Human Pose Estimation and Action Recognition

Gang Yu, Megvii (Face++)

Junsong Yuan, SUNY Buffalo

Zicheng Liu, Microsoft

Overview

- Part1: Human Pose Estimation
  - 2D Skeleton
    - Top-Down
    - Bottom-Up
  - 3D Skeleton
    - 2D -> 3D Skeleton
    - 2D -> 3D Shape
  - Application

- Part2: Action Recognition
  - Datasets
    - RGB
    - RGB-D
  - Skeleton based approaches
    - 2D and 3D skeletons
  - Video based approaches
    - 2D/3D CNN features

Human Pose Estimation Algorithm and Application

Gang Yu
yugang@megvii.com

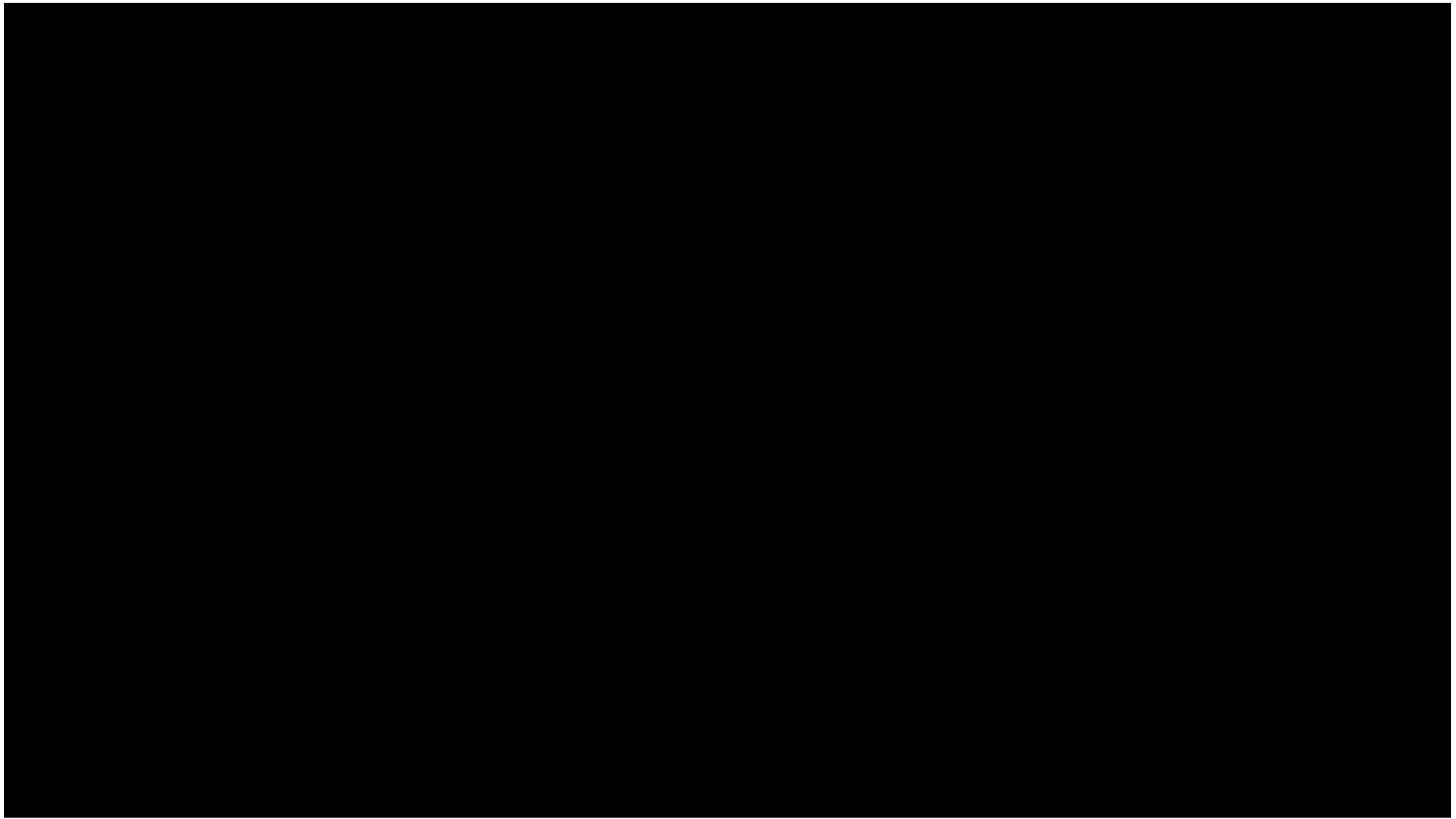
Outline

- Introduction to Human Pose Estimation
- 2D Skeleton
  - Top-Down
  - Bottom-Up
- 3D Skeleton
  - 2D -> 3D Skeleton
  - 2D -> 3D Shape
- Application
- Conclusion

Outline

- Introduction to Human Pose Estimation
- 2D Skeleton
  - Top-Down
  - Bottom-Up
- 3D Skeleton
  - 2D -> 3D Skeleton
  - 2D -> 3D Shape
- Application
- Conclusion

What is Human Pose Estimation?



