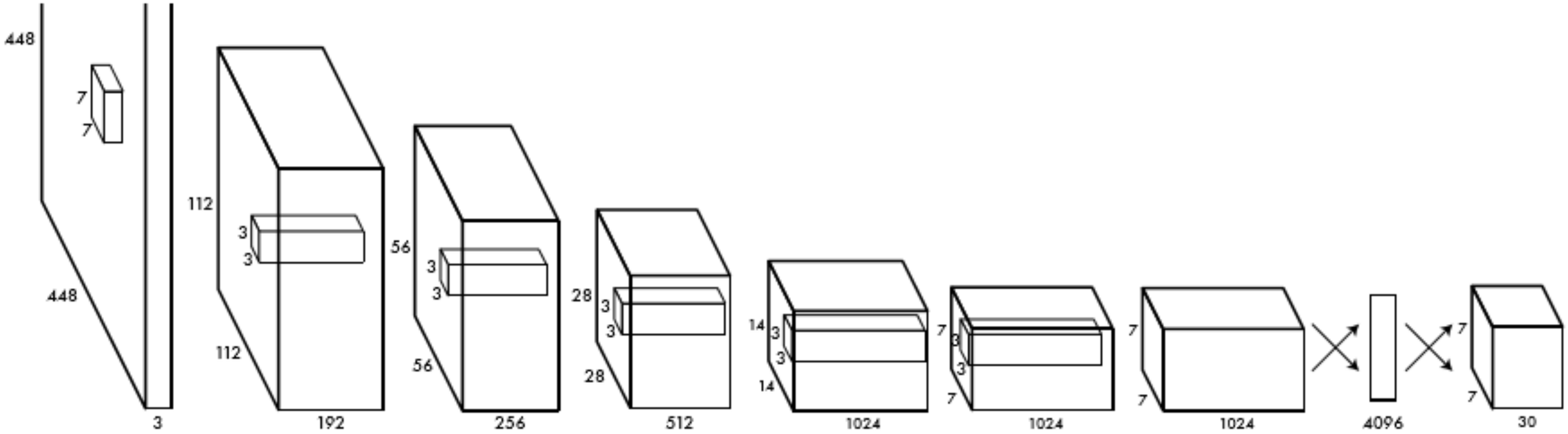
Introduction

❑ Generic Object Detection



Introduction

- From Proposal to Regression
  - Region Proposal v.s Unified Network
    (Classification v.s Regression)
    (R-CNN variants v.s YOLO/SSD)



Introduction

- Human Pose Estimation
  - Naturally a regression problem
  - Coordinates v.s Heatmaps

Mizzou
University of Missouri - Columbia

Chapter. 2
Human Detection and Tracking

Human Detection and Tracking

❑Problem Definition

- Visual Object Tracking： the process of localizing a single target in a video or sequential images, given the target position in the first frame.

Human Detection and Tracking

- Significance

  - It has a wide range of applications such as motion analysis, activity recognition, surveillance, and human-computer interaction.

  - It can be a prerequisite or a necessary component of another system.

Human Detection and Tracking

- **Appearance Variations:**
  - Target deformations
  - Fast and abrupt motion
  - Scale changes
  - Background Clutters

- **2. Occlusion**

- **3. Difficulties Introduced by Camera**
  - Uneven lighting, Illumination
  - Blur, Low resolution
  - Perspective distortion

Human Detection and Tracking

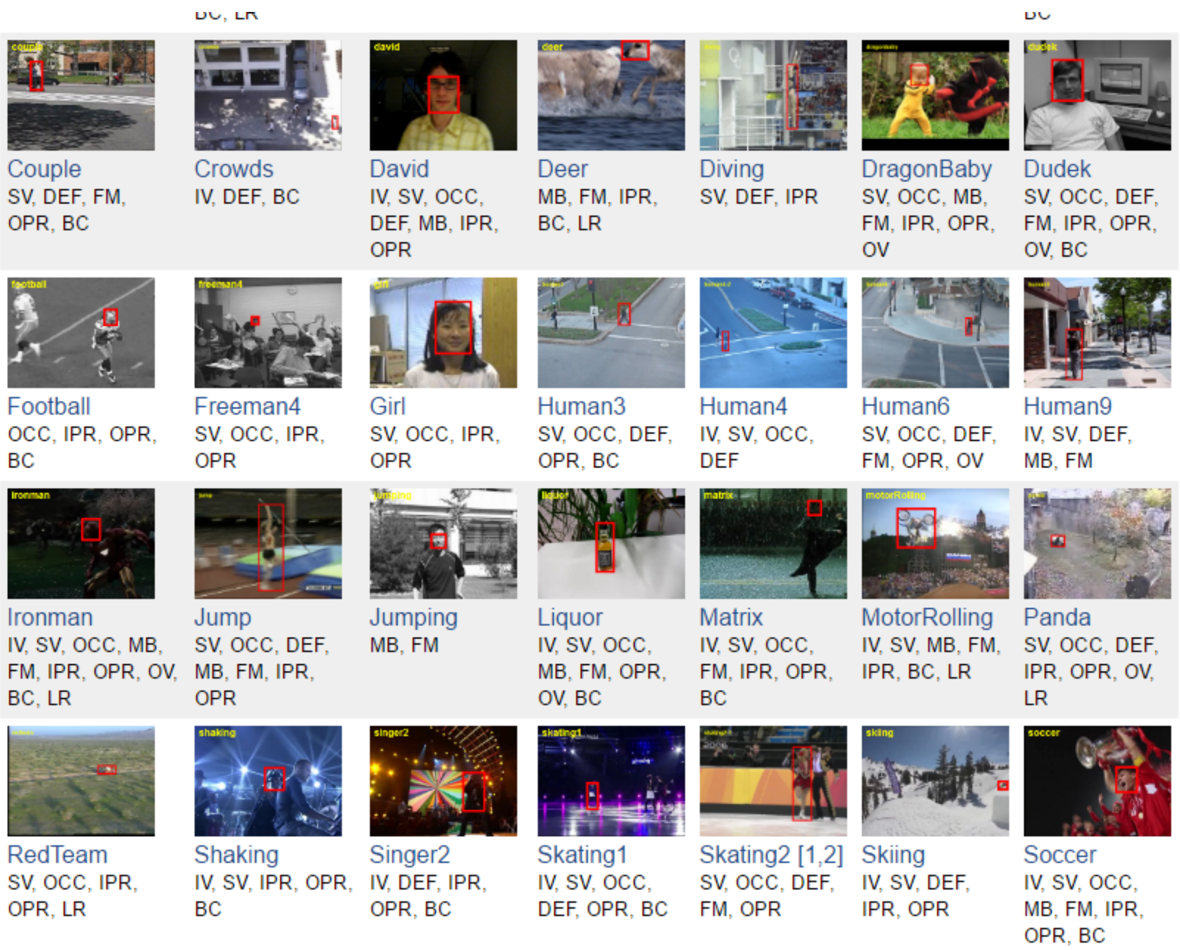


Figure 1: OTB dataset

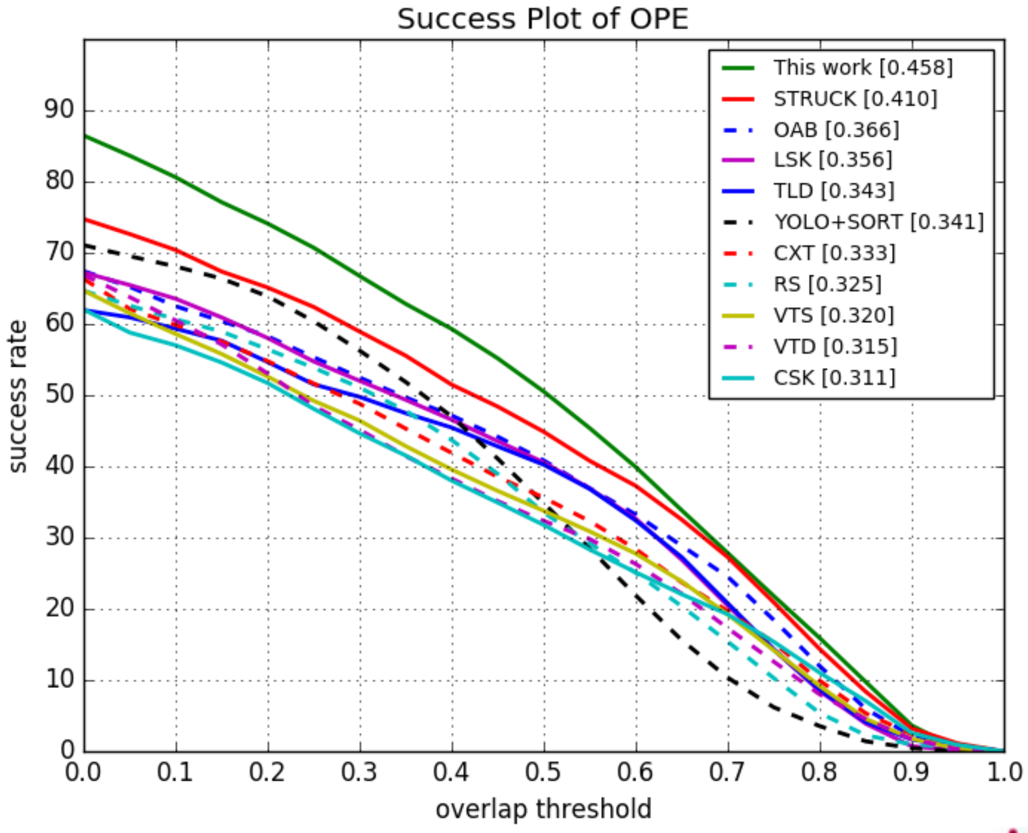
OTB is one of the most commonly used datasets. Each video is annotated with one or more attributes:

- IV: Illumination Variation
- SV: Scale Variation
- OCC: Occlusion
- DEF: Deformation
- MB: Motion Blur
- FM: Fast Motion
- IPR: In-plane Rotation
- OPR: Out-of-Plane Rotation
- OV: Out-of-View
- BC: Background Clutters
- LR: Low Resolution

Human Detection and Tracking

❑ Evaluation (OPE)

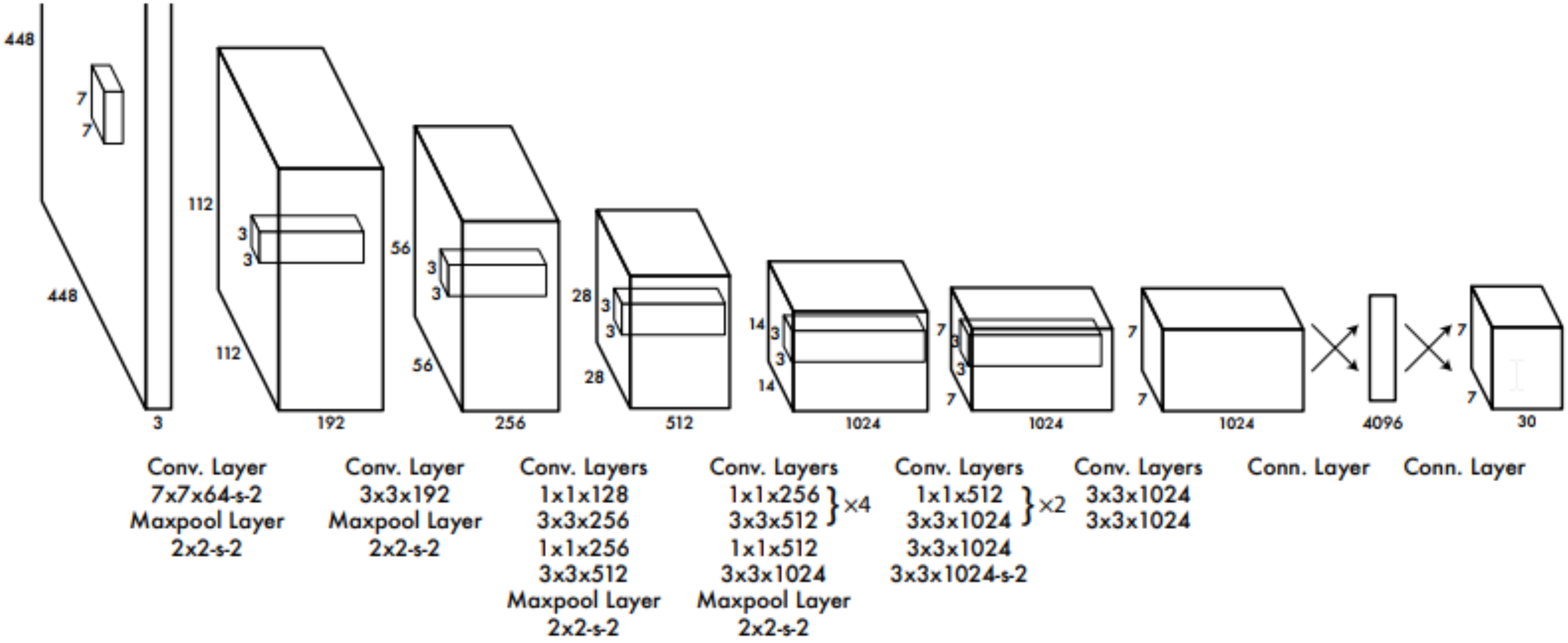
- Average Precision
- AUC of a Success Plot



