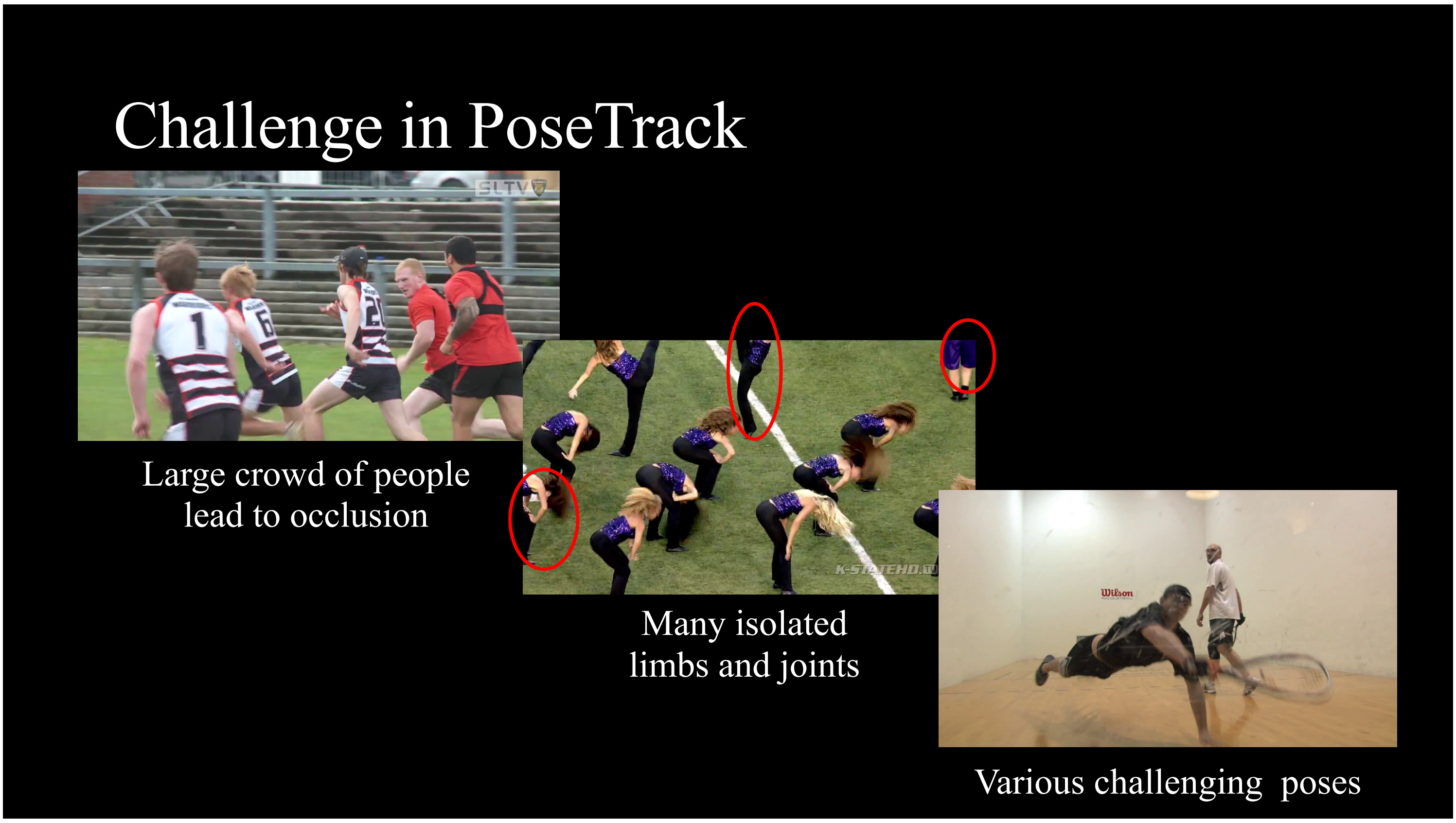
Bottom-up Multi-person Pose Estimation with Multiscale Features

Sheng Jin, Xujie Ma, Wentao Liu, Wei Yang, Chen Qian, Wanli Ouyang
SenseTime Group Limited,
Tsinghua University,
The Chinese University of Hong Kong