Benchmark and Evaluation

- Benchmark
  - Single-person Estimation
    - MPII, FLIC, LSP, LIP
  - Multi-person Keypoint Detection
    - COCO, CrowdPose
  - Video
    - PoseTrack
  - 3D
    - Human3.6M, DensePose
- Evaluation on COCO

$$OKS = \Sigma_i[\exp(-d_i^2/2s^2\kappa_i^2)\delta(v_i>0)] / \Sigma_i[\delta(v_i>0)]$$

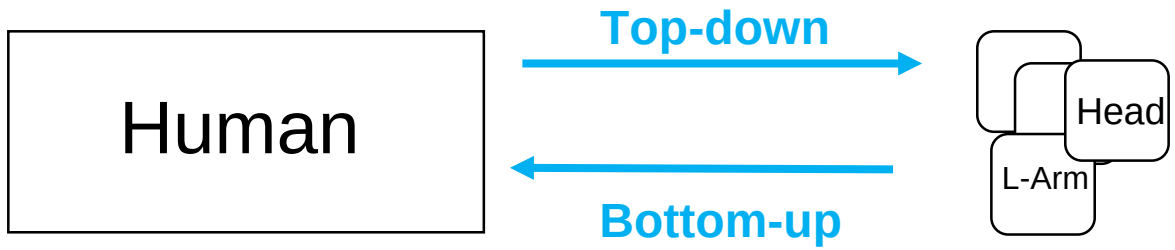
```
Average Precision (AP):
  AP                  % AP at OKS=.50:.05:.95 (primary challenge metric)
  AP^OKS=.50          % AP at OKS=.50 (loose metric)
  AP^OKS=.75          % AP at OKS=.75 (strict metric)
AP Across Scales:
  AP^medium           % AP for medium objects: 32^2 < area < 96^2
  AP^large            % AP for large objects: area > 96^2
Average Recall (AR):
  AR                  % AR at OKS=.50:.05:.95
  AR^OKS=.50          % AR at OKS=.50
  AR^OKS=.75          % AR at OKS=.75
AR Across Scales:
  AR^medium           % AR for medium objects: 32^2 < area < 96^2
  AR^large            % AR for large objects: area > 96^2
```

Outline

- Introduction to Human Pose Estimation
- 2D Skeleton
  - Top-Down
  - Bottom-Up
- 3D Skeleton
  - 2D -> 3D Skeleton
  - 2D -> 3D Shape
- Application
- Conclusion

2D Skeleton: How to Do Pose Estimation

- Top-down Approach VS Bottom-up Approach



  - Top-down
    - Mask R-CNN, CPN, MSPN
    - High Performance (good localization ability), High Recall
  - Bottom-up
    - Openpose, Associative Embeding
    - Clean framework, potentially fast speed

Mask R-CNN, Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, ICCV 2018
Cascaded Pyramid Network for Multi-Person Pose Estimation, Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, Jian Sun, CVPR 2018
Rethinking on Multi-Stage Networks for Human Pose Estimation, Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, Jian Sun
OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh,
Associative Embedding: End-to-End Learning for Joint Detection and Grouping, Alejandro Newell, Zhiao Huang, Jia Deng, NIPS 2017

Challenges

- Ambiguous Appearance
- Crowd Case
- Large Pose
- Inference Speed

Top-Down: Mask R-CNN

- Motivation:
  - Multi-task learning
  - ROI Pool -> ROI Align

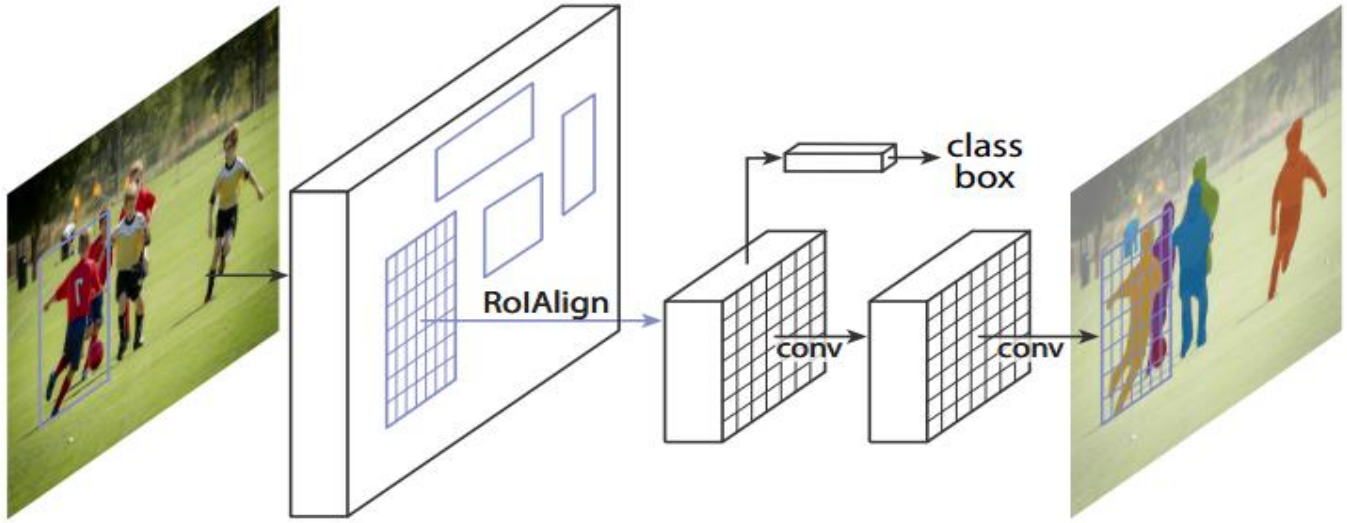


Figure 1. The **Mask R-CNN** framework for instance segmentation.