Human Detection and Tracking

❑ Major Related Works

- **YOLO**

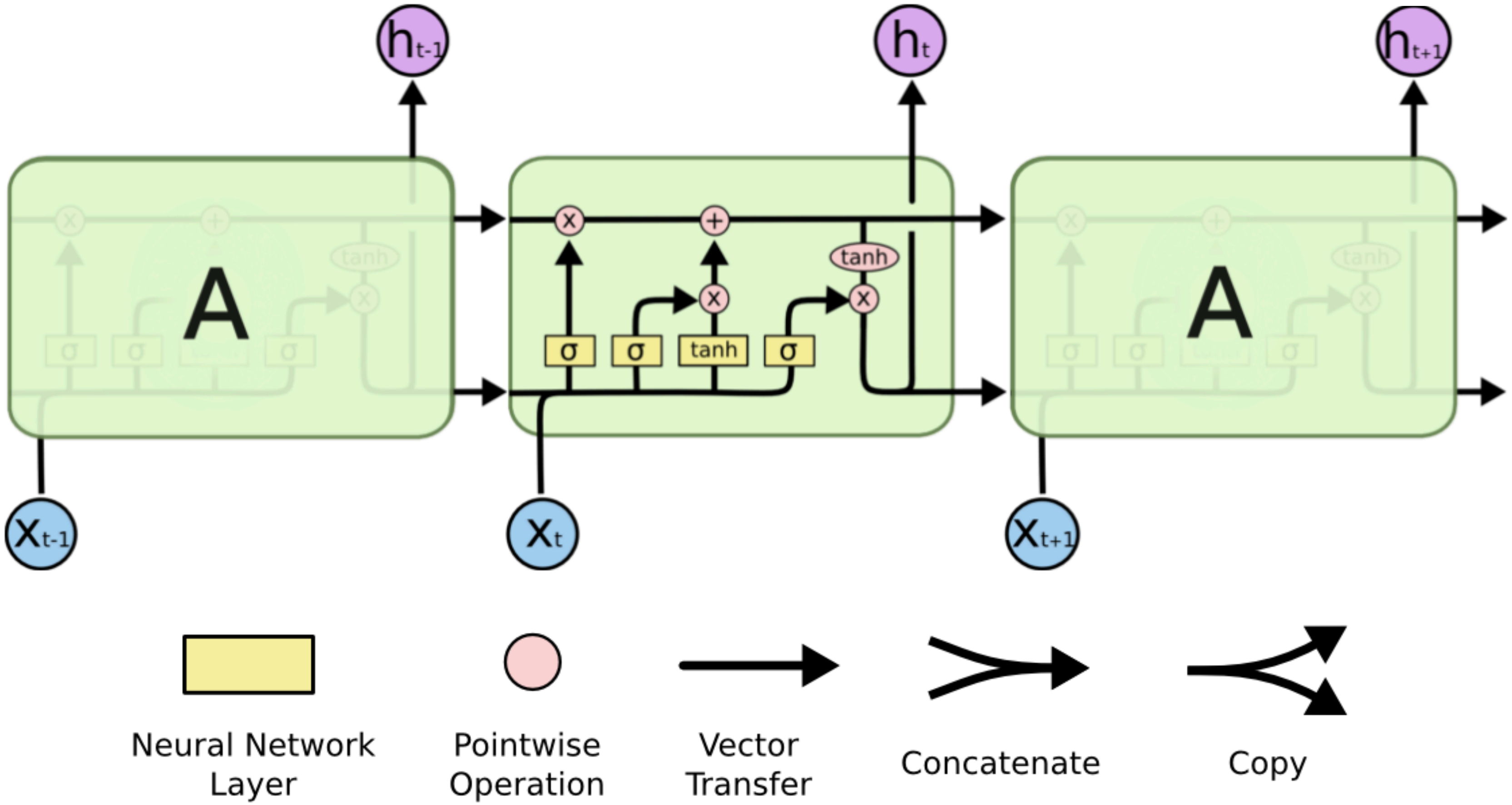


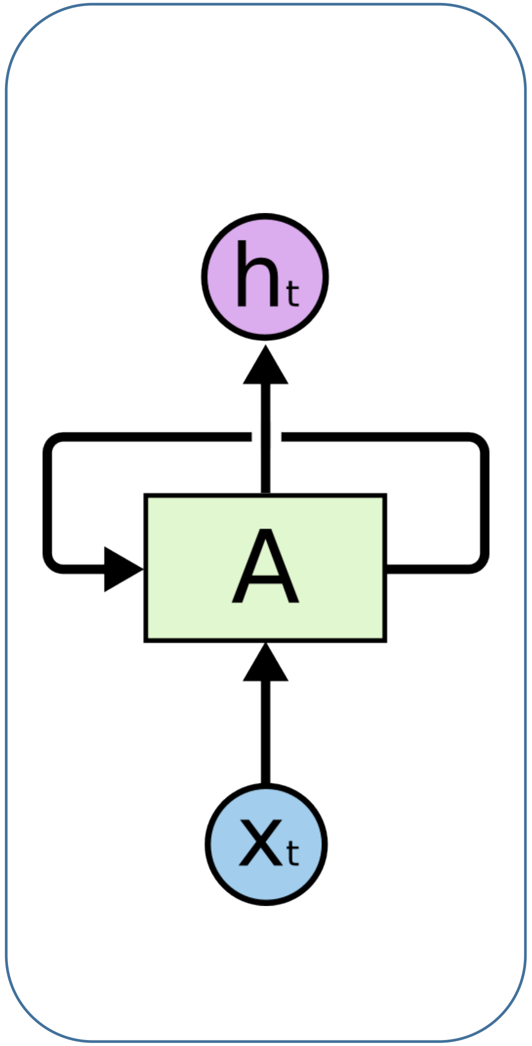
Human Detection and Tracking

- Major Related Works
  - **LSTM**

$$\sigma = (1 + e^{-x})^{-1}$$

$$\phi(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$



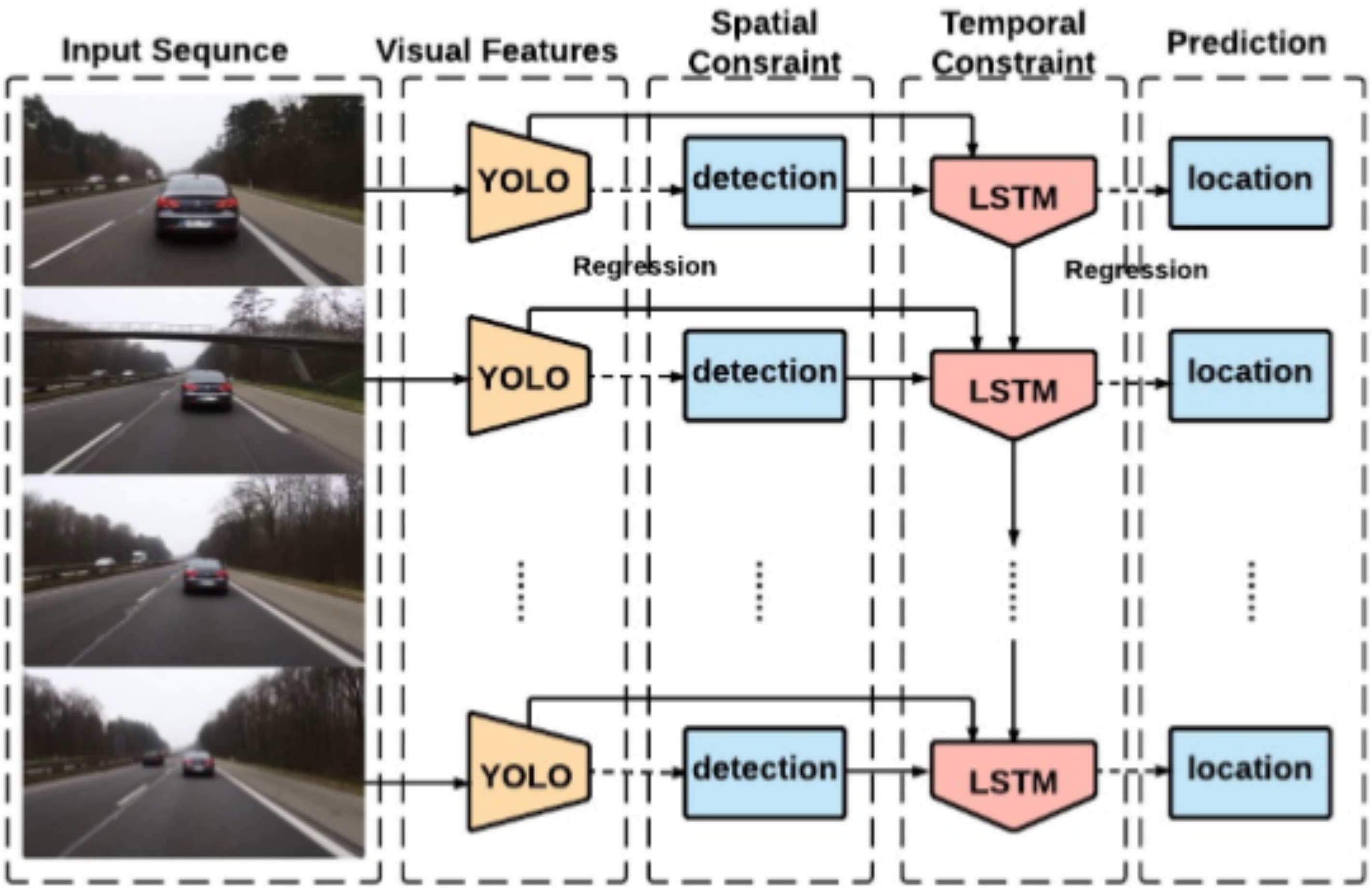


Human Detection and Tracking

- Motivation
  - Existing Tracking Methods do not handle full occlusion very well
  - Existing Tracking Methods based on CNN is slow. (1~2 fps)
  - No RNN-based tracking method on real data had been proposed

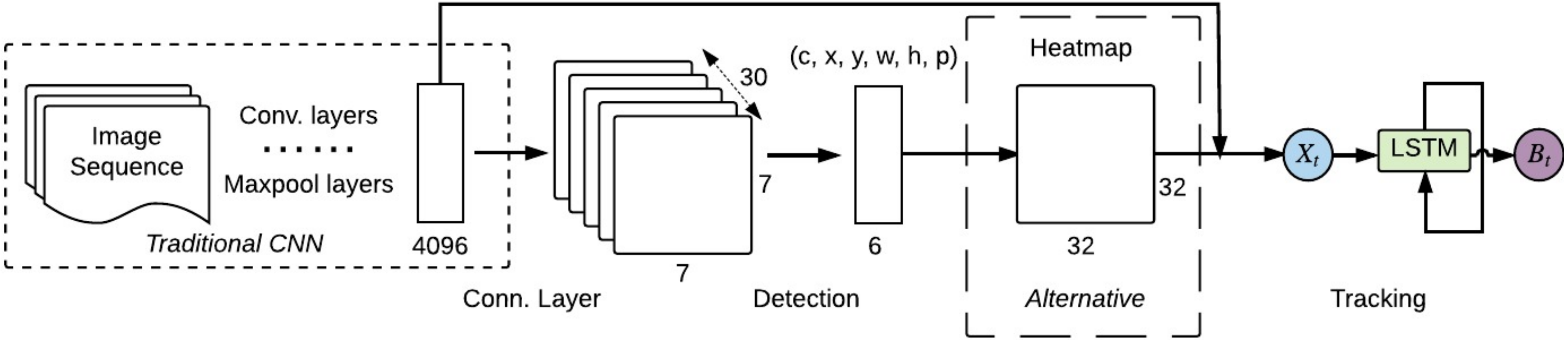
Human Detection and Tracking

❑ Simplified Overview



Human Detection and Tracking

❑ Architecture



Human Detection and Tracking

❑ Simple loss Functions

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n} ||B_{target} - B_{pred}||_2^2,$$

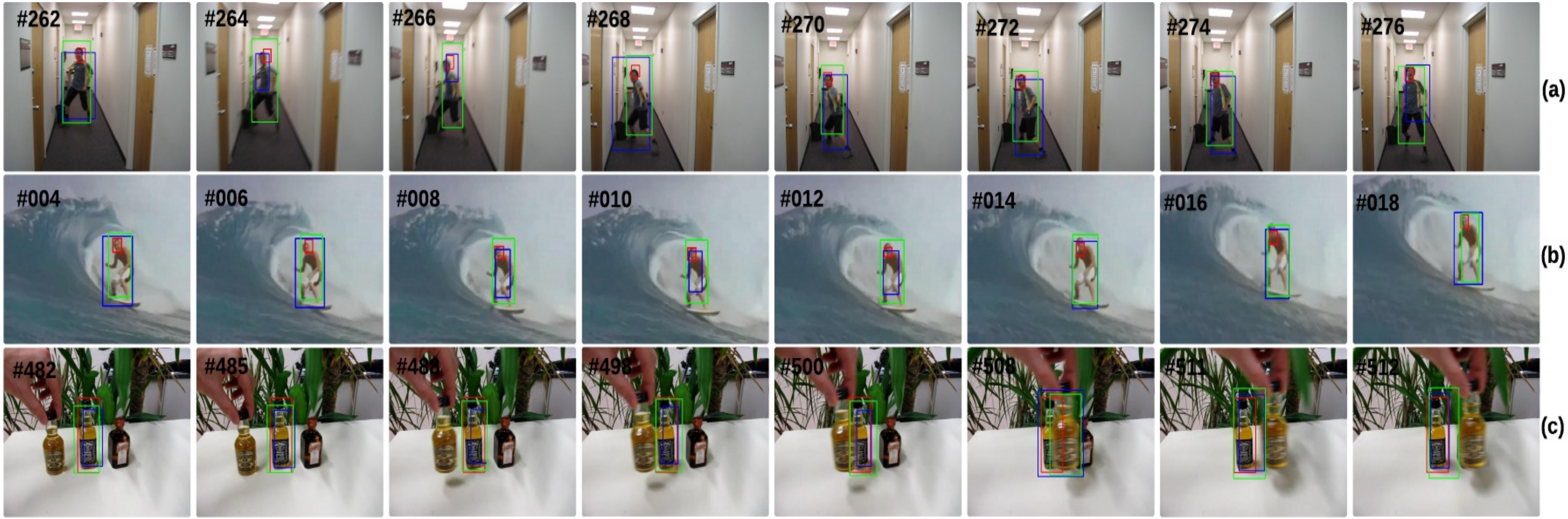
$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n} ||H_{target} - H_{pred}||_2^2,$$

Human Detection and Tracking

- ❑ **Qualitative Results for Sequences**

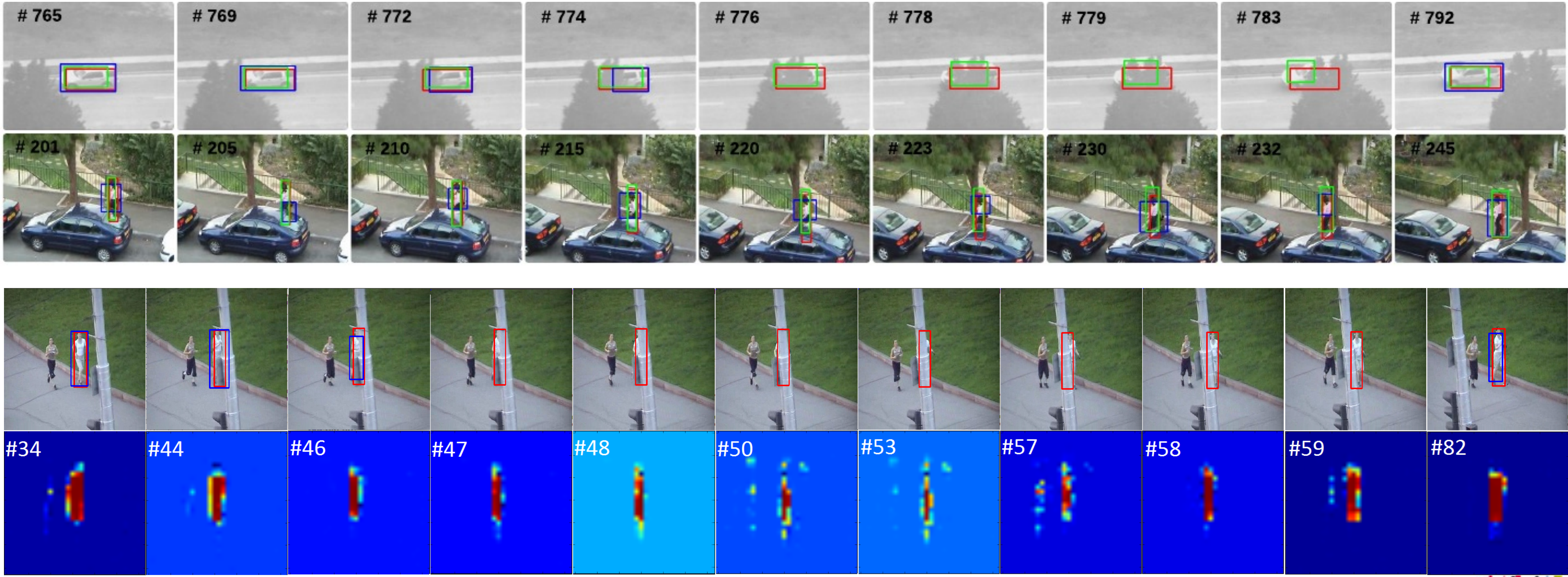
Human Detection and Tracking

❑ Qualitative Results over time



Human Detection and Tracking

❑ Spatio-temporal Robustness against Occlusion



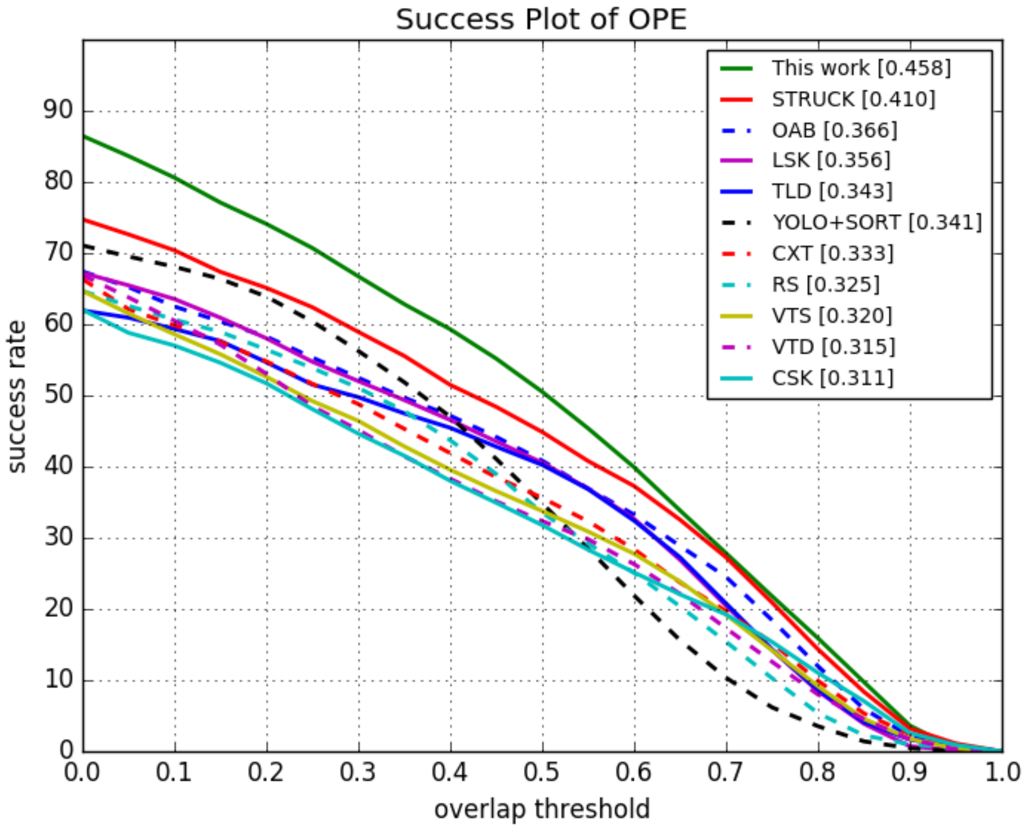
Human Detection and Tracking

❑ ROLO is effective due to several reasons:

- (1) the representation power of the high-level visual features from convNets,
- (2) the feature interpretation power of LSTM, therefore the ability to detect visual objects,
- (3) spatially supervised by a location or heatmap vector,
- (3) the capability of LSTM in regressing effectively with spatio-temporal information.