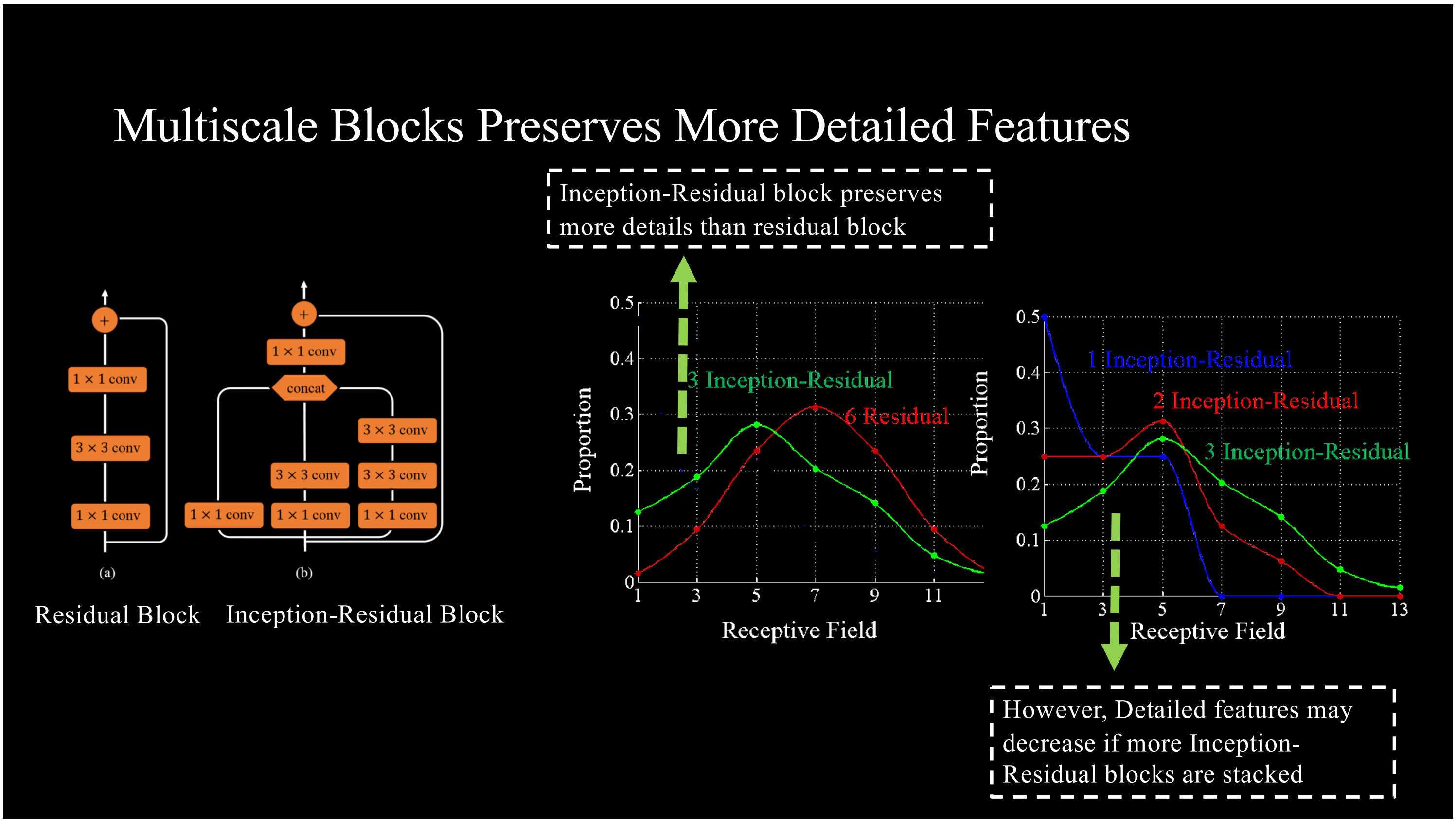

Challenge in PoseTrack
Large crowd of people
lead to occlusion
Many isolated
limbs and joints
Various challenging poses

Inception of Inception Network with Attention Modulated Feature Fusion for Human Pose Estimation

- Motivation
  - Accurate keypoint localization of human pose needs diversified features
    - High level for contextual dependencies
    - Low level for detailed refinement of joints
  - The importance of the two factors varies from case to case

Occluded left wrist needs high level context

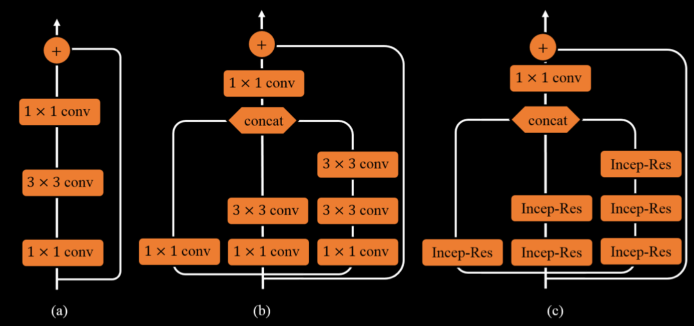
Partially occluded left ankle could be accurate located if more detailed features are preserved



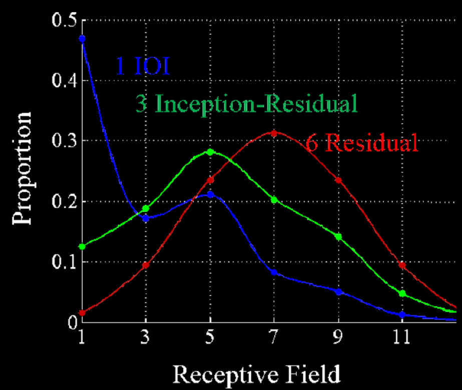
Multiscale Blocks Preserves More Detailed Features
Inception-Residual block preserves more details than residual block
1 × 1 conv
3 × 3 conv
1 × 1 conv
(a)
Residual Block
1 × 1 conv
concat
3 × 3 conv
3 × 3 conv
3 × 3 conv
1 × 1 conv
1 × 1 conv
1 × 1 conv
(b)
Inception-Residual Block
3 Inception-Residual
6 Residual
Proportion
Receptive Field
1 Inception-Residual
2 Inception-Residual
3 Inception-Residual
Proportion
Receptive Field
However, Detailed features may decrease if more Inception-Residual blocks are stacked