Mask R-CNN, Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, ICCV 2017

Top-Down: Mask R-CNN

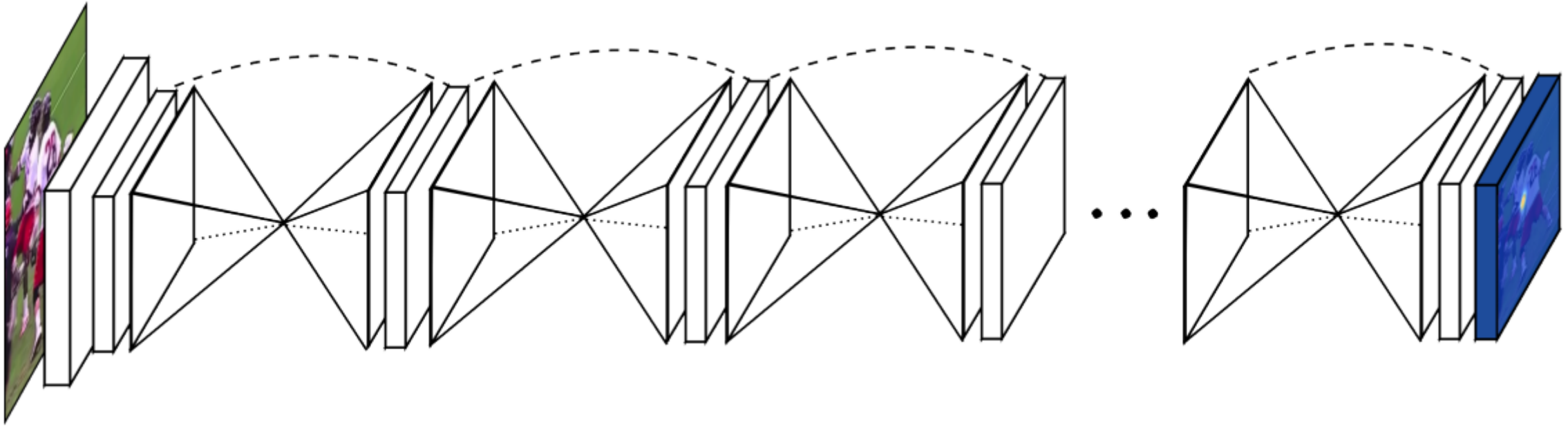
- Experiments on COCO Skeleton:

	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
CMU-Pose+++ [6]	61.8	84.9	67.5	57.1	68.2
G-RMI [32]†	62.4	84.0	68.5	**59.1**	68.1
Mask R-CNN, keypoint-only	62.7	87.0	68.4	57.4	71.1
Mask R-CNN, keypoint & mask	**63.1**	**87.3**	**68.7**	57.8	**71.4**

Mask R-CNN, Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, ICCV 2017

Top-Down: Hourglass

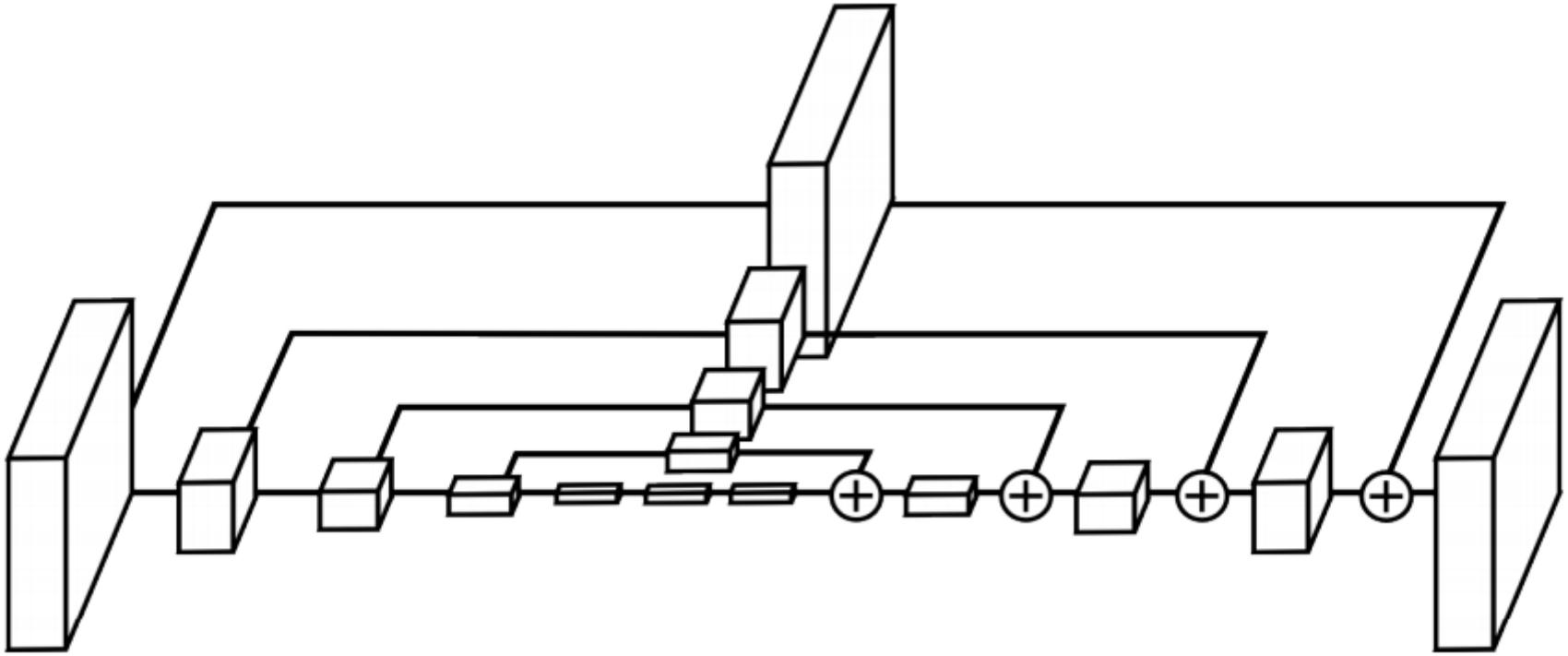
- Motivation:
  - Crop & Single Person Skeleton
  - Multi-stage context refinement