Human Detection and Tracking

❑ Performance



- Due to fast motions, occlusions, and therefore poor detections, YOLO with the kalman filter perform inferiorly lacking knowledge of the visual context.
- LSTM is capable of regressing both visual context and location histories, performing better than [YOLO + Kalman]

Area Under Curve (AUC) score reflected on right-top.

Human Detection and Tracking

- Summary of Contributions

  - Our proposed ROLO method extends the deep neural network learning and analysis into the spatiotemporal domain. It is the first work that proposes to incorporate CNN and LSTM for object tracking.

  - We have studied LSTM’s interpretation and regression capabilities of high-level visual features.

  - Our proposed tracker is both spatially and temporally deep, and can effectively tackle problems of major occlusion and severe motion blur.

Mizzou
University of Missouri - Columbia

Chapter. 3
Single-Person Human Pose Estimation

Single-Person Human Pose Estimation

❑ Contents

- 0. Problem Definition
- 1. Evaluation Criterion
- 2. Datasets
- 3. Benchmarks
- 4. Our Performance
- 5. The Proposed Method
    - 5.1 Overview
    - 5.2 Implementation Details
- 6. Qualitative Results
- 7. Discussion

Single-Person Human Pose Estimation

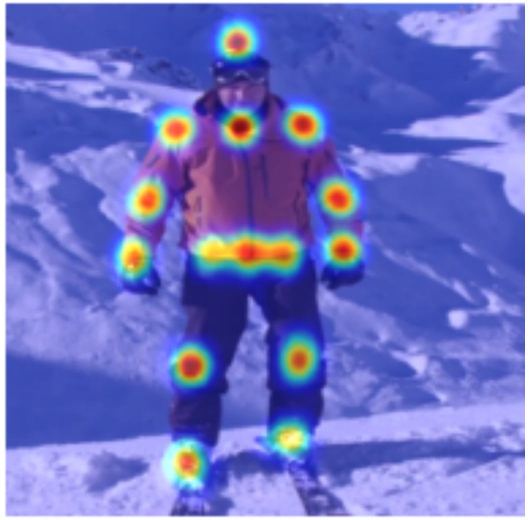
- Problem Definition

2D vs 3D

Image vs Video

Single-person vs Multi-Person

Original image →



→

Single-Person Human Pose Estimation

❑ Evaluation Criterion

(1) PCK Measure: (Percentage of Correct Keypoints)

(2) PCKh Measure

- a specific kind of PCK measure that uses the matching threshold as a certain percentage of the head segment length.

(3) AUC (Area Under Curve)

- Draw a curve with different α values, calculate the area under curve

Single-Person Human Pose Estimation

- (1) MPII

- 25,000 images
- 40,000 people with annotated joints
- 410 human activities

# of joints: 16

Single-Person Human Pose Estimation

❑ (2) LSP

❑ 1000 training images
And 10000 extended images

❑ 1000 testing images

# of joints: 14

Single-Person Human Pose Estimation

Related Works

Methods	School / Lab	Publication	Similarity	Differences
CPM	Carnegie Mellon	CVPR 2016	Fully Convolutional; Implemented in Caffe	Everything else
Hourglass	University of Michigan	ECCV 2016	Hourglass design	They use resnet as basic module; We propose a more robust module as the basic building block
Part Heatmap Regression	University of Nottingham	ECCV 2016	Improve on Limbs	They focus on part detection to aid heatmap regression; We focus on implicitly learning better limb prior.

Single-Person Human Pose Estimation

❑ Major Contributions

- We designed a novel building block for more robust feature representation
- We proposed a novel feature injection technique to guide the CNN model to learn limb prior
- We proposed a novel 3D cross-heatmap NMS technique for human pose estimation

Single-Person Human Pose Estimation

- (1) Performance on MPII dataset

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total	
Ours	98.1	96.3	92.2	87.8	90.6	87.6	82.7	**91.2**	
Newell et al., ECCV'16[55]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9	Hourglass (Michigan)
Bulat&Tzimiropoulos, ECCV'16 [84]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	Part Heatmap Regression
Wei et al., CVPR'16 [81]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	CPM (CMU)
Insafutdinov et al., ECCV'16 [76]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5	
Rafi et al., BMVC'16 [97]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3	
Gkioxary et al., ECCV'16 [98]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1	
Lifshitz et al., ECCV'16 [99]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0	
Pishchulin et al., CVPR'16 [75]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	
Hu&Ramanan, CVPR'16 [100]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4	
Tompson et al., CVPR'15 [74]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0	
Carreira et al., CVPR'16 [82]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3	
Tompson et al., NIPS'14 [54]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6	
Pishchulin et al., ICCV'13 [65]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1	

Table 3.2: Comparisons of PCKh@0.5 score on the MPII test set.

Single-Person Human Pose Estimation

❑ (2) Performance on LSP dataset

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total	
Ours	98.2	94.4	91.8	89.3	94.7	95.0	93.5	**93.9**	
Bulat&Tzimiropoulos. ECCV'16 [84],	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7	Part Heatmap Regression
Wei et al. CVPR'16 [81],	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5	CPM (CMU)
Insafutdinov et al. ECCV'16 [76],	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1	DeeperCut
Pishchulin et al. CVPR'16 [75],	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1	
Lifshitz et al. ECCV'16 [99],	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7	
Belagiannis&Zisserman FG'17 [80],	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2	
Yu et al. ECCV'16 [101],	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3	
Rafi et al. BMVC'16 [97],	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8	
Yang et al. CVPR'16 [102],	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6	
Chen&Yuille NIPS'14 [73],	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	
Fan et al. CVPR'15 [103],	92.4	75.2	65.3	64.0	76.7	68.3	70.4	73.0	
Tompson et al. NIPS'14 [54],	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	
Pishchulin et al. ICCV'13 [65],	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1	
Wang&Li et al. CVPR'13 [104],	84.7	57.1	43.7	36.7	56.7	52.4	50.8	54.6	

Table 3.3: Comparisons of PCK@0.2 score on the LSP test set.

Single-Person Human Pose Estimation

- (2) Performance on LSP dataset

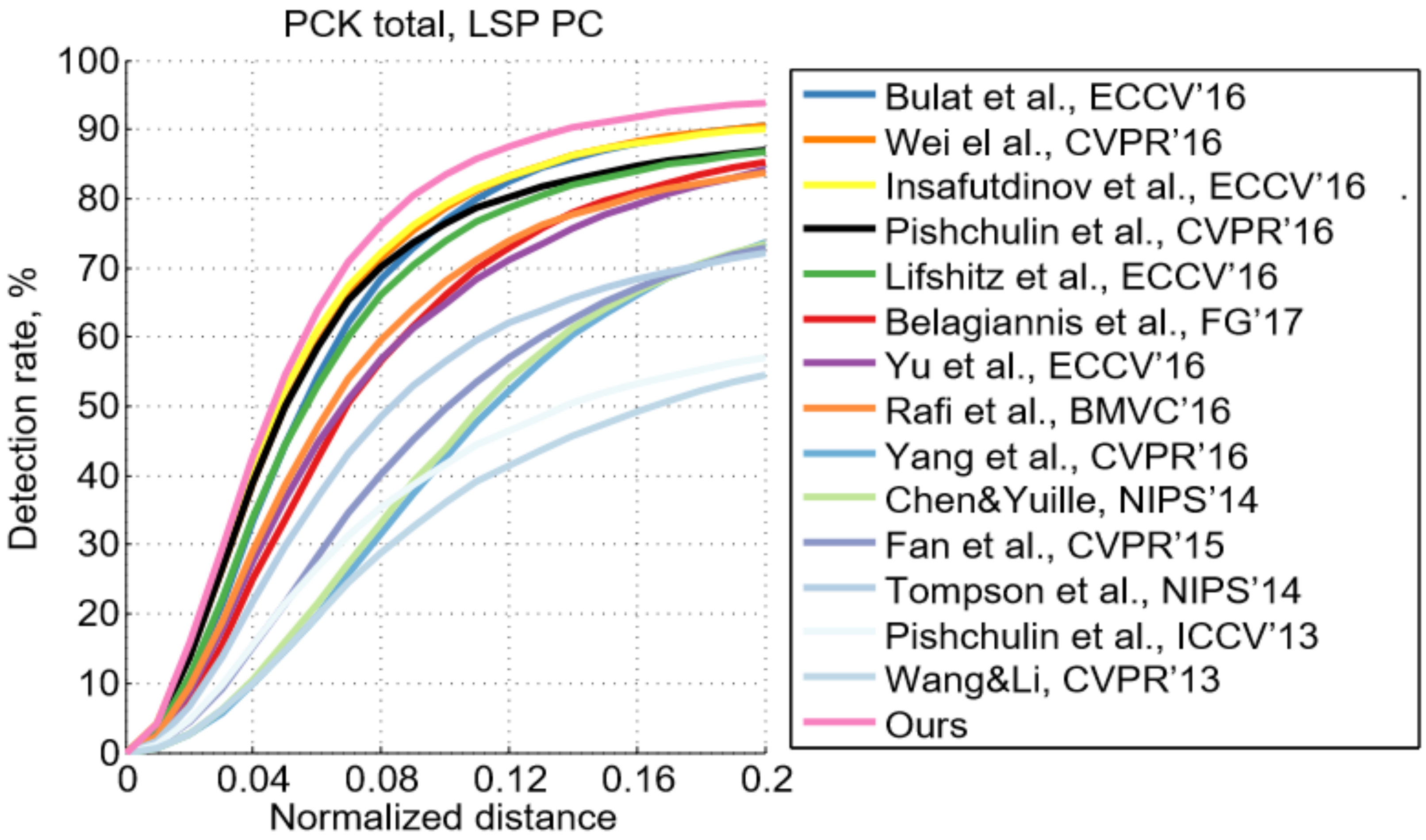
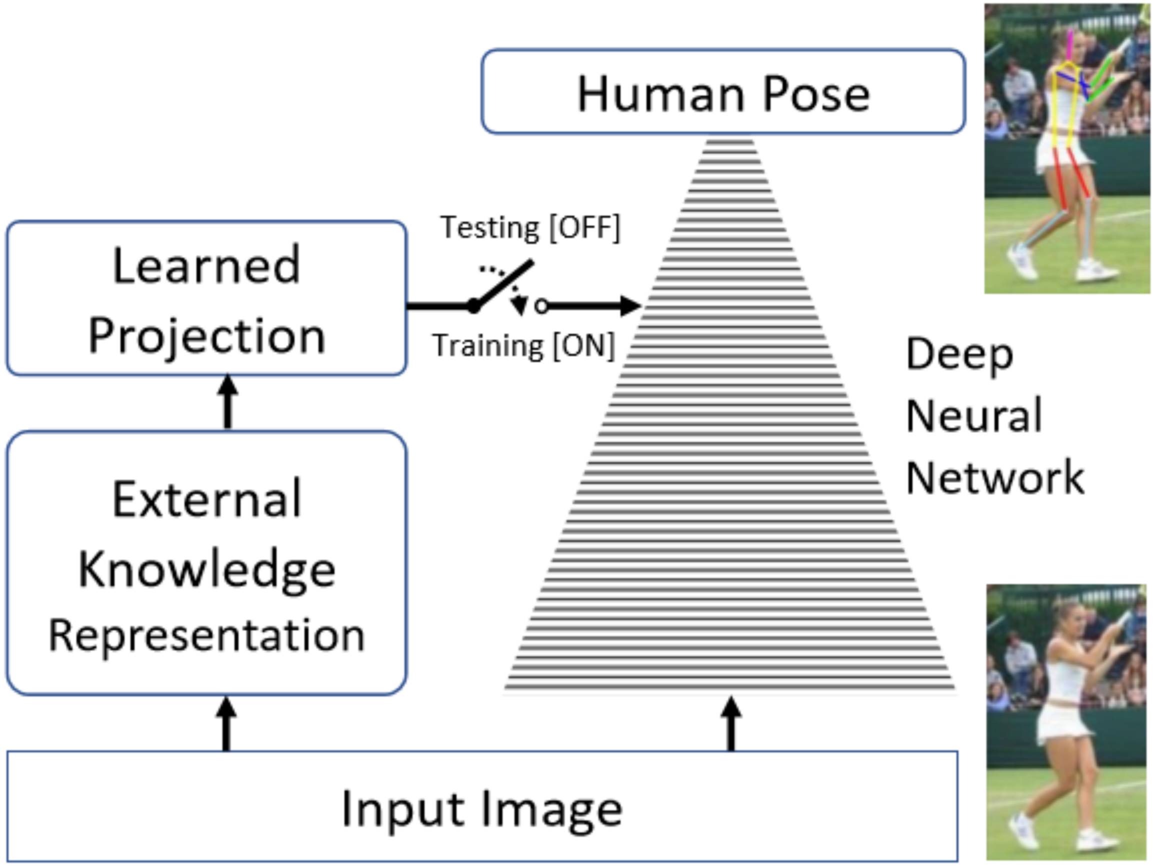


Figure 3.9: Person-Centric (PC) PCK curves on the LSP test set. Ours is on top.