Inception of Inception Block

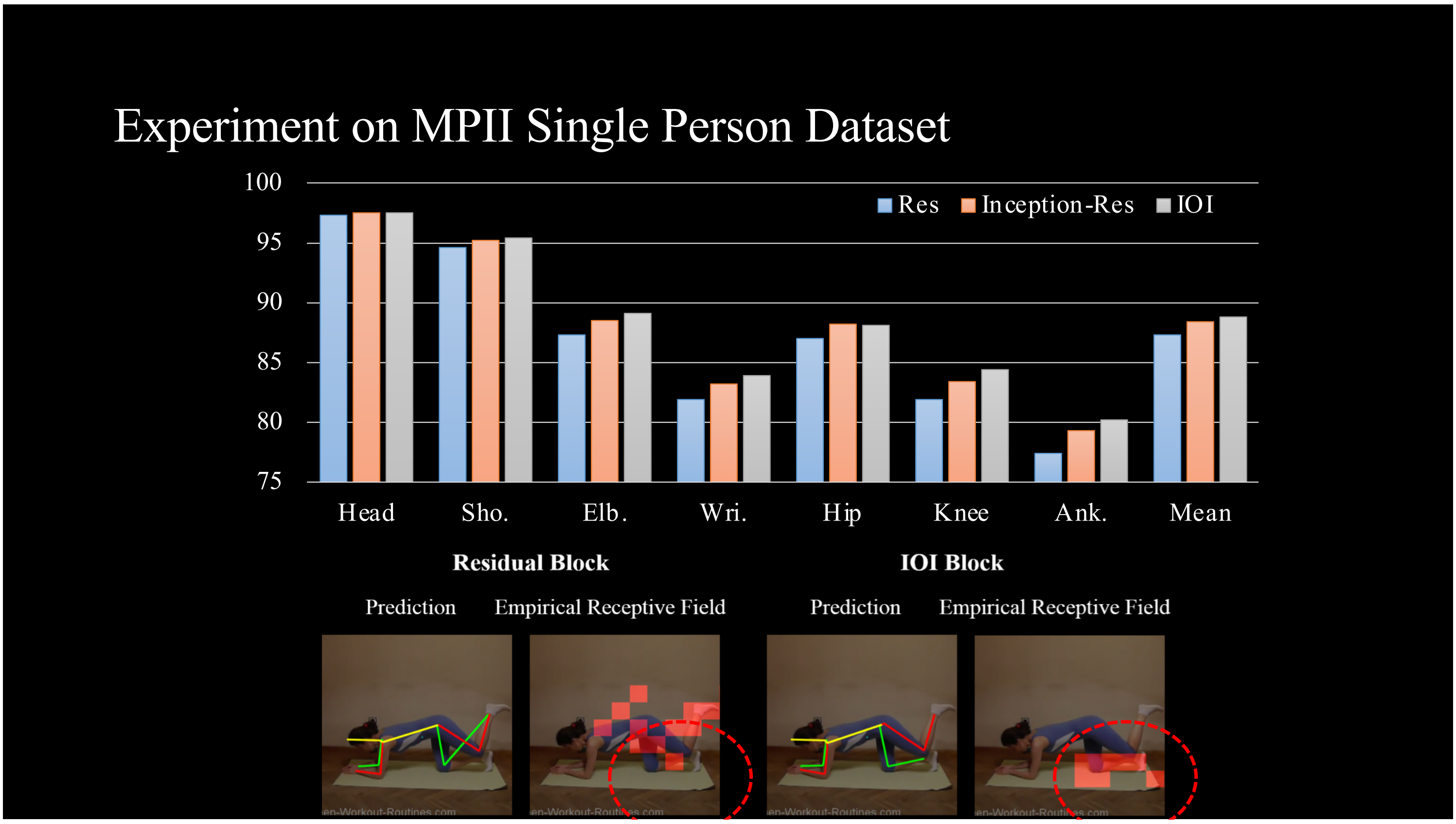
- We proposed Inception of Inception (IOI) Block to preserve scale diversity in deeper network