Stacked Hourglass Networks for Human Pose Estimation, Alejandro Newell, Kaiyu Yang, and Jia Deng, ECCV 2016

Top-Down: Hourglass

- Structure of a one block



Stacked Hourglass Networks for Human Pose Estimation, Alejandro Newell, Kaiyu Yang, and Jia Deng, ECCV 2016

Top-Down: Hourglass

- Experiments

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al. [16], CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al. [19], CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al. [17], CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al. [27], CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al. [18], CVPR'16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	**98.2**	**96.3**	**91.2**	**87.1**	**90.1**	**87.4**	**83.6**	**90.9**

Table 2. Results on MPII Human Pose (PCKh@0.5)

Stacked Hourglass Networks for Human Pose Estimation, Alejandro Newell, Kaiyu Yang, and Jia Deng, ECCV 2016

Top-Down: Single Person Skeleton: CPM

- Motivation:
  - Multi-stage context refinement
  - Large receptive Field -> long range spatial relationship

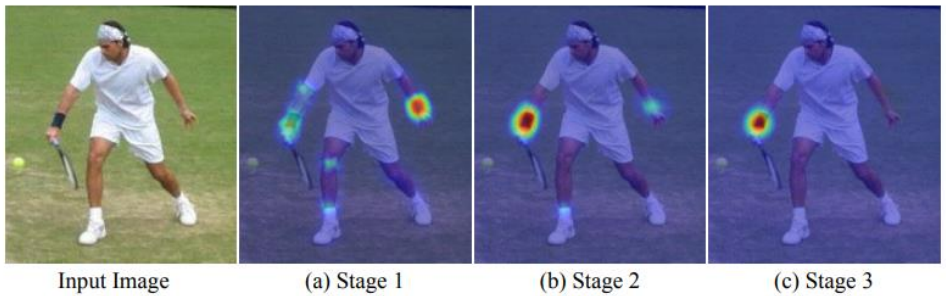


Figure 1: A **Convolutional Pose Machine** consists of a sequence of predictors trained to make dense predictions at each image location. Here we show the increasingly refined estimates for the location of the *right elbow* in each stage of the sequence. (a) Predicting from local evidence often causes confusion. (b) Multi-part context helps resolve ambiguity. (c) Additional iterations help converge to a certain solution.

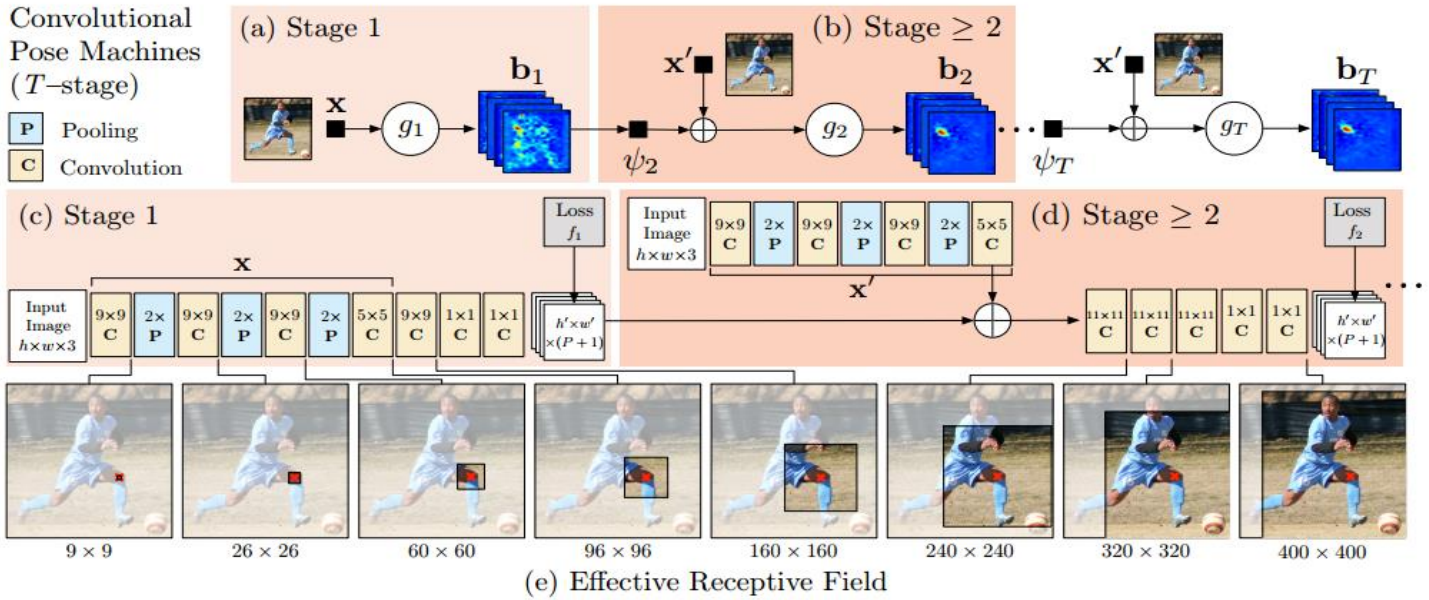
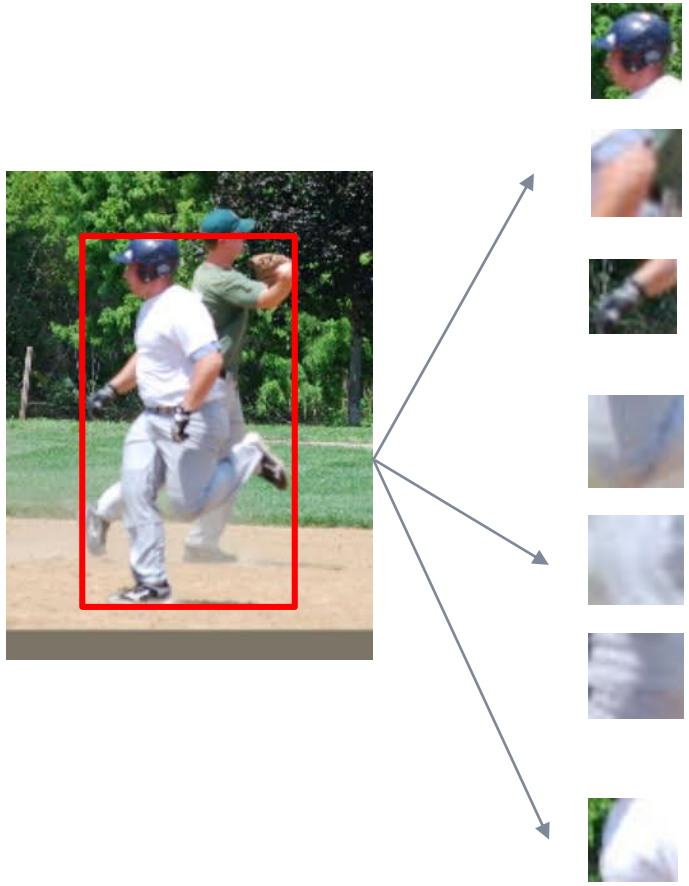