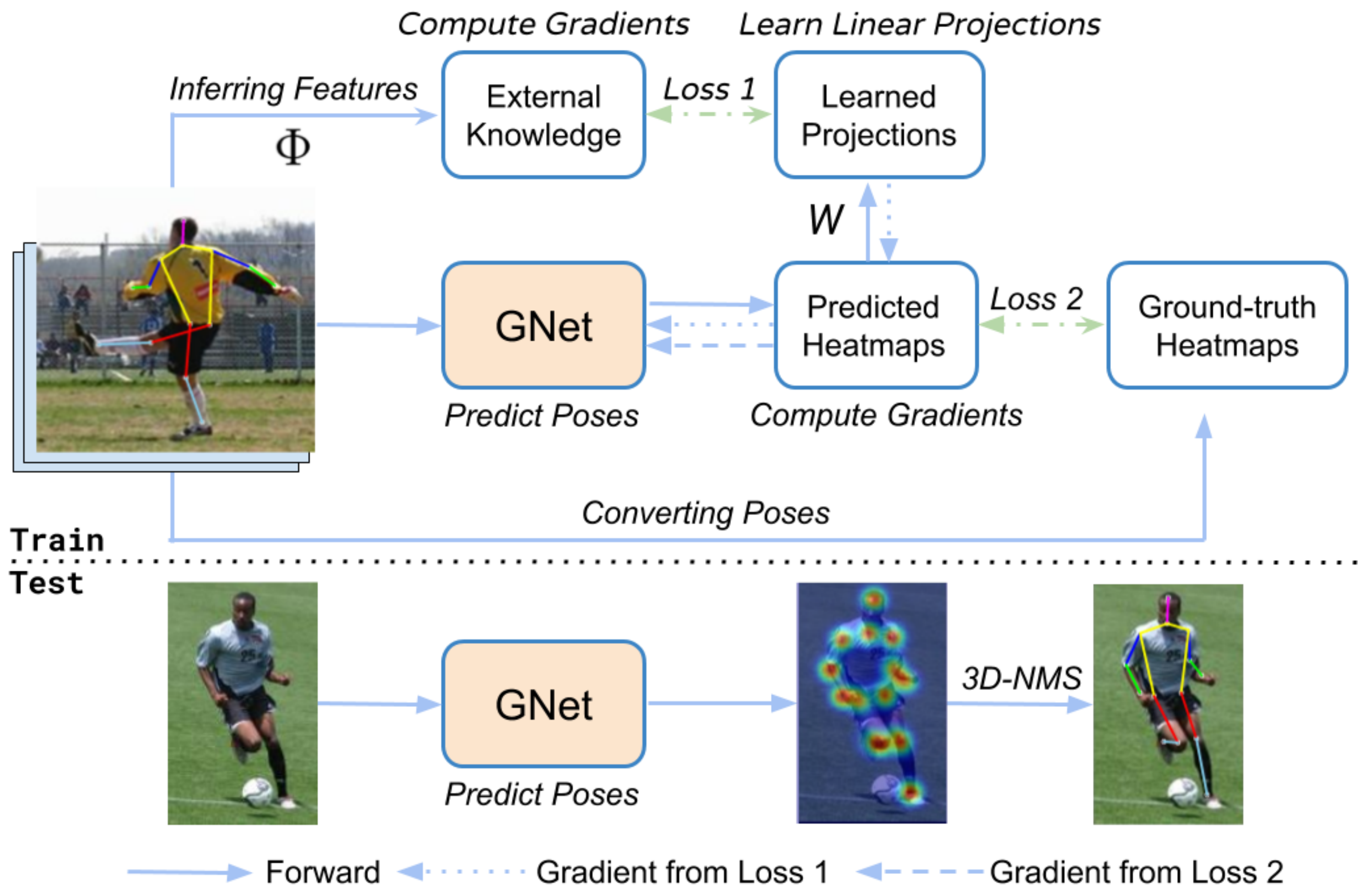
Single-Person Human Pose Estimation

❑ Motivation



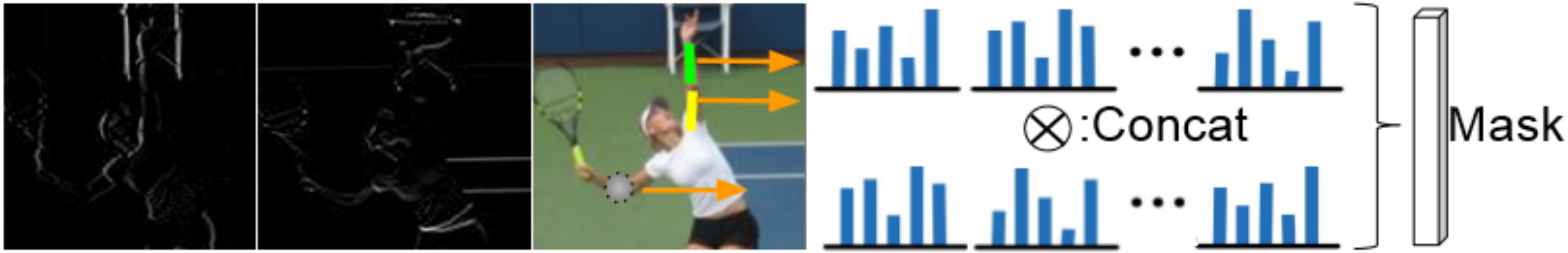
Single-Person Human Pose Estimation

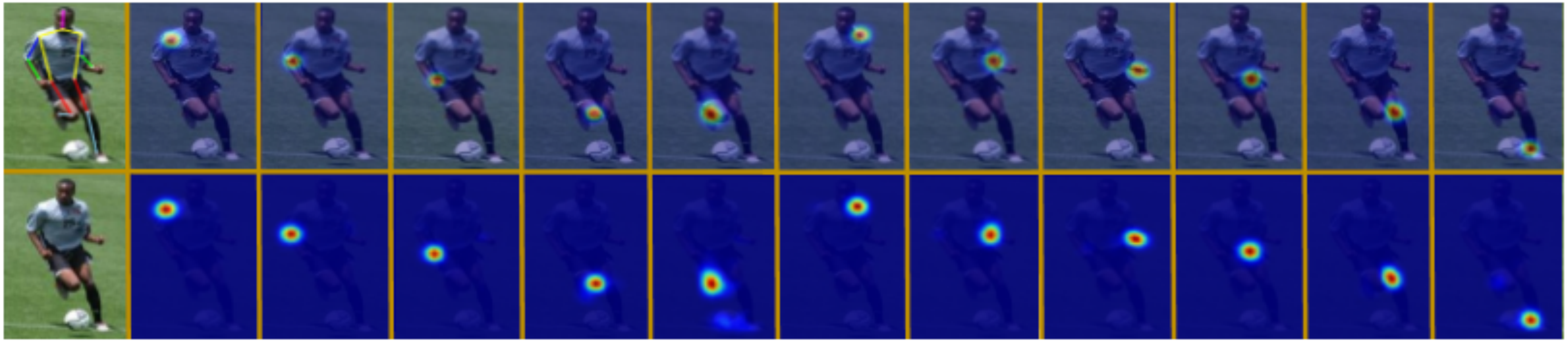
- Methodology



Single-Person Human Pose Estimation

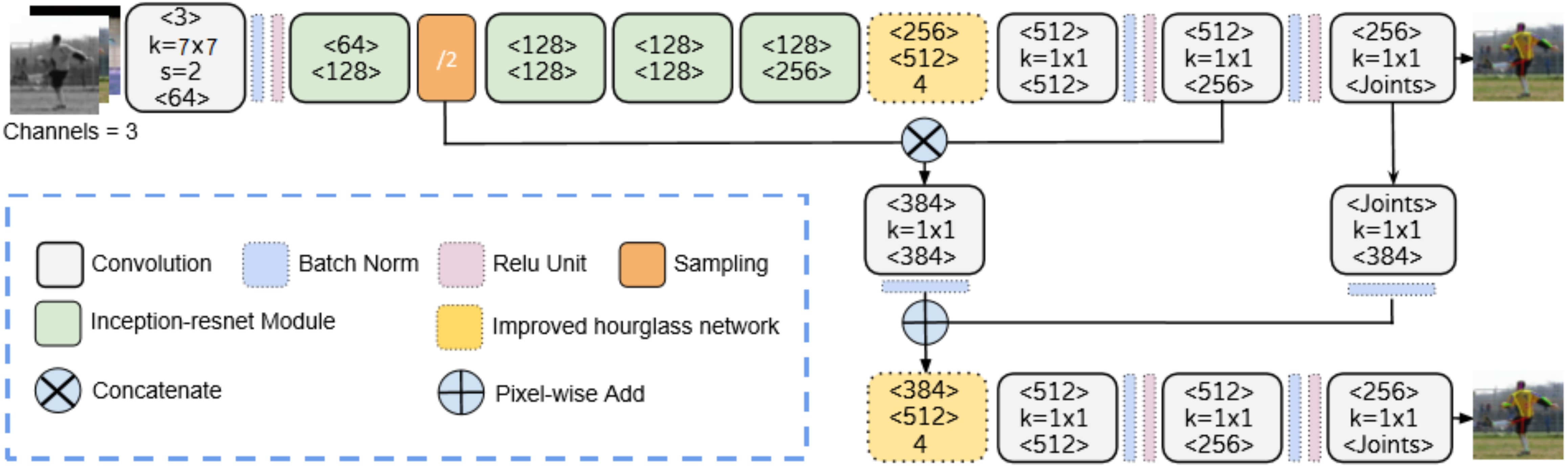
- Methodology





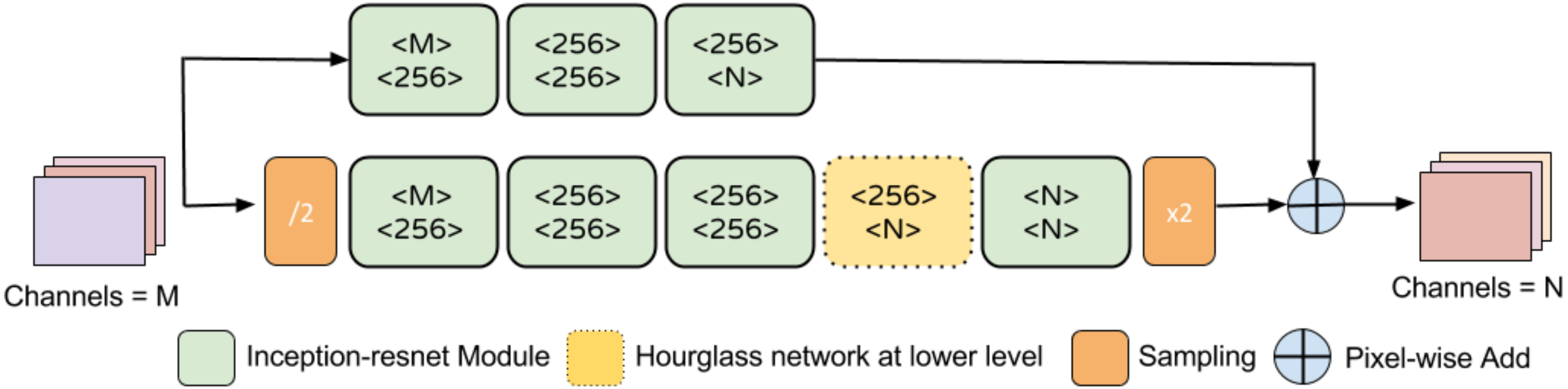
Single-Person Human Pose Estimation

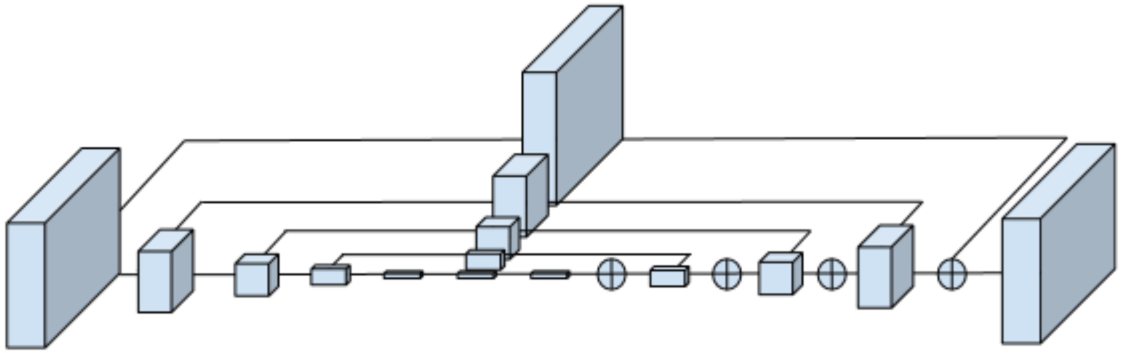
❑ Overview of the Network



Single-Person Human Pose Estimation

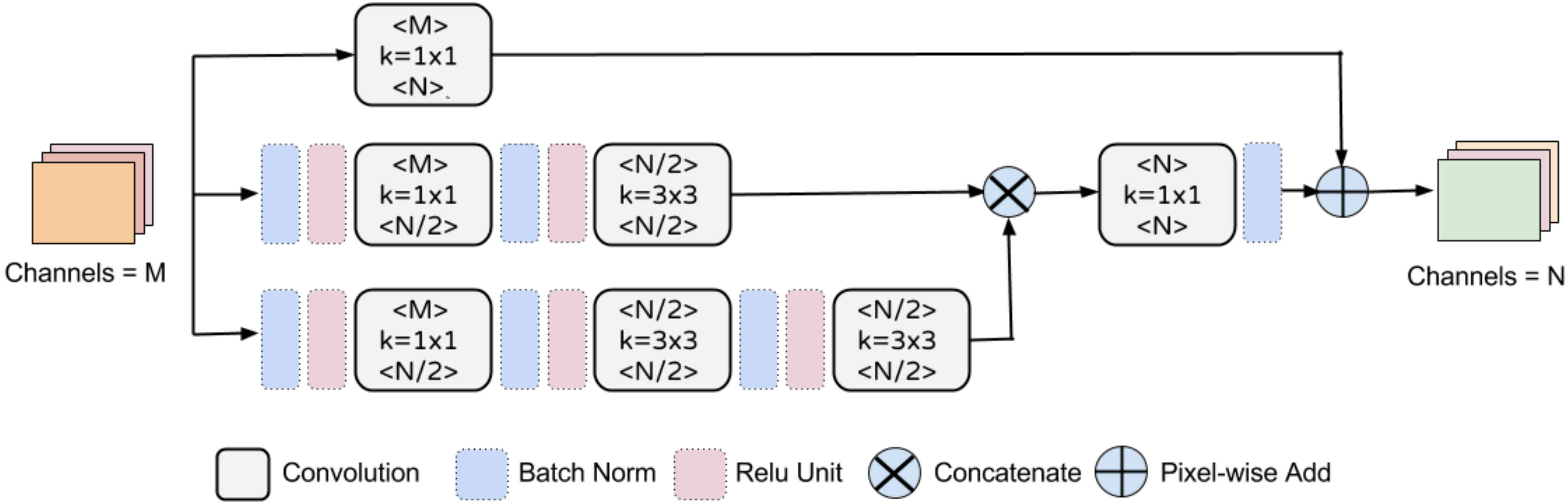
- Mid-level Abstraction: Hourglass Design