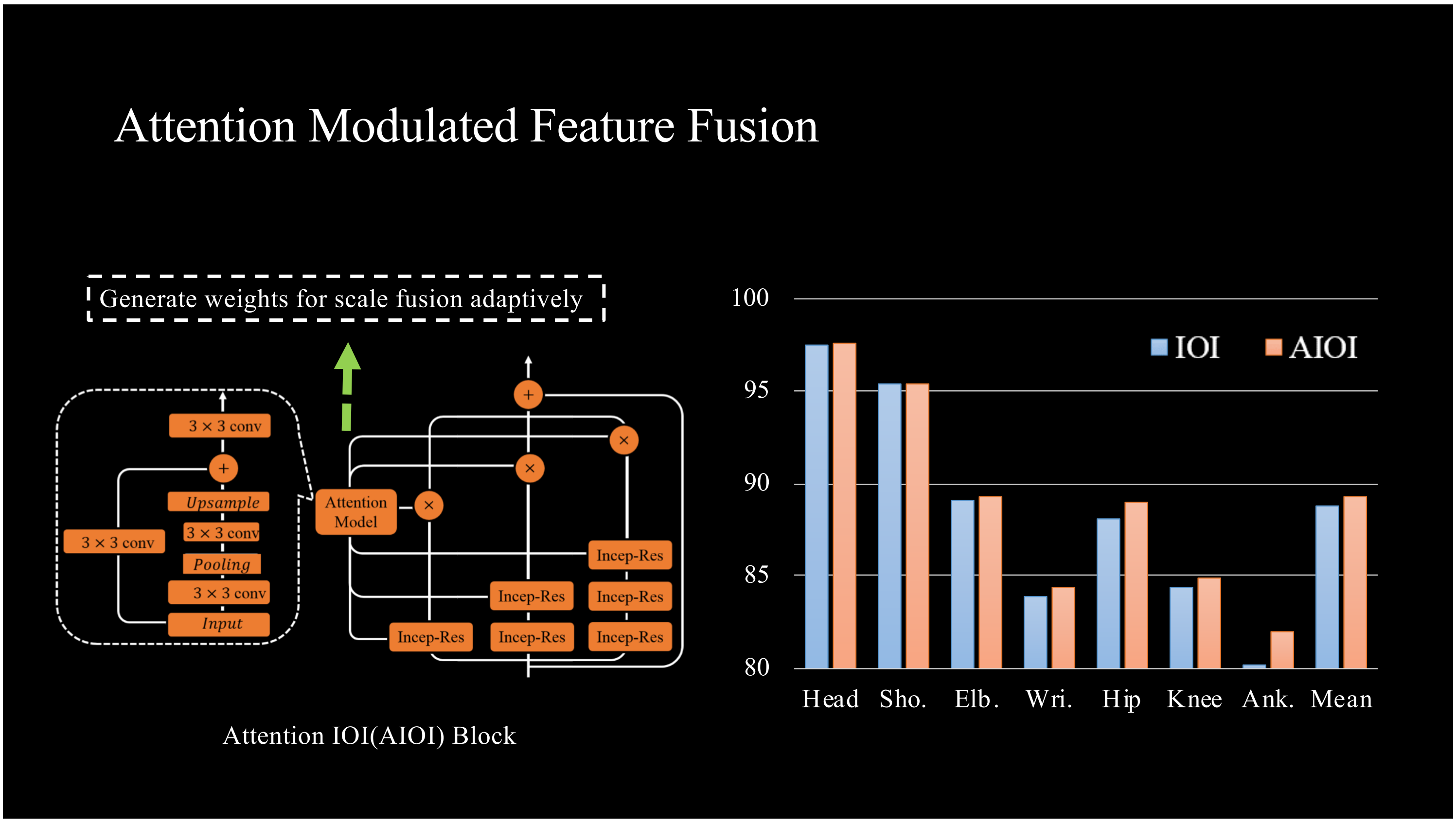


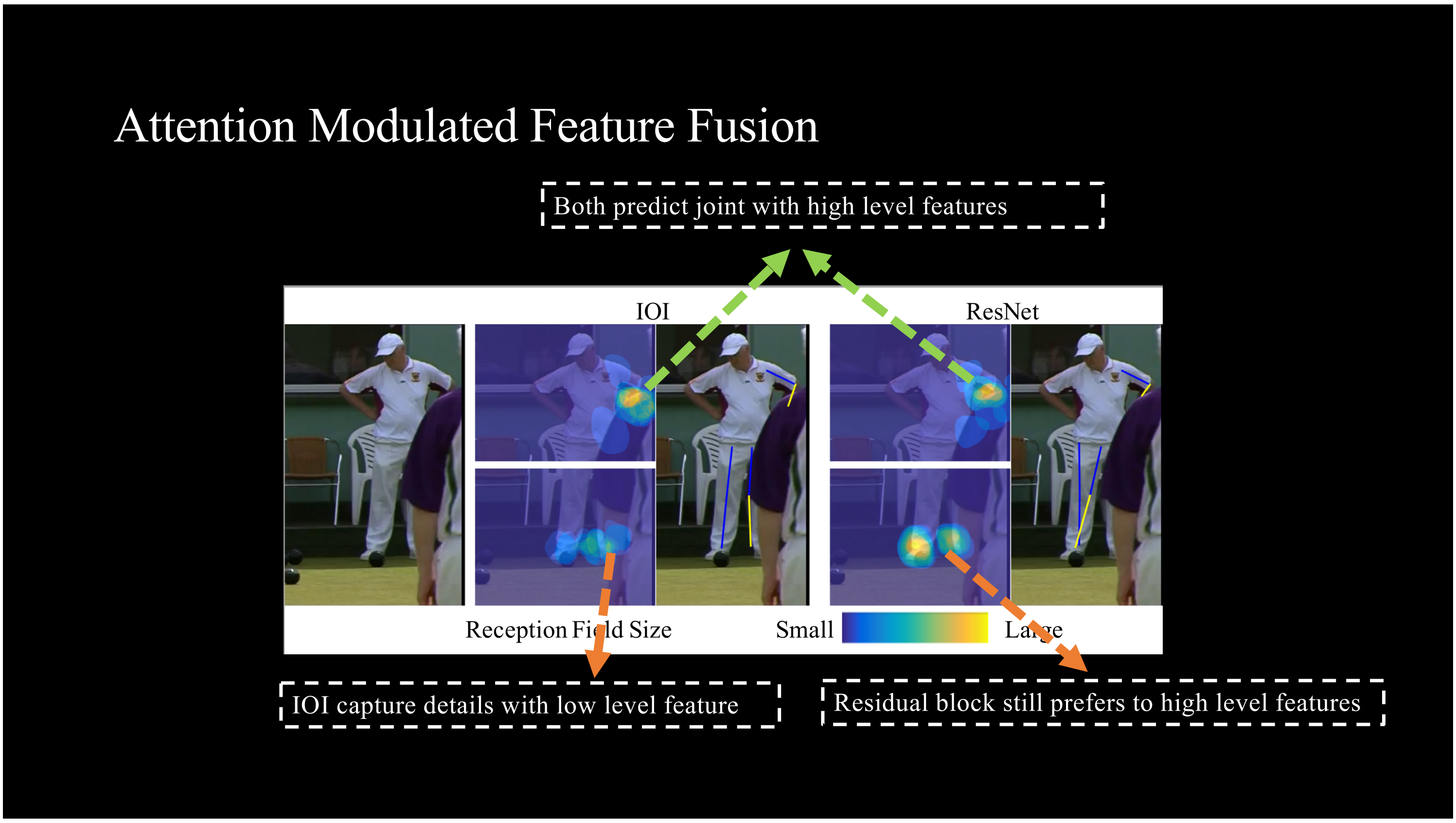
Residual Block Inception-Residual Block IOI