Figure 2: Architecture and receptive fields of CPMs. We show a convolutional architecture and receptive fields across layers for a CPM with any T stages. The pose machine [29] is shown in insets (a) and (b), and the corresponding convolutional networks are shown in insets (c) and (d). Insets (a) and (c) show the architecture that operates only on image evidence in the first stage. Insets (b) and (d) shows the architecture for subsequent stages, which operate both on image evidence as well as belief maps from preceding stages. The architectures in (b) and (d) are repeated for all subsequent stages (2 to T). The network is locally supervised after each stage using an intermediate loss layer that prevents vanishing gradients during training. Below in inset (e) we show the effective receptive field on an image (centered at left knee) of the architecture, where the large receptive field enables the model to capture long-range spatial dependencies such as those between head and knees. (Best viewed in color.)

Convolutional Pose Machines, Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, CVPR 2016

Top-Down: Cascade Pyramid Network

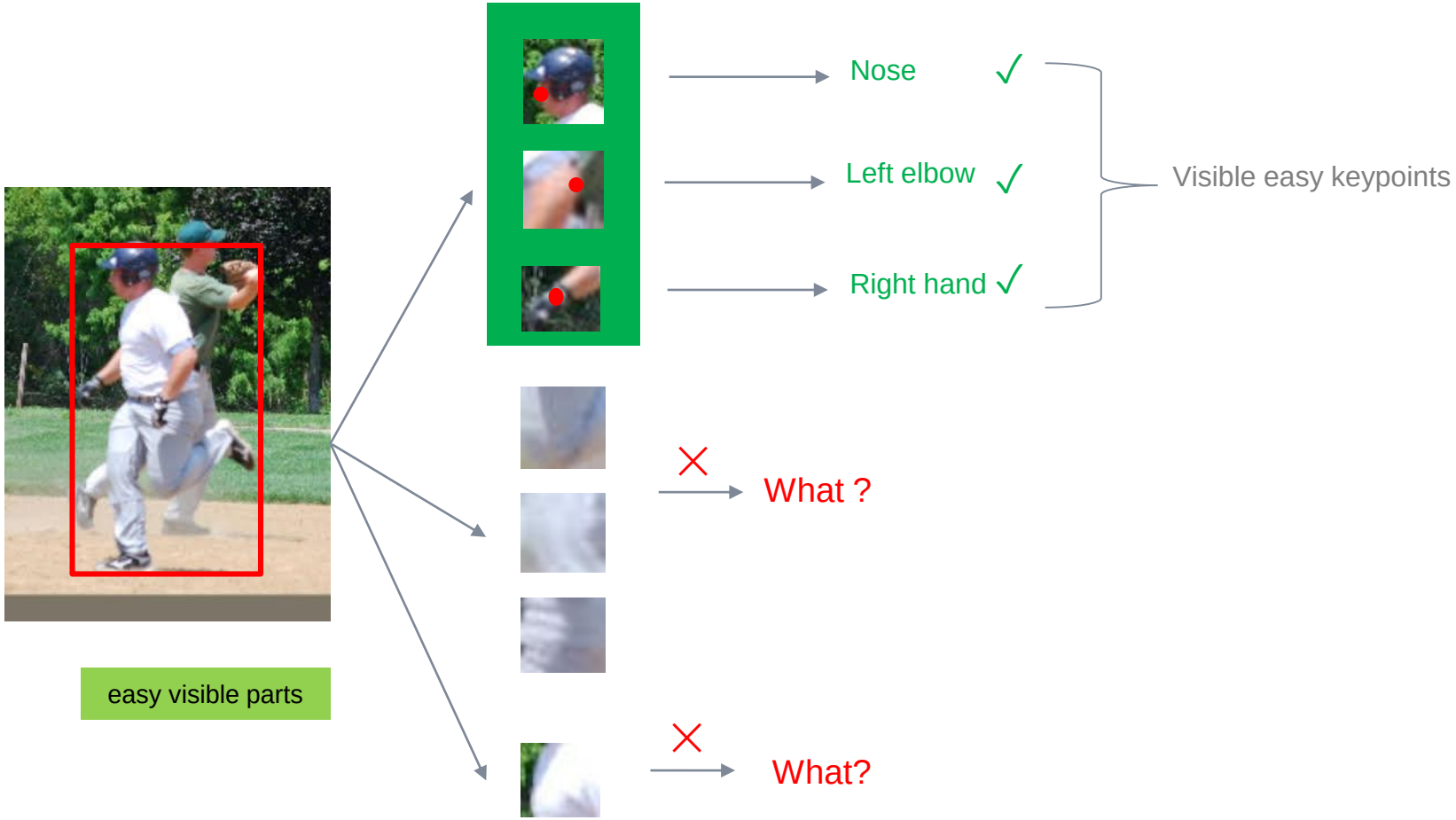
- Motivation: How to locate the “hard” joints
- Human perspective