Single-Person Human Pose Estimation

- ❑ Basic Module: Inception-resnet module



Single-Person Human Pose Estimation

❑ Loss Functions

$$W_f^* \leftarrow \arg\min_{W_f}(\lambda \times \mathcal{L}_{KP} + (1-\lambda) \times \mathcal{L}_f)$$

where

$$\mathcal{L}_{KP} = ||K - W \times H_J||_2^2 + \beta \times ||W||_2^2$$

$$\mathcal{L}_f = \sum_{j \in J} ||H_j(p) - H_j^*(p)||_2^2$$

Single-Person Human Pose Estimation

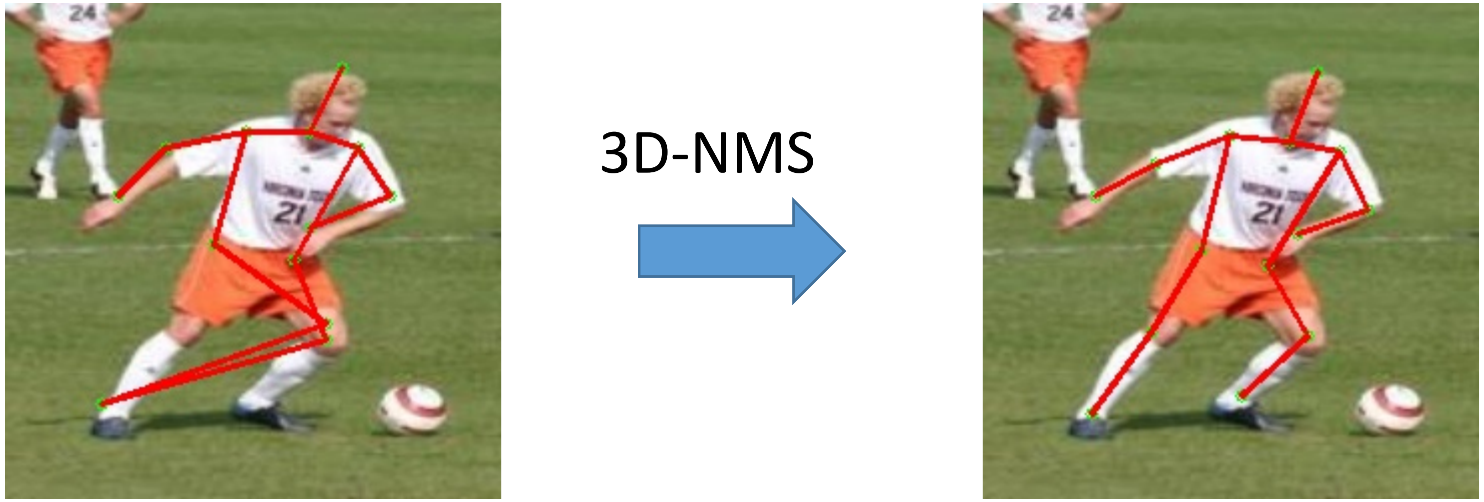
❑ Cross-Heatmap Non-maximum Suppression

- For each heatmap channel, detect blobs
- Rank all of them to a list by confidence. From the top blob:
    - Make this blob the final detection of its heatmap
    - Suppress other blobs from the same heatmap
    - Suppress blobs from other heatmap channels that are close to this blob in image coordinate system
- Until no blobs can be further removed
- If a channel has no blob left, find the maximum pixel in this heatmap

Single-Person Human Pose Estimation

❑ Cross-Heatmap Non-maximum Suppression

- Results are from our earliest poor pose estimator



Single-Person Human Pose Estimation

❑ Implementation Details

- Preprocessing
  - Input image is normalized by mean subtraction at each channel
  - Data augmentation by: rotation, flipping, cropping

- Training Details
  - RMSProp Optimization
  - Learning rate = 10^4, then step-decrease to 10^-6
  - Momentum (not applicable)
  - Batchsize = 12
  - Epoches >= 300
  - Heatmap: Gaussian with variance of 1.3
  - Weight gradient responses on background and joints, otherwise the network converge to zero.

- Training time:
  - 3 days to reach 93.9% (4 TITAN X)

Single-Person Human Pose Estimation

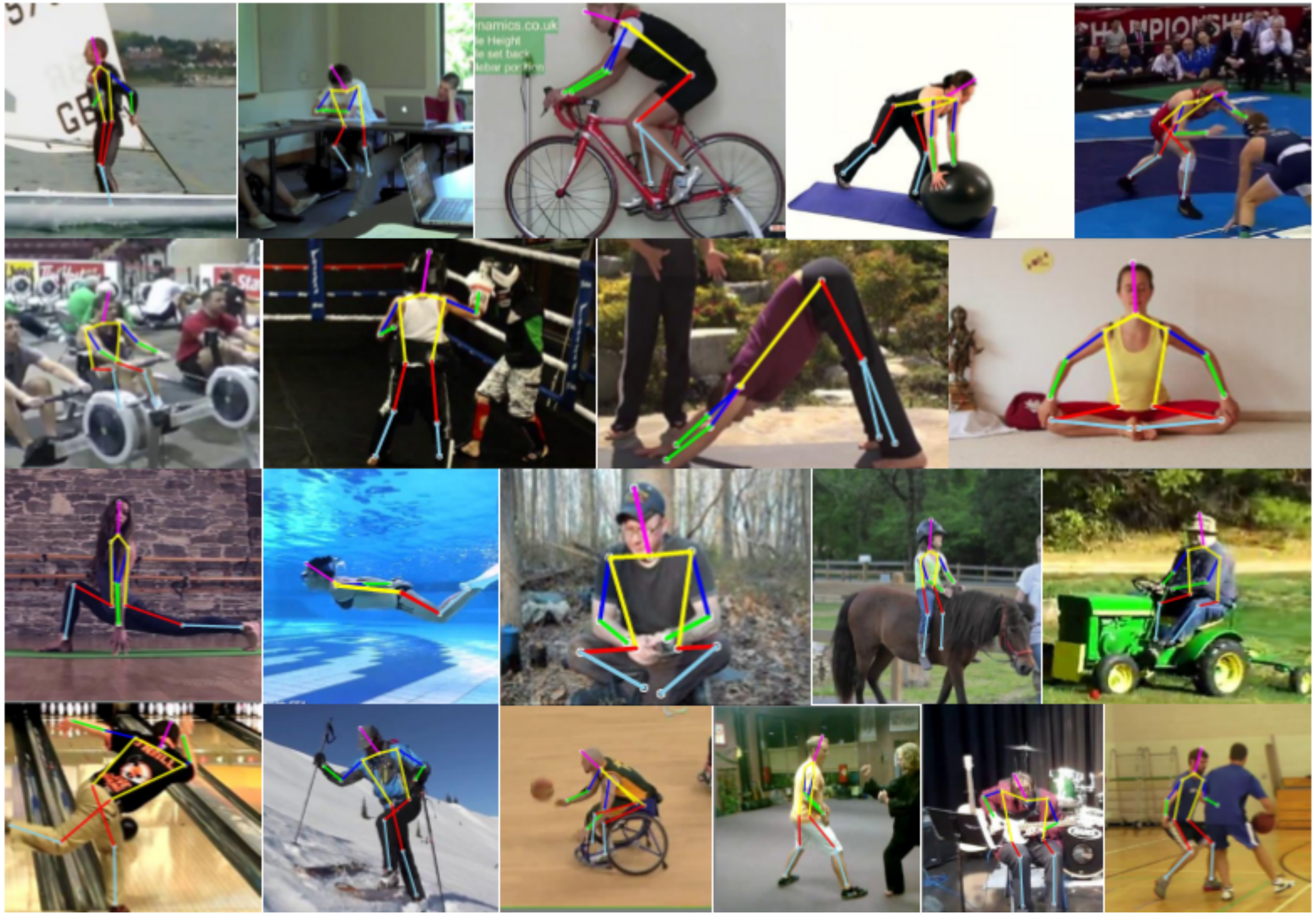
❑ Ablation Study

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total
Hourglass	97.0	93.0	88.8	85.6	92.2	93.0	90.9	**91.5**
Ours (no guidance)	97.9	93.2	89.1	86.4	94.5	93.8	92.9	**92.6**
Ours (with guidance)	98.2	94.4	91.8	89.3	94.7	95.0	93.5	**93.9**
Plain testing	97.4	92.7	88.8	86.7	92.2	93.8	92.2	92.0
+ flipping	97.7	93.3	90.4	87.5	93.2	94.2	92.8	92.7
+ scaling	98.1	93.7	91.3	88.7	94.0	94.6	93.2	93.4
+ 3D-NMS	98.2	94.4	91.8	89.3	94.7	95.0	93.5	**93.9**

Table 3.1: **Component analysis** on the LSP Dataset of PCK@0.2 score. Note that numbers in bold indicate the method has employed all techniques during testing.

Single-Person Human Pose Estimation

- Qualitative Results of MPII



Single-Person Human Pose Estimation

- Qualitative Results of LSP

Mizzou
University of Missouri - Columbia

Chapter. 4

Human Pose Estimation with Adversarial Training

HPE with Adversarial Training

❑ **Remaining Problems**

HPE with Adversarial Training

❑ Motivation

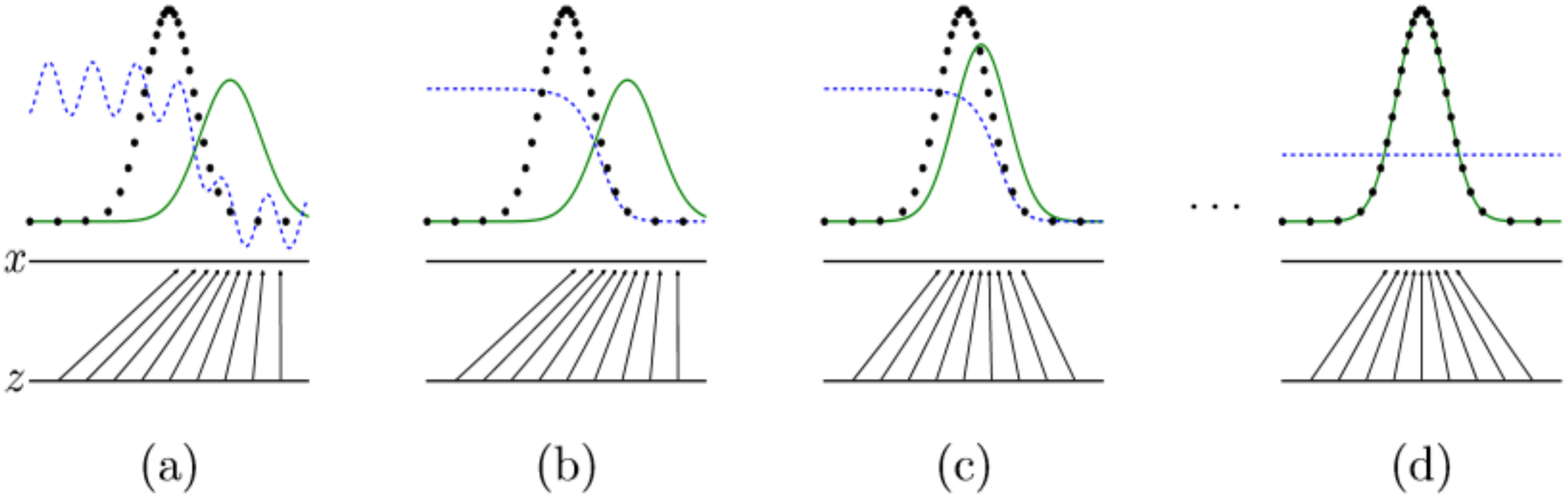
- For the first kind of failure, full occlusion of 2 or more adjacent body parts are hard to recover, as visual information from the RGB image is inadequate to resolve the ambiguity.

- For the second kind of failure, the mistakes are partly due to the body part noises from other persons and partly due to the occlusion of a single body part. These weak ambiguities can surely be mitigated with proper pose prior.

HPE with Adversarial Training

- Introduction to GAN

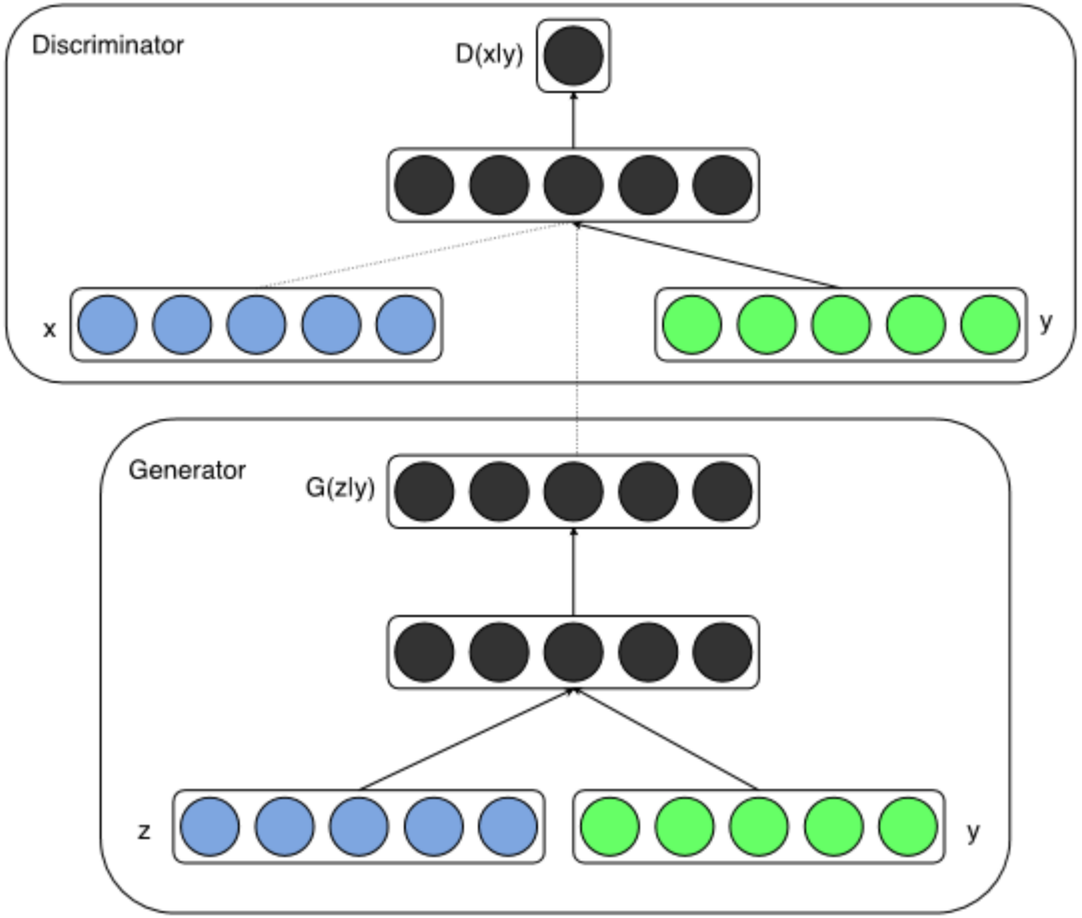
$$\min_G \max_D V(G,D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$