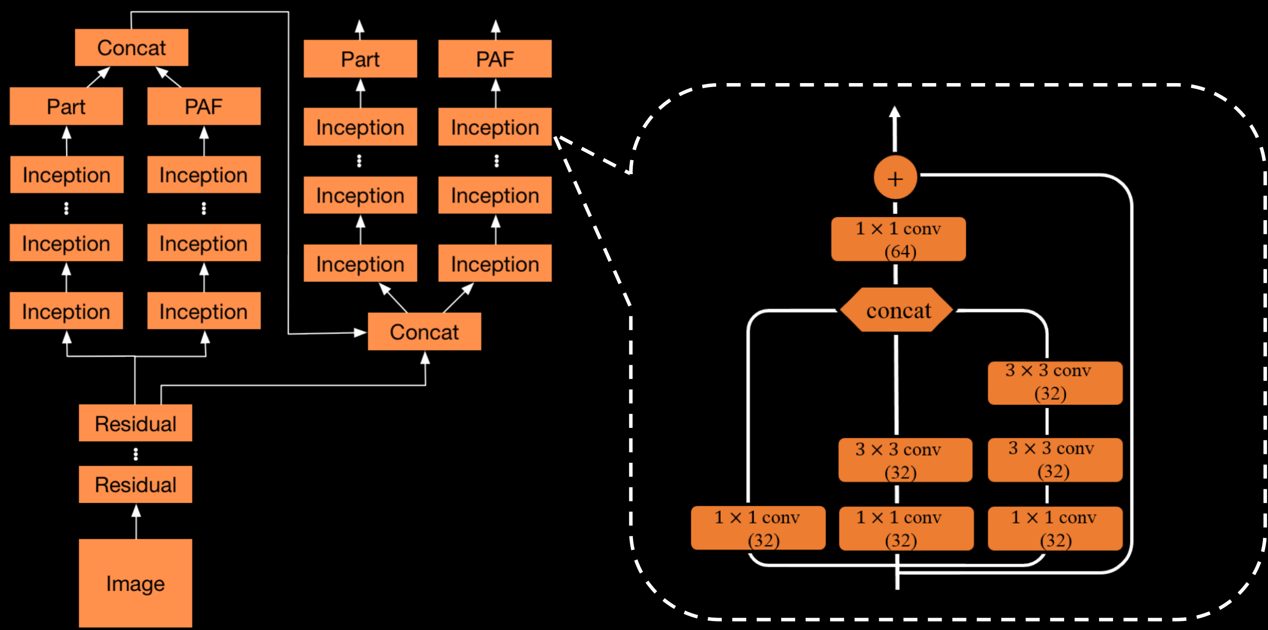
The proposed IOI block presents to have more detailed features among all the blocks and is able to construct deeper human pose estimation network


Experiment on MPII Single Person Dataset
100
95
90
85
80
75
Res
Inception-Res
IOI
Head
Sho.
Elb.
Wri.
Hip
Knee
Ank.
Mean
Residual Block
IOI Block
Prediction
Empirical Receptive Field
Prediction
Empirical Receptive Field


Attention Modulated Feature Fusion
Generate weights for scale fusion adaptively
3 × 3 conv
+
Upsample
3 × 3 conv
3 × 3 conv
Pooling
3 × 3 conv
Input
Attention Model
×
+
Incep-Res
Incep-Res
Incep-Res
Incep-Res
Incep-Res
Incep-Res
Attention IOI(AIOI) Block
100
95
90
85
80
IOI
AIOI
Head
Sho.
Elb.
Wri.
Hip
Knee
Ank.
Mean


Attention Modulated Feature Fusion
Both predict joint with high level features
IOI
ResNet
Reception Field Size
Small
Large
IOI capture details with low level feature
Residual block still prefers to high level features

Multiscale Part Affinity Fields (MSPAF)



Experiment of Bottom-up Methods on COCO

Table 1. Comparison of paf and our proposed multiscale paf method

Methods	Parameter Number(M)	TFLOPS	Forward Time on TitanX (ms)	mAP
PAF 6 Stages	52.31	0.05	104.91	58.50
MSPAF 3 Stage	9.66	0.03	56.96	61.10

Results of Top-down and Bottom-up Methods on COCO

- Top-down method outperforms bottom-up method on COCO

Table 2. Comparison of Top-down and Bottom-up method on COCO

	Methods	**mAP**
Bottom-up	PAF 6 Stages[1]	58.5
	MSPAF 3stage	61.1
Top-down	Mask R-CNN[2]	62.7
	Mask R-CNN (our implementation)	63.1

1. Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *arXiv preprint arXiv:1611.08050* (2016).
2. He, Kaiming, et al. "Mask r-cnn." *arXiv preprint arXiv:1703.06870* (2017).