Cascaded Pyramid Network for Multi-Person Pose Estimation, Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, Jian Sun, CVPR 2018

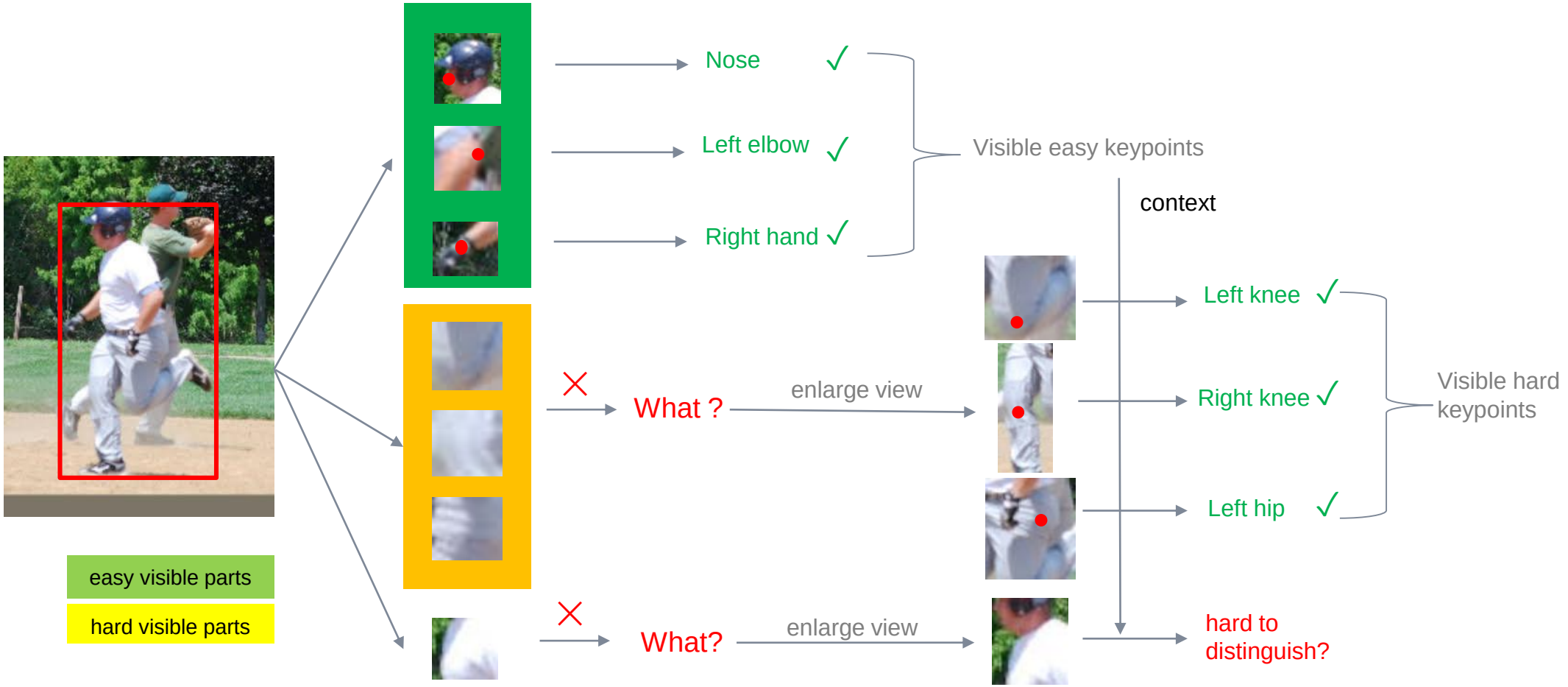
Top-Down: Cascade Pyramid Network

- Motivation: How to locate the “hard” joints
- Human perspective



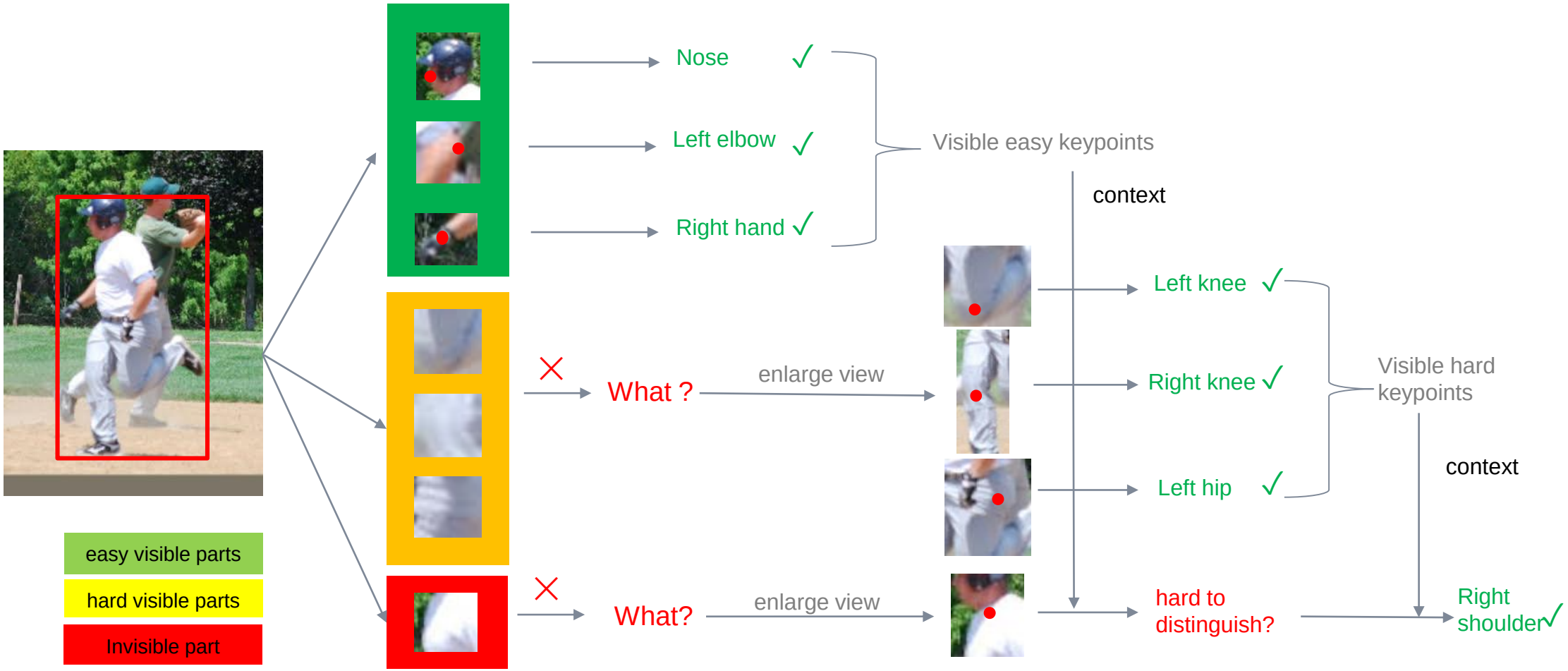
Top-Down: Cascade Pyramid Network

- Motivation: How to locate the “hard” joints
- Human perspective



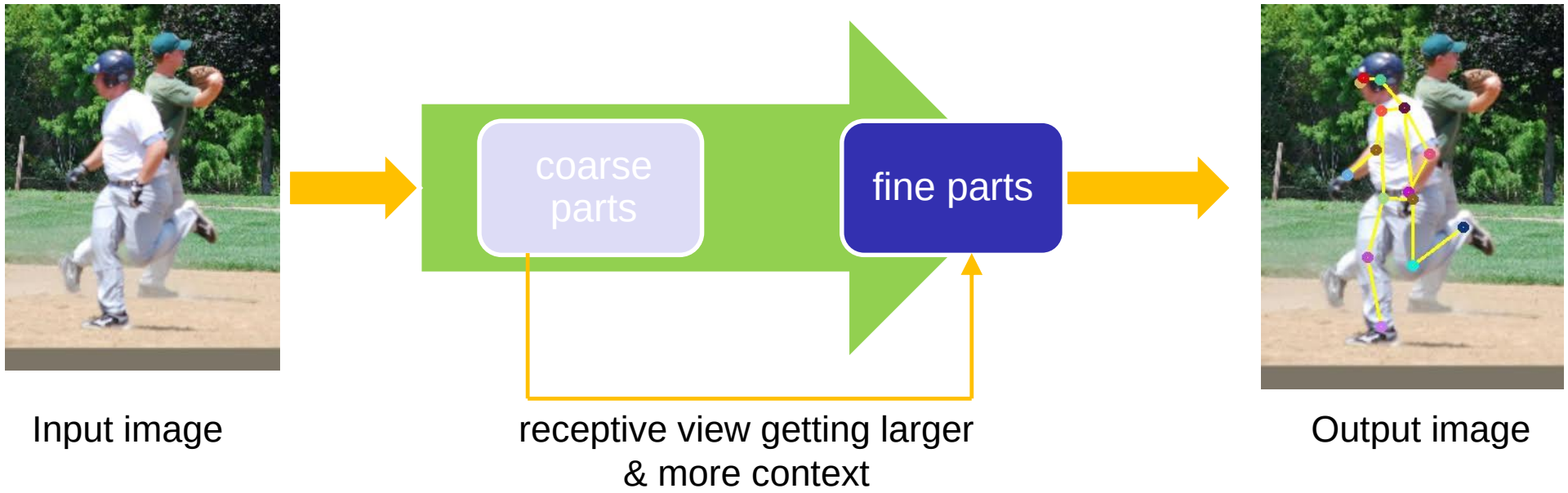
Top-Down: Cascade Pyramid Network

- Motivation: How to locate the “hard” joints
- Human perspective

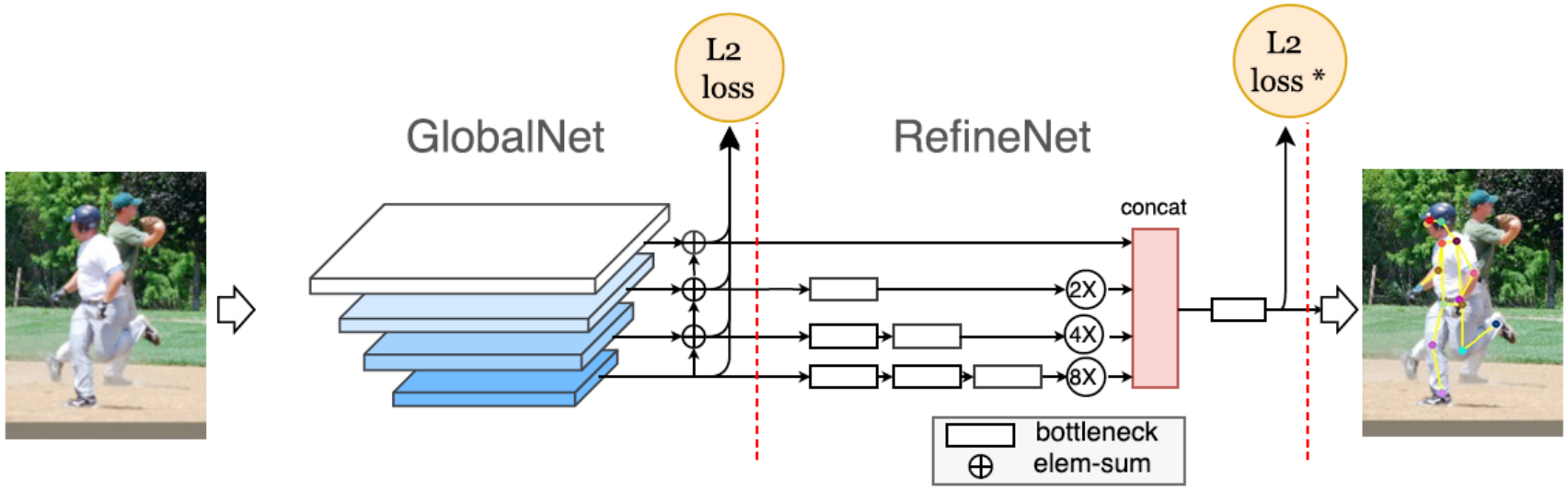