HPE with Adversarial Training

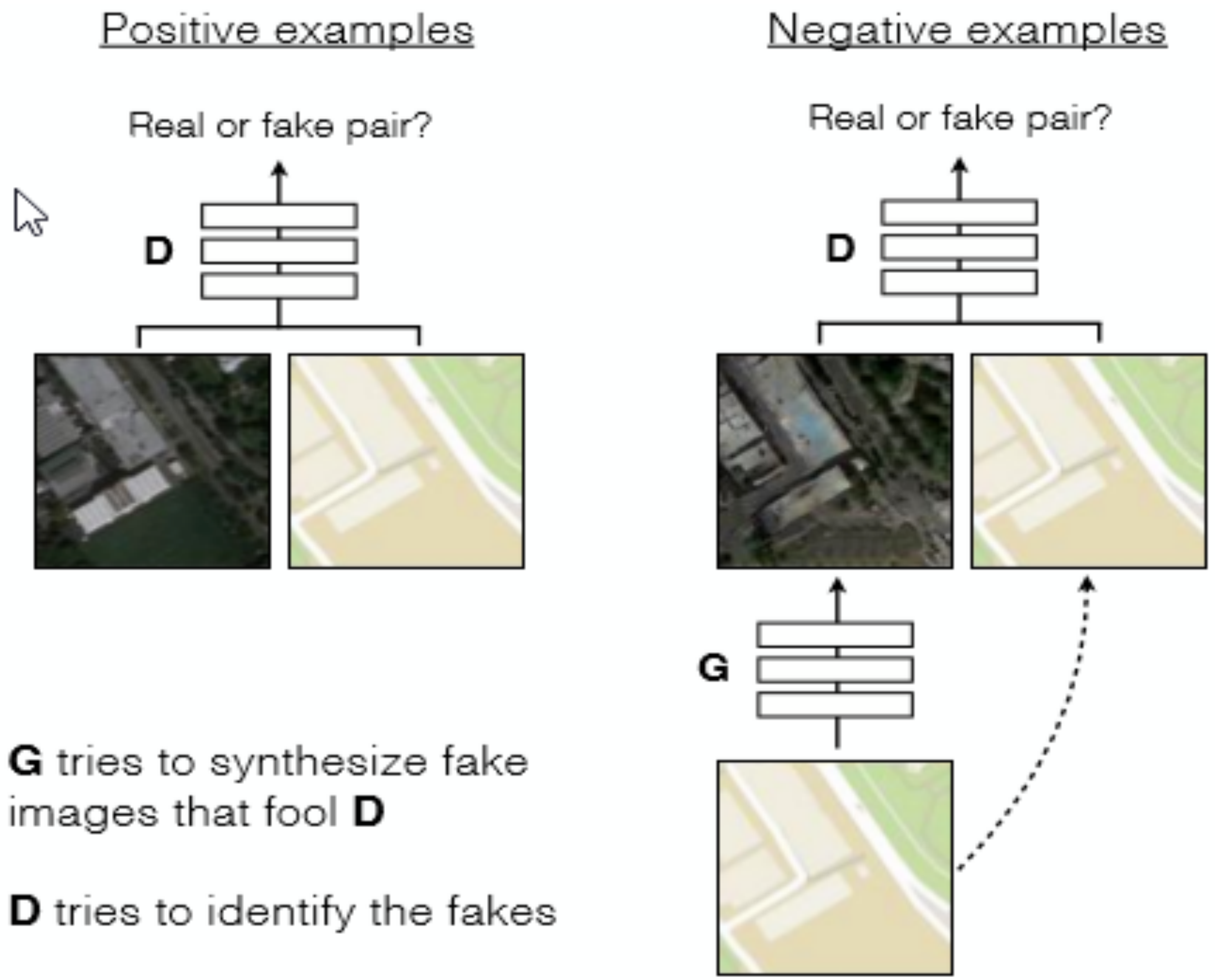
- ❑ Introduction to cGAN

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))].$$



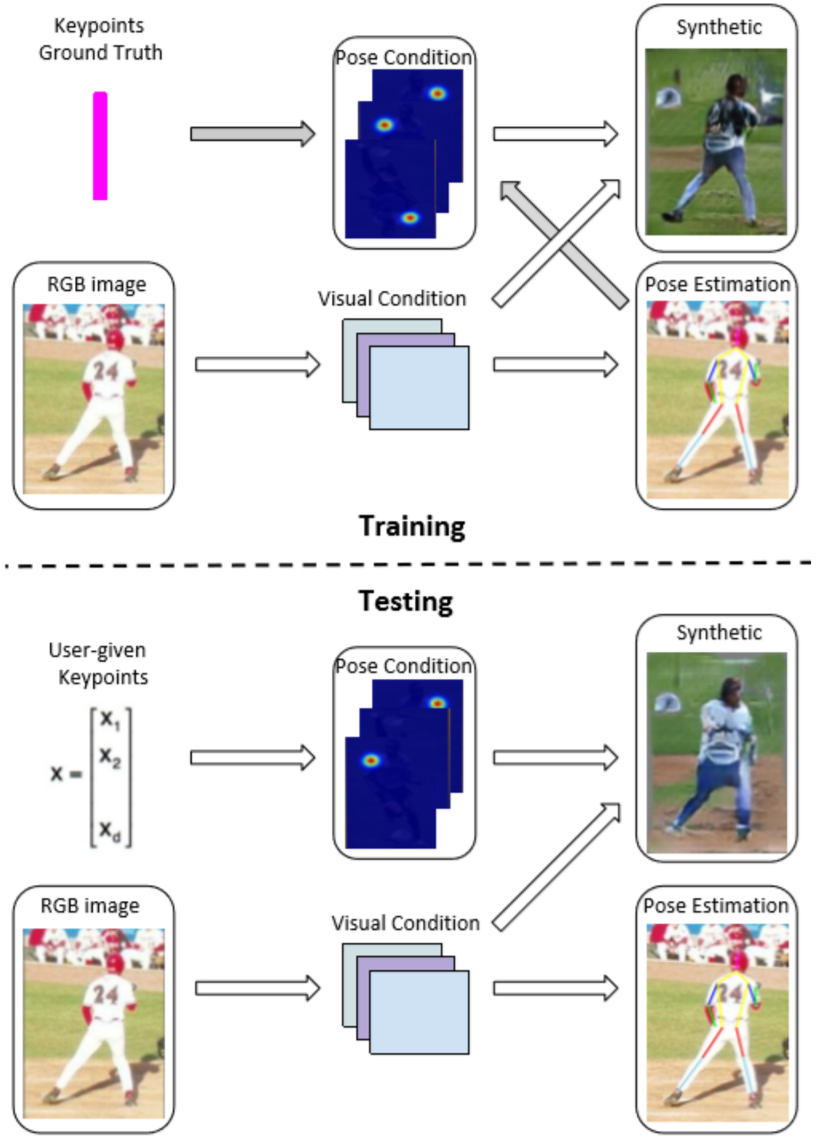
HPE with Adversarial Training

- Introduction to cGAN
  - Image-to-image: Pixel Level Translation



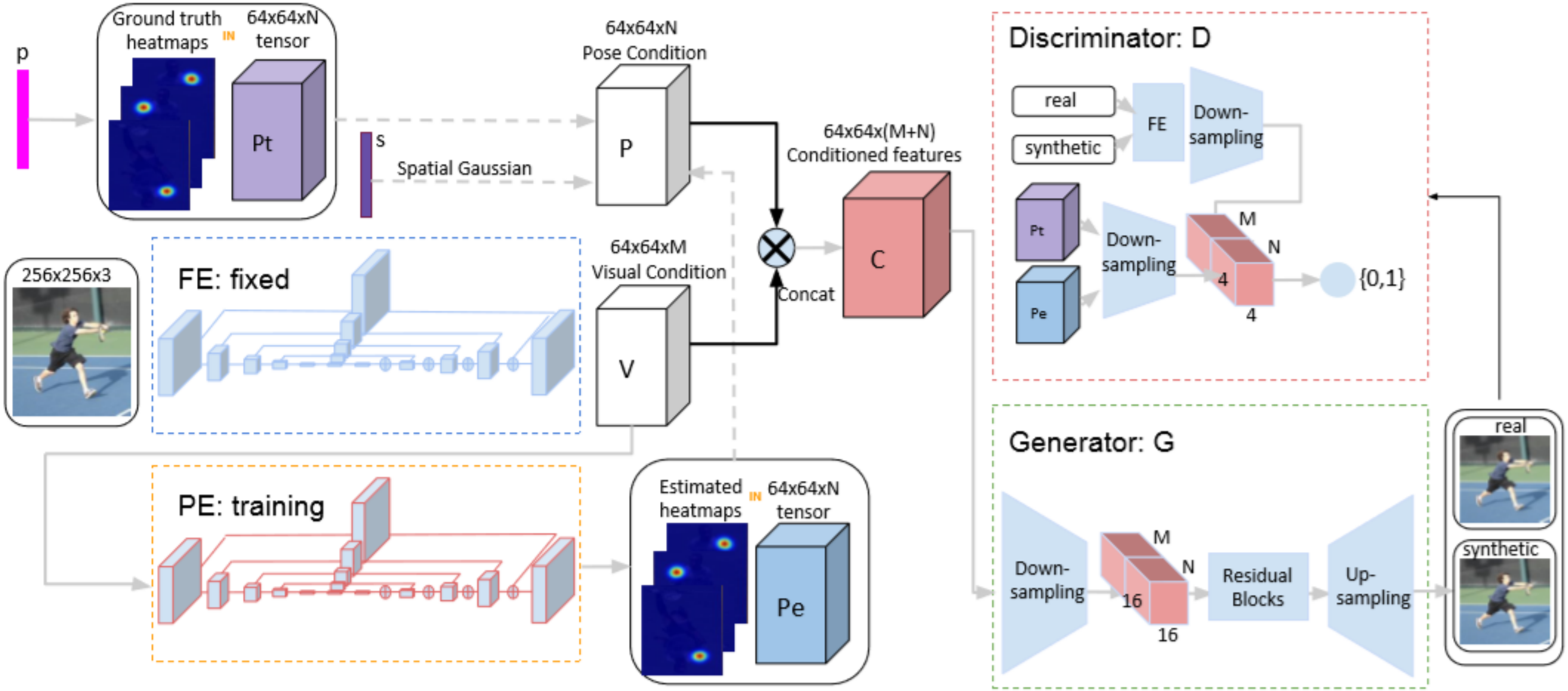
HPE with Adversarial Training

(1) Proposed Method



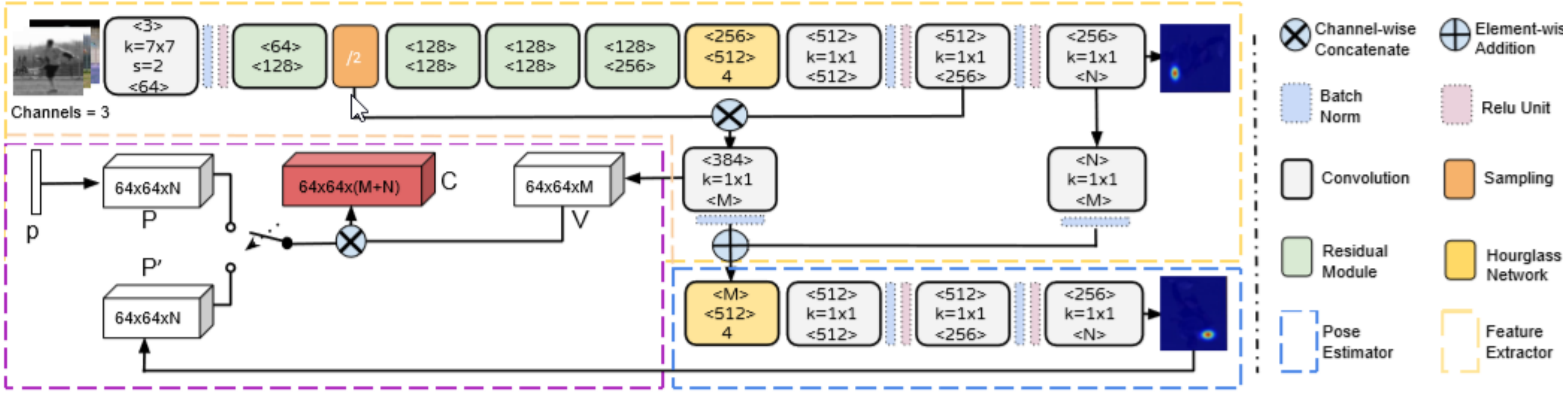
HPE with Adversarial Training

(2) Modules with Input and Output Details:



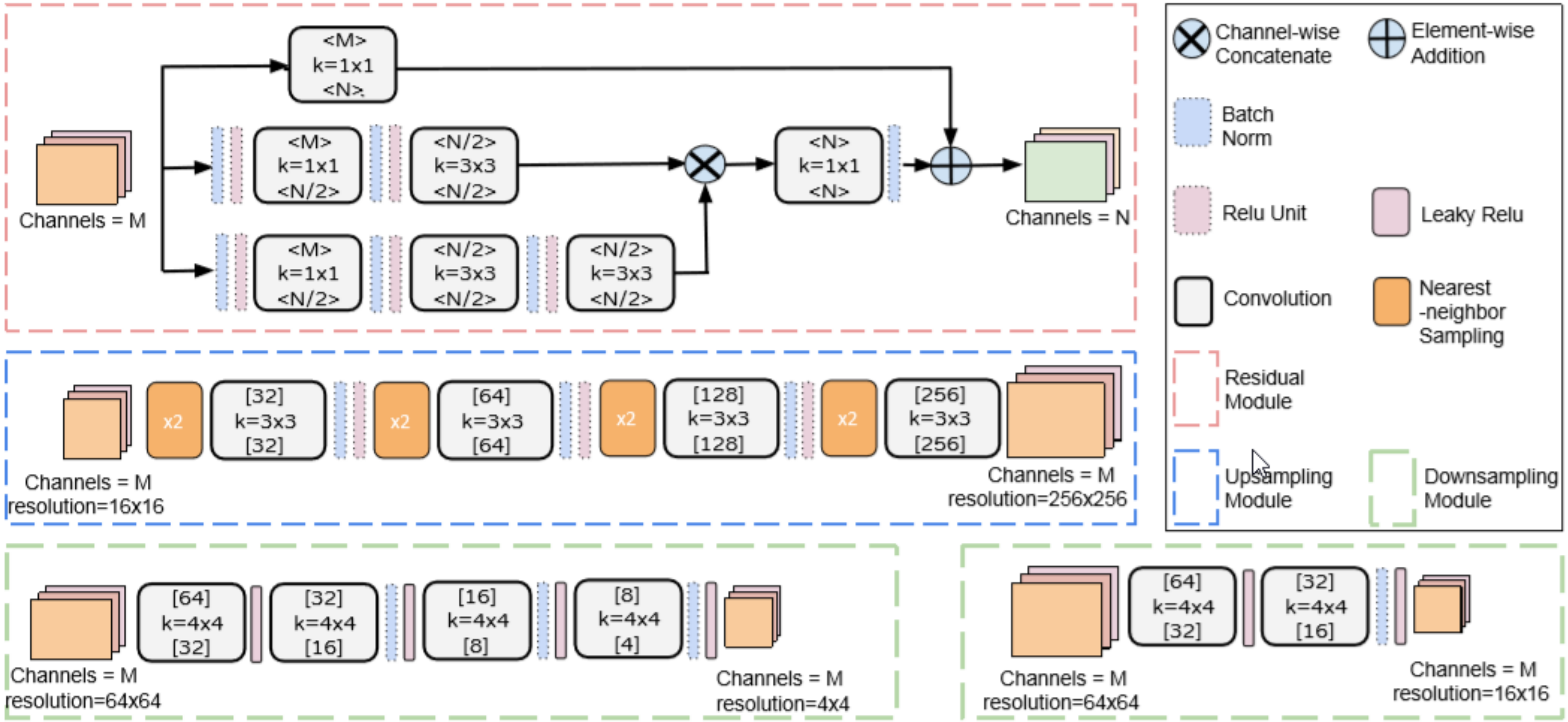
HPE with Adversarial Training

(3) Modules with Implementation Details:



HPE with Adversarial Training

(3) Modules with Implementation Details (Continued):