Comparison of Bottom-up and Top-Down Methods on PoseTrack

- Bottom-up method outperforms top-down method on PoseTrack

Table 3. Experiment results on PoseTrack

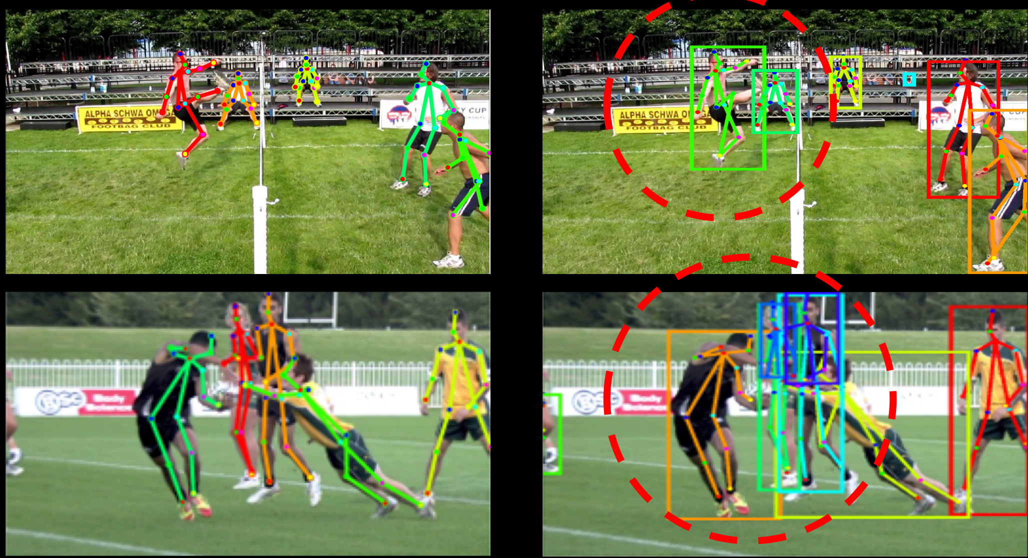
	Methods	**mAP**
Bottom-up	MSPAF	67.8
Top-down	Mask R-CNN	57.4

Failure Cases of Top-down method

Bottom-up | Top-down

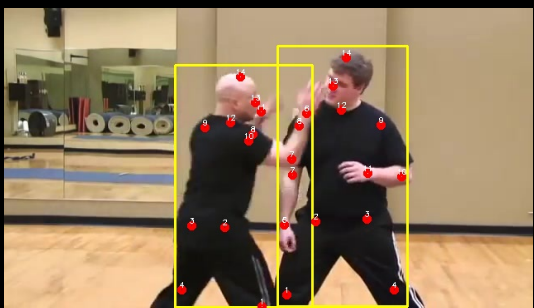
Complex poses

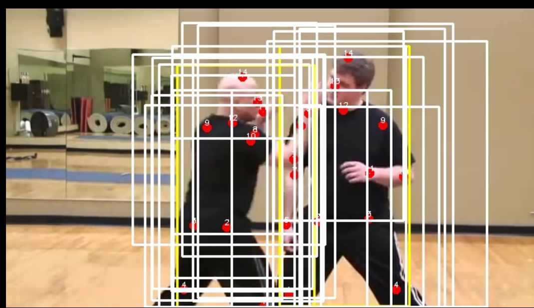
Crowding



Person detection + Single-person Pose Estimation

- Train with detection boxes
  - Inaccurate person bounding boxes
  - Poor performance
- Train with RPN proposal boxes
  - Data augmentation
  - Better performance