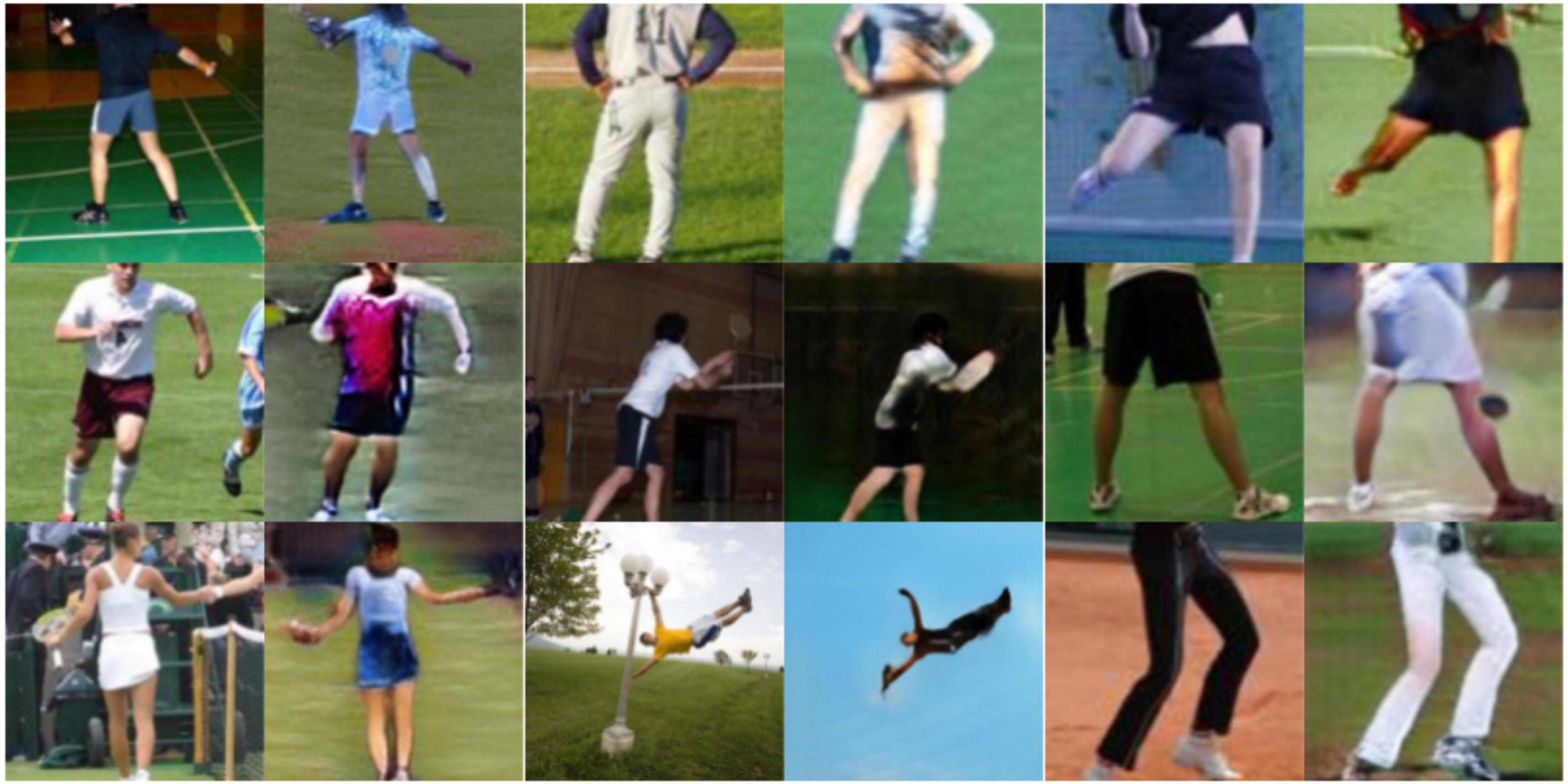


HPE with Adversarial Training

Mitigated Results

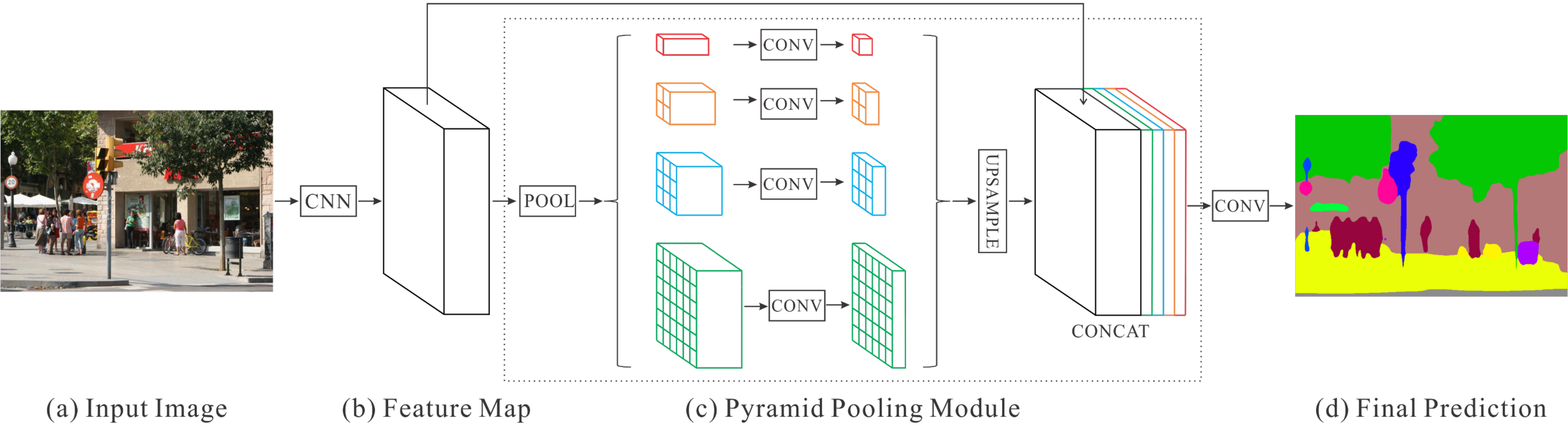
HPE with Adversarial Training

Pose-Conditioned Image Synthesis Results



HPE with Adversarial Training

Semantic Segmentation and Human Parsing



HPE with Adversarial Training

Parsing-Conditioned Image Synthesis Results

HPE with Adversarial Training

Parsing results comparison: Original image VS synthetic image

HPE with Adversarial Training

Parsing results on ATR test set:

Table 4.2: Per-class-accuracy on the ATR test set.

Method	Hat	Hair	Sunglass	Upper	Skirt	Pants	Dress	Belt	L-shoe	R-shoe	Face	L-leg	R-leg	L-arm	R-arm	Bag	Scarf	Avg
Ours	85.67	85.33	81.47	87.91	85.03	84.44	72.99	61.32	70.93	74.50	87.83	81.38	82.32	82.65	87.52	83.32	56.24	**79.46**
PSPNet,CVPR'17 [152]	81.71	83.64	77.70	87.44	77.33	84.22	76.66	61.60	67.95	60.40	88.89	76.69	83.52	78.84	80.33	85.61	50.75	76.66
CO-CNN, ICCV'15 [135]	72.07	86.33	72.81	85.72	70.82	83.05	69.95	37.66	76.48	76.8	89.02	85.49	85.23	84.16	84.04	81.51	44.94	75.65

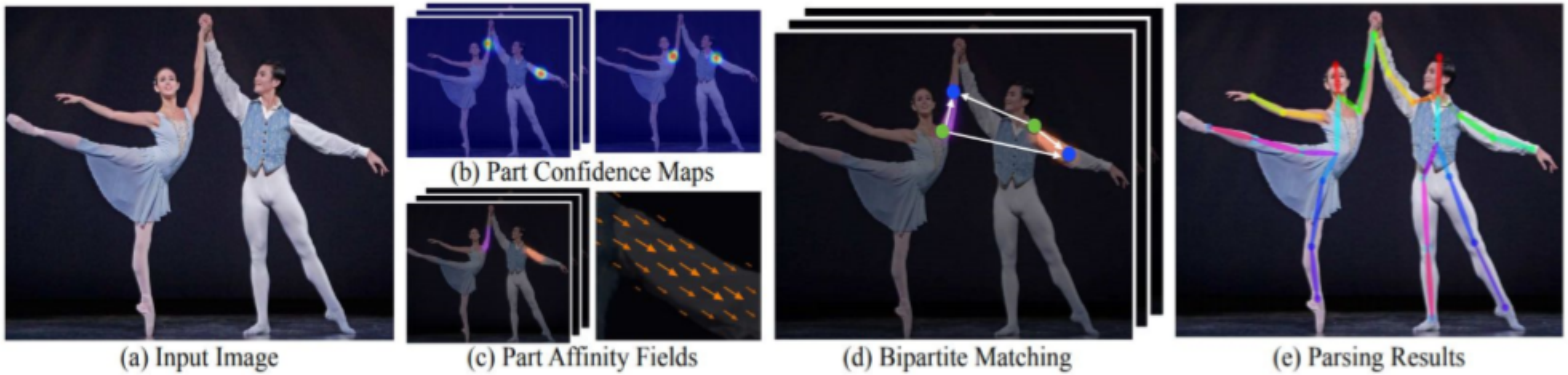
Mizzou
University of Missouri - Columbia

Chapter. 5

Multi-person Human Pose Estimation: PoseTrack Challenge

Multi-Person Human Pose Estimation

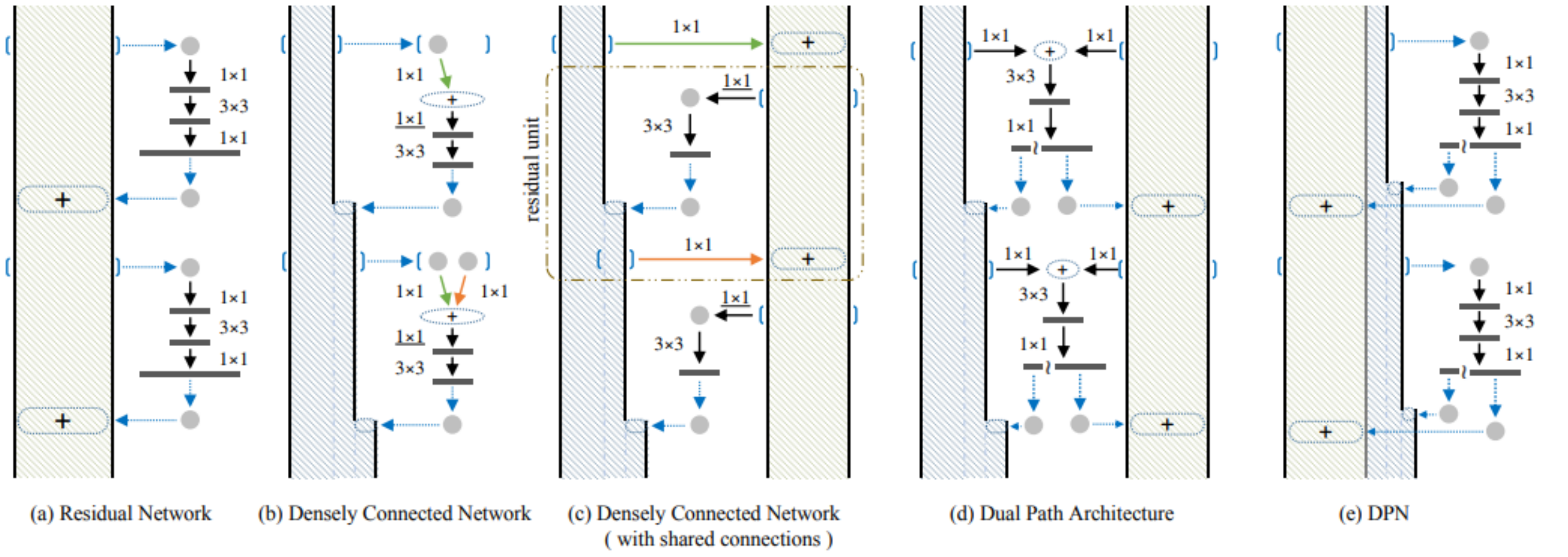
❑ Related Work: PAF



Part affinity fields (PAF) [7] for multi-person human pose estimation. An entire image is taken as the input for a two-branch CNN to jointly predict confidence maps for body part detection, shown in (b), and part affinity fields for parts association, shown in (c). The parsing step performs a set of bipartite matchings to associate body parts candidates (d). These body parts candidates are finally assembled into full body poses for all people in the image (e).

Multi-Person Human Pose Estimation

❑ Related Work: Dual-path networks



Architecture comparison of different networks. (a) The residual network. (b) The densely connected network, where each layer can access the outputs of all previous micro-blocks. Here, a 1×1 convolutional layer (underlined) is added for consistency with the micro-block design in (a). (c) By sharing the first 1×1 connection of the same output across micro-blocks in (b), the densely connected network degenerates to a residual network. The dotted rectangular in (c) highlights the residual unit. (d) The proposed dual path architecture, DPN. (e) An equivalent form of (d) from the perspective of implementation, where the symbol "≀" denotes a split operation, and "+" denotes element-wise addition.

Multi-Person Human Pose Estimation

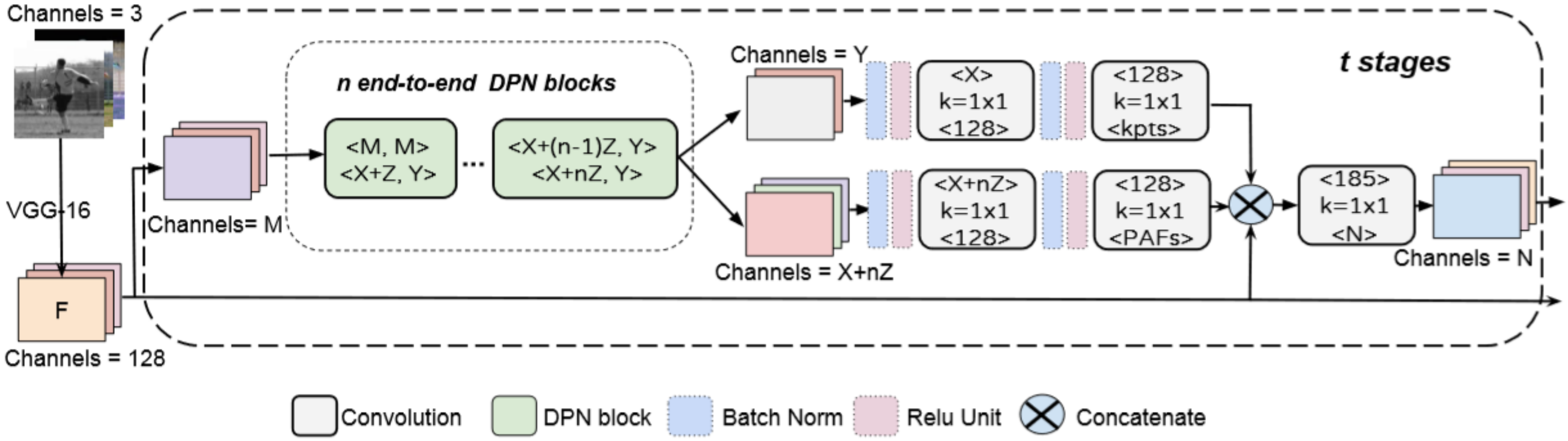
- PoseTrack Challenge Results

Challenge 1: Single-Frame Person Pose Estimation

No.	Entry	MF	Track	AP	Primary Affiliation
1	FractalNet	N	N	62.4151	University of Missouri-Columbia
2	SOPT-PT	N	Y	62.4764	Hikvision Research Institue
3	ML_Lab	N	Y	70.3338	Samsung Research Beijing
4	ProTracker	N	Y	59.5597	CMU
5	SSDHG	N	N	60.0265	South China University of Technology
6	NTHU-test	N	N	38.1329	National Tsing Hua University
7	ICG	N	Y	51.1658	Graz University of Technology
8	BUTDS	N	N	64.4616	The Chinese University of Hong Kong