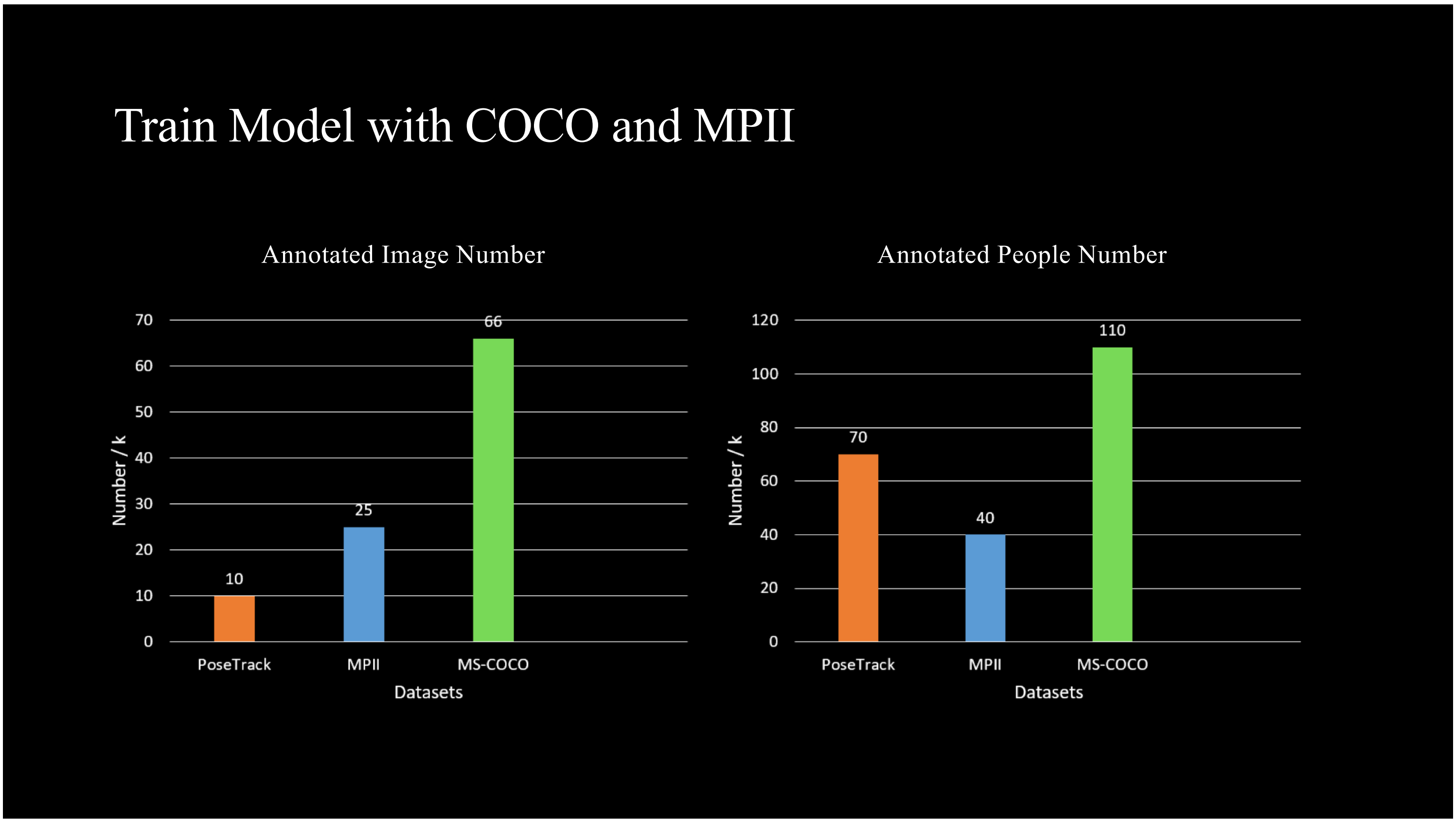




Person detection + Single-person Pose Estimation

Table 4. Improvement on Top-down based method

	Methods	**mAP**
Bottom-up	MSPAF	67.8
Top-down	Mask R-CNN	57.4
	Person detection + Single-person Stacked Hourglass	60.3


Train Model with COCO and MPII
Annotated Image Number
70
60
50
40
30
20
10
0
Number / k
10
25
66
PoseTrack
MPII
MS-COCO
Datasets
Annotated People Number
120
100
80
60
40
20
0
Number / k
70
40
110
PoseTrack
MPII
MS-COCO
Datasets

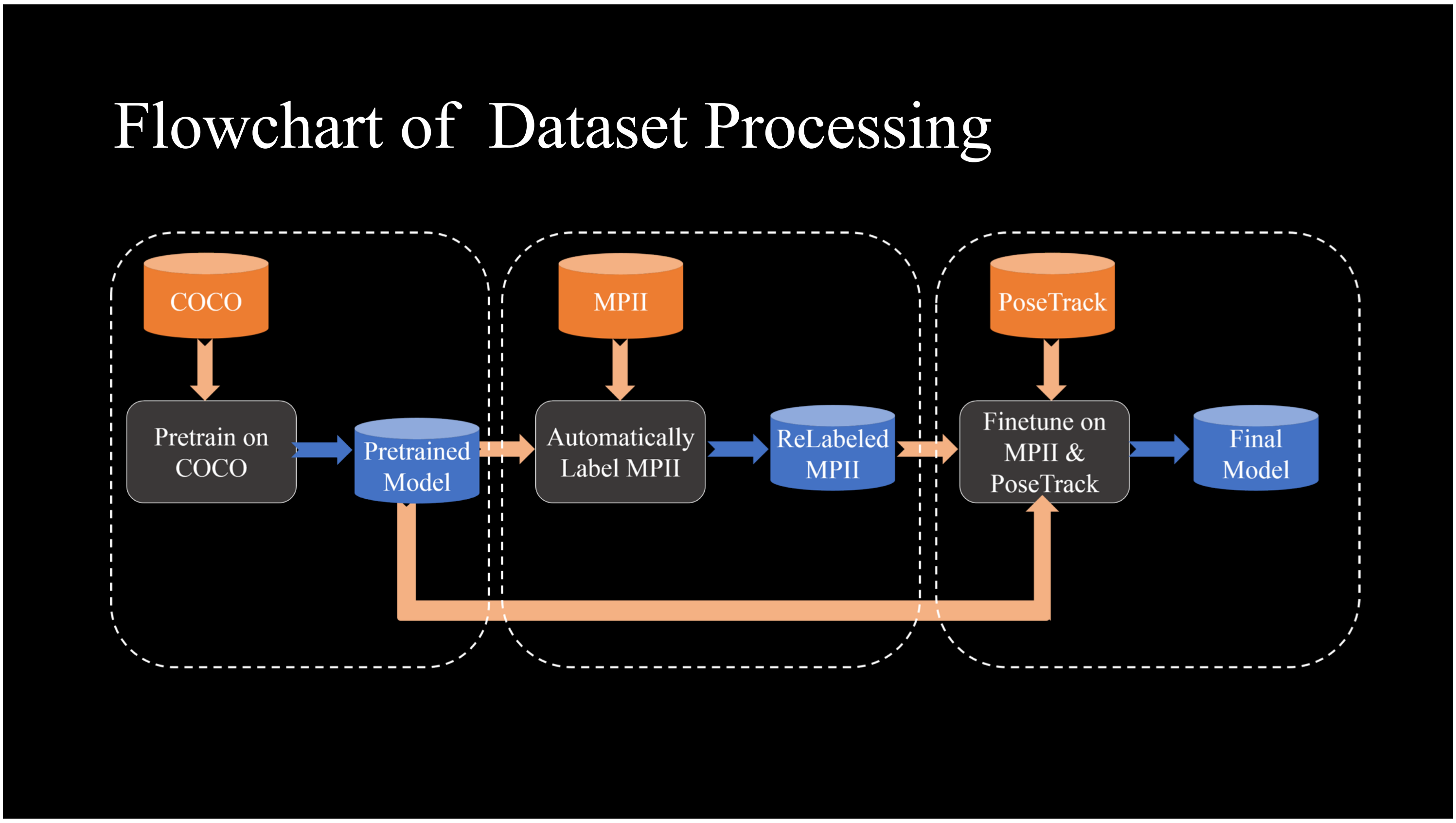
Handling Annotation Difference

• Annotation difference between three dataset

PoseTrack: 15 joints
(head-top, nose)

MPII: 14 joints
(head-top)

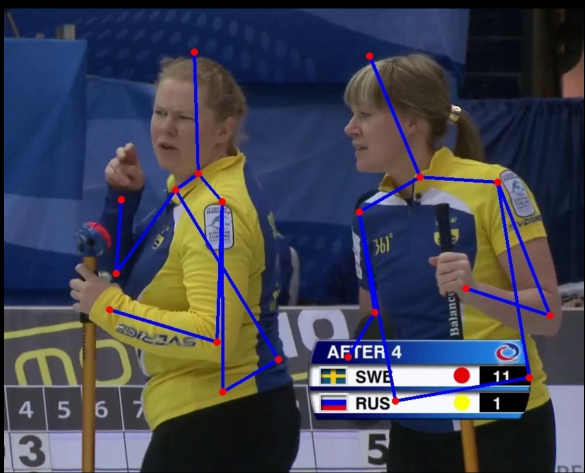
MS-COCO: 17 joints
(nose, eyes, ears)

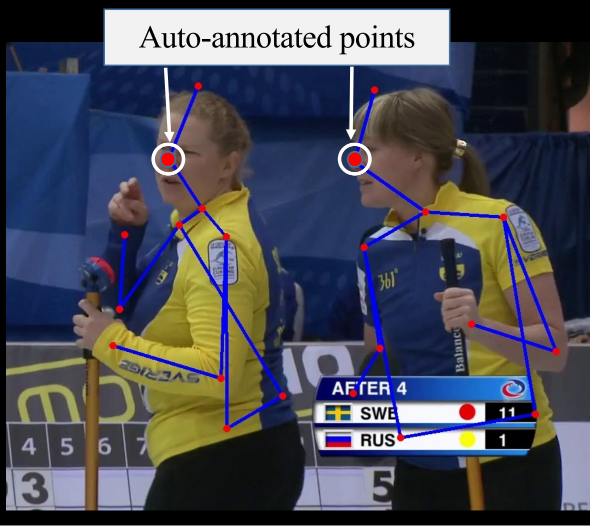


Flowchart of Dataset Processing
COCO
Pretrain on COCO
Pretrained Model
MPII
Automatically Label MPII
ReLabeled MPII
PoseTrack
Finetune on MPII & PoseTrack
Final Model

Handling Annotation Difference

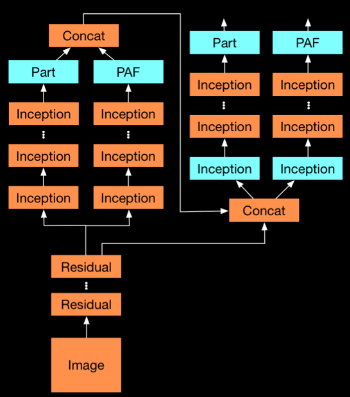
- Align MPII with PoseTrack
    - Automatically annotate nose point on MPII using model trained with MS-COCO dataset





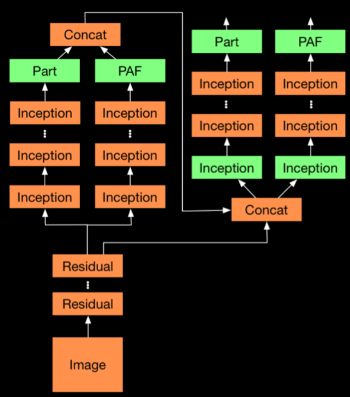
Handling Annotation Difference

Finetune on MPII and PoseTrack from COCO



Pretrain on COCO

Share Weights



Finetune on PoseTrack and MPII

Improvements to Performance

Table 5. Experiment results on combination of different dataset

Dataset	Iteration(w)	mAP
PoseTrack	5	37.5
PoseTrack+MPII+COCO	5	63.8
PoseTrack+MPII+COCO	11	67.8

Examples

Thanks for your attention!