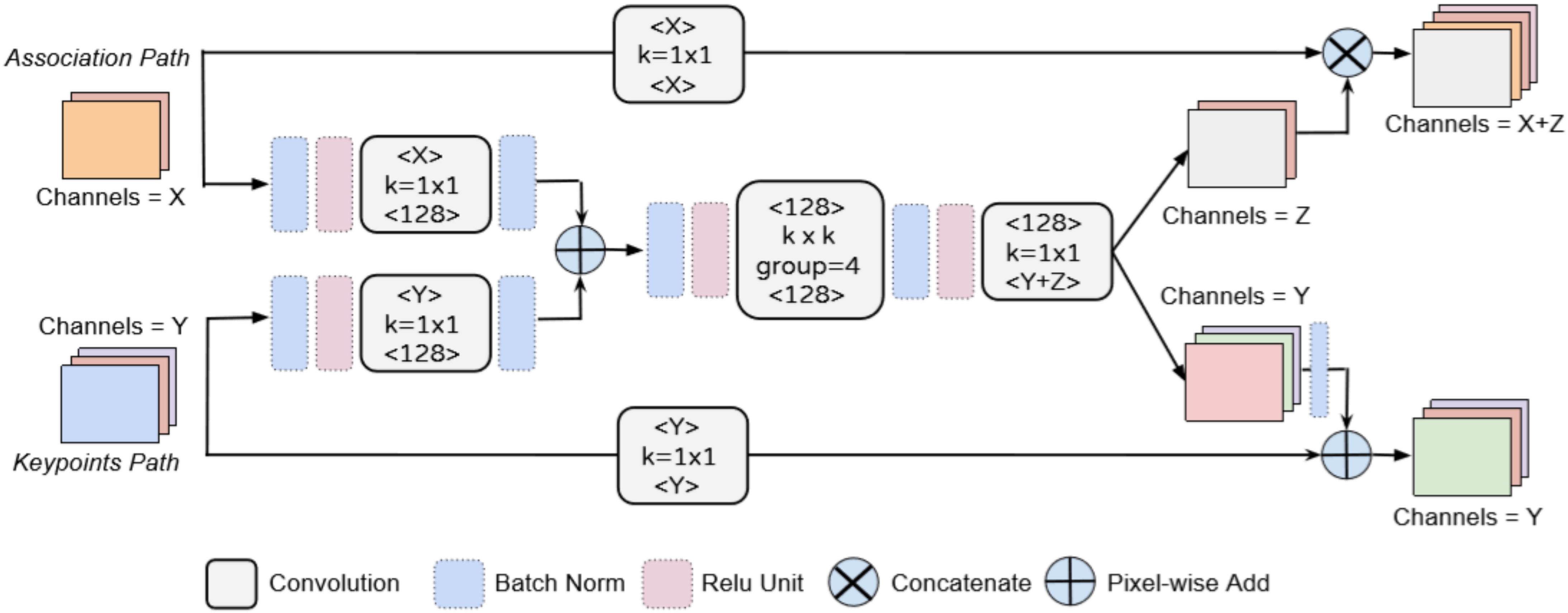
Multi-Person Human Pose Estimation

❑ Proposed Network



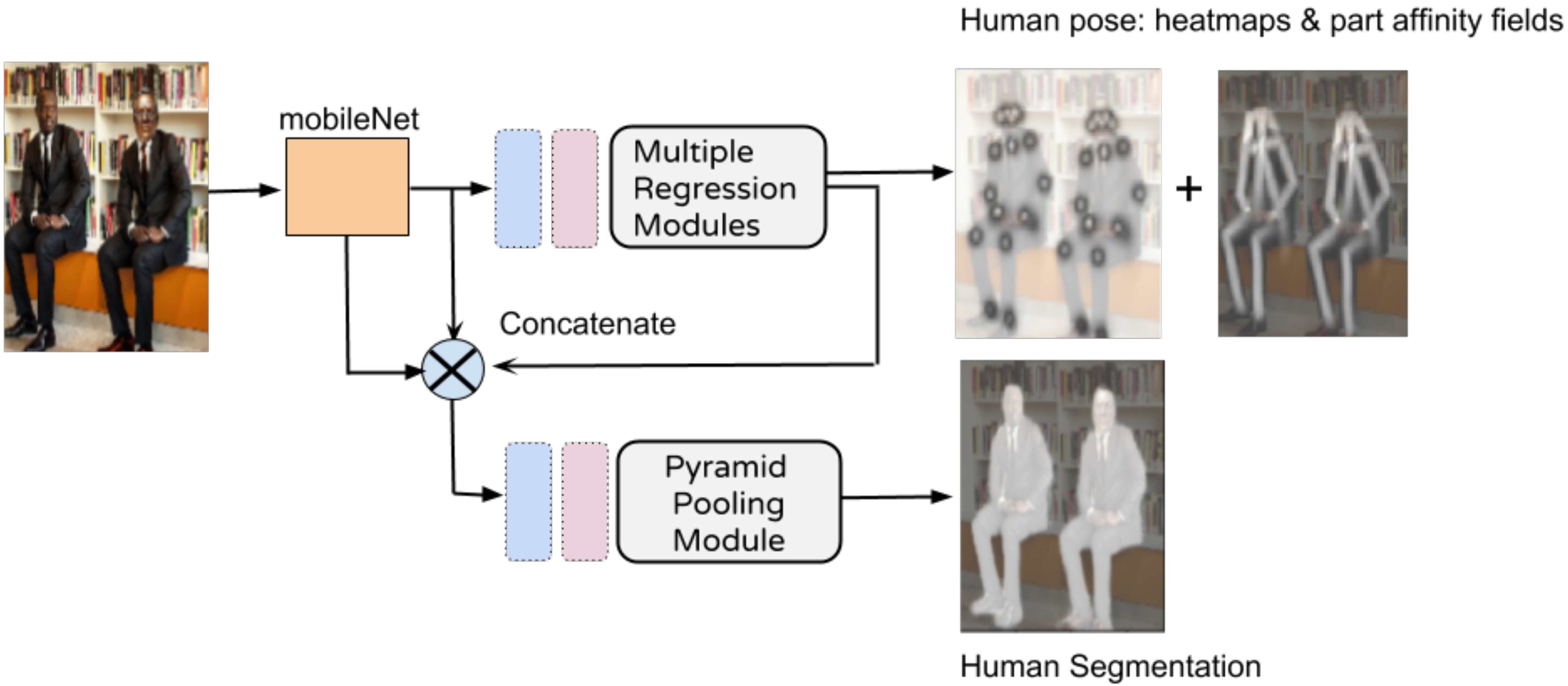
Multi-Person Human Pose Estimation

- Proposed DPN Block



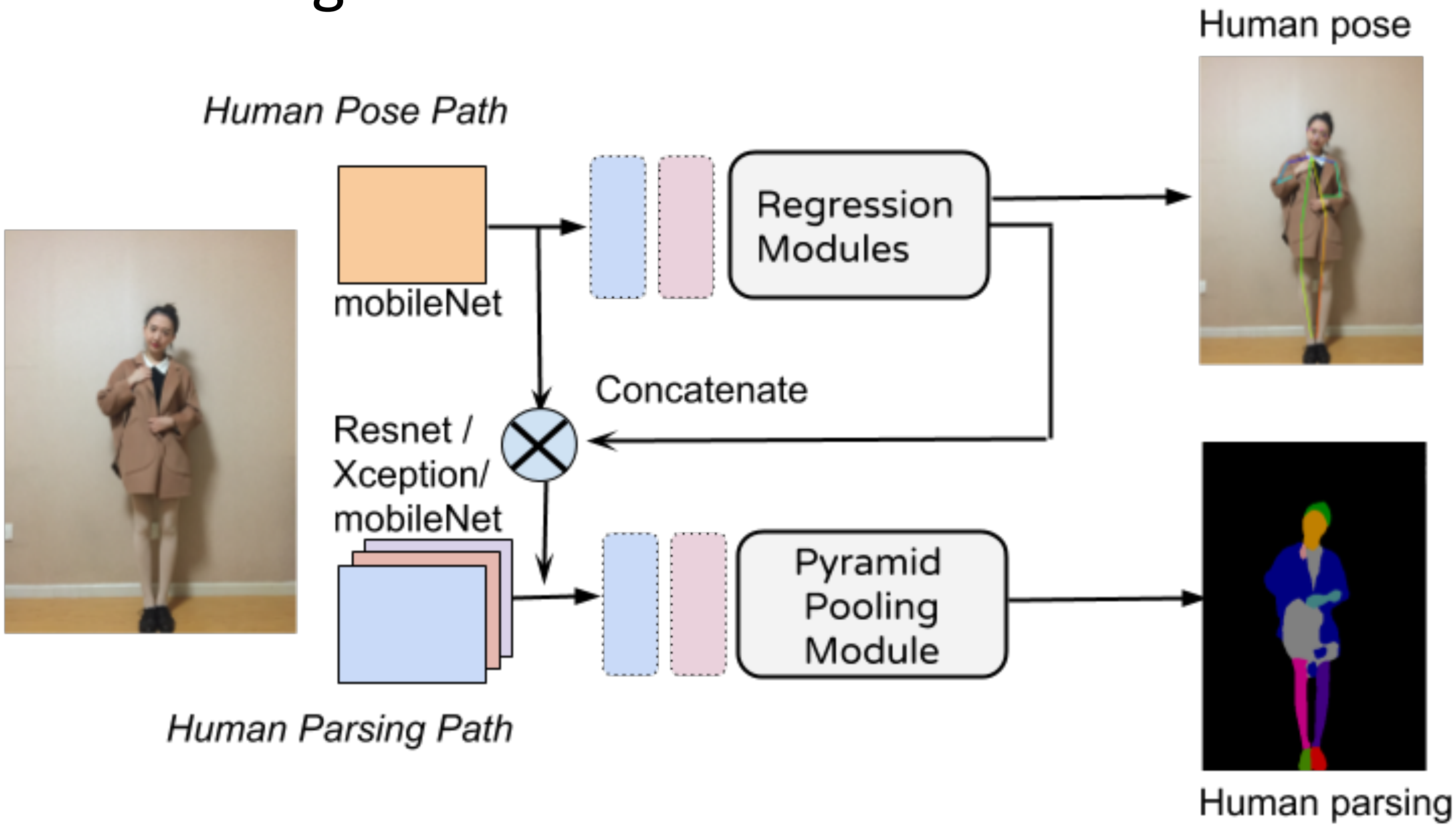
Multi-Person Human Pose Estimation

- ❑ PoSeg: Pose & Segmentation



HPE with Adversarial Training

- Human Parsing



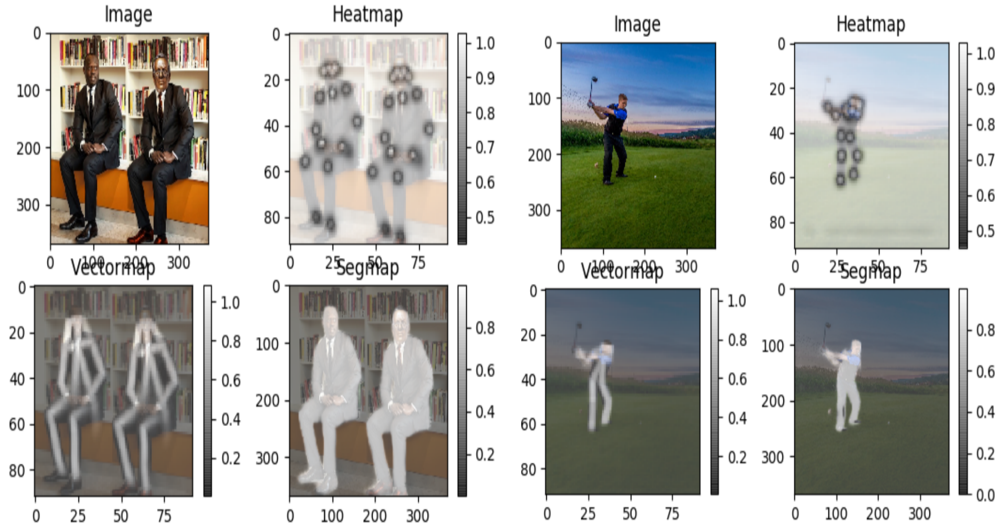
Human Pose Estimation and Applications

Guanghan Ning, JD.COM

What is Human Pose Estimation?

➢ Detect the keypoints of human joints

- ## PoSeg Network Results



- Innovative PoSeg Network:

Joint Pose & Segmentation

(1). Train a single joint network that does two jobs, designed for higher speed and used for AR / mobile applications.

Model training
- › PoSeg on Tensorflow
- › Nvidia TITANx2 GPUs
- › Only 50ms inference
- › Single backbone

(2). Train Human Pose Estimation to aid Human Parsing, aiming at higher accuracy and used for fashion clothing item segmentation and retrieval.

Model training
- › PoSeg on Pytorch
- › Nvidia P40x4 GPUs
- › Two backbones

- ## AR / Mobile Applications

(1) WingAdder:

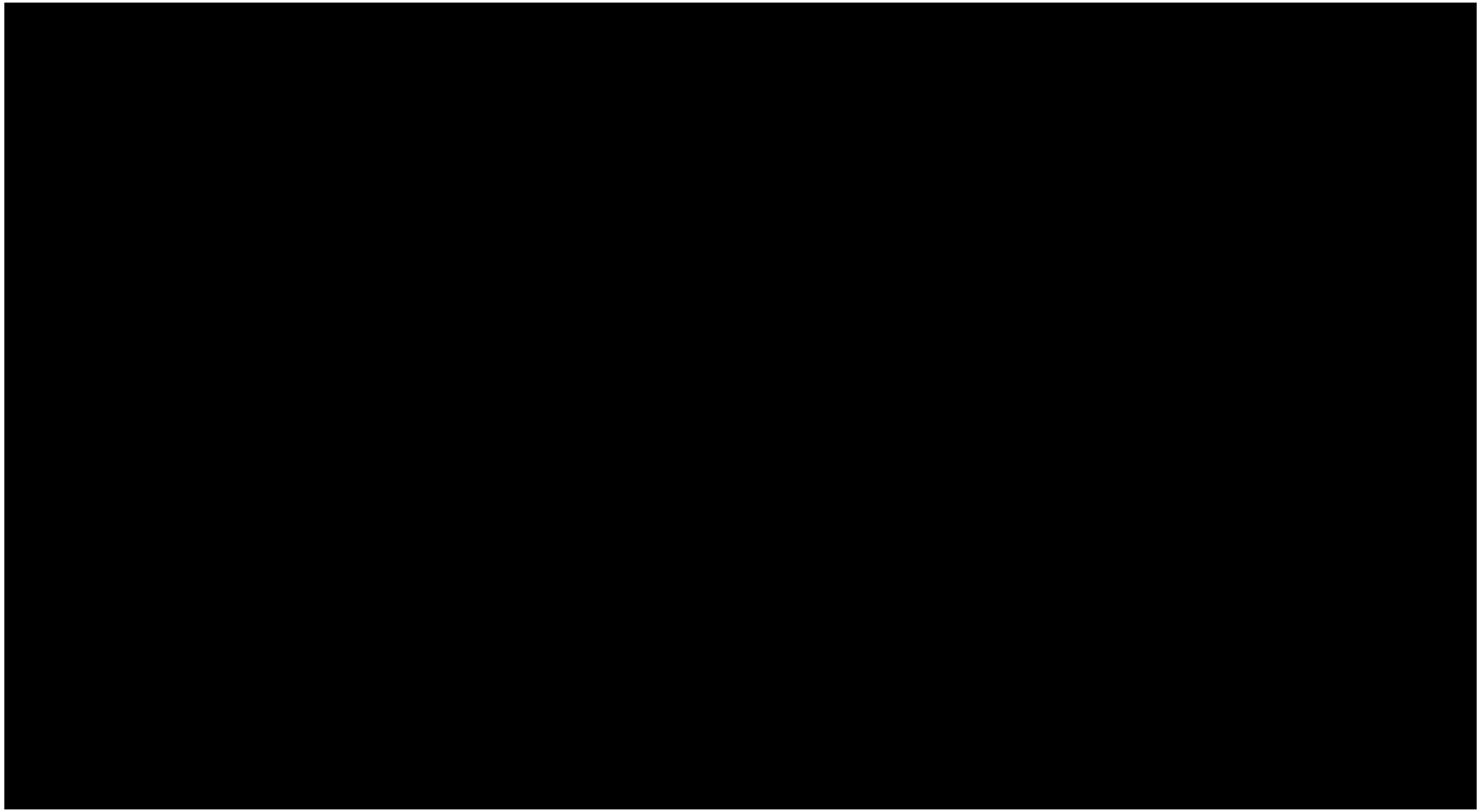
➢ Add Special Effects

(2) Thinner:

➢ Make Person Look Thinner

Pose Estimation DEMO

- Video Demo: https://youtu.be/f5hbo7lnuLI



The End

THANK YOU!