Top-Down: Cascade Pyramid Network

- Motivation: How to locate the “hard” joints
- Human perspective： Coarse to Fine



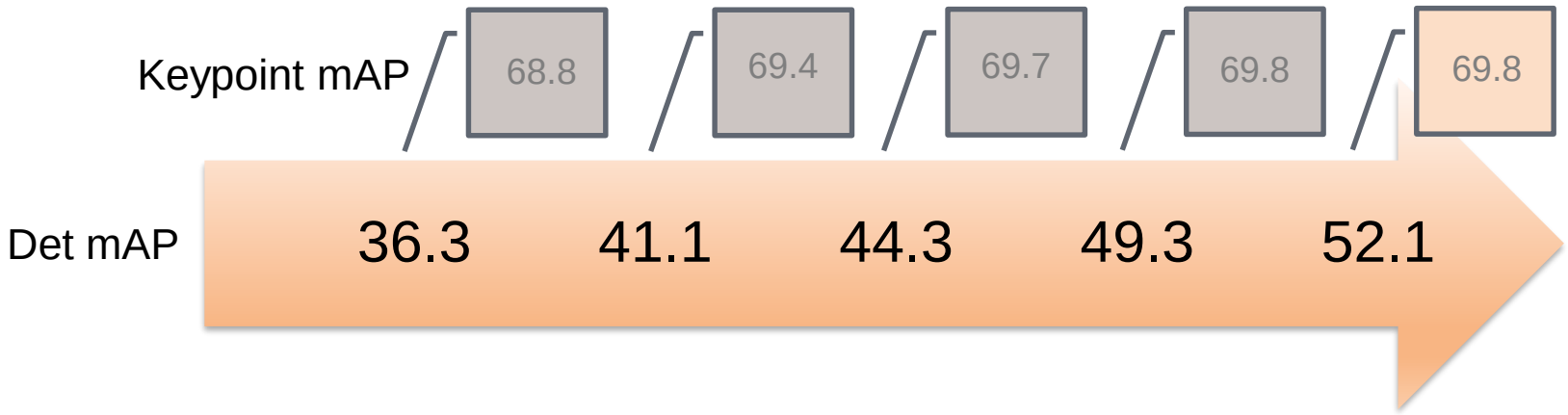
Network Architecture



Network Design Principles:

- Inspired by the process of human locating keypoints and adjusted to CNN network
  - locate easy parts => locate hard parts
- Two stages
  - GlobalNet: to locate the easy parts (Vanilla L2 loss)
  - RefineNet: to locate hard parts (deep layers) with online hard keypoint mining(Hard Mining Loss)

Experiments: Person Detector



Det Methods	AP(all)	AP(H)	AR(H)	AP(OKS)
FPN-1	36.3	49.6	58.5	68.8
FPN-2	41.1	55.3	67.0	69.4
FPN-3	44.3	58.4	71.3	69.7
ensemble-1	49.3	61.4	71.8	69.8
ensemble-2	52.1	62.9	74.7	69.8

Table 2. Comparison between detection performance and key-points detection performance. FPN-1: FPN with the backbone of res50; FPN-2: res101 with Soft-NMS and OHEM [38] applied; FPN-3: resnext101 with Soft-NMS, OHEM [38], multiscale training applied; ensemble-1: multiscale test involved; ensemble-2: multiscale test, large batch and SENet [18] involved. H is short for Human.

Experiments: Online Hard Keypoints Mining

M	6	8	10	12	14	17
AP (OKS)	68.8	69.4	69.0	69.0	69.0	68.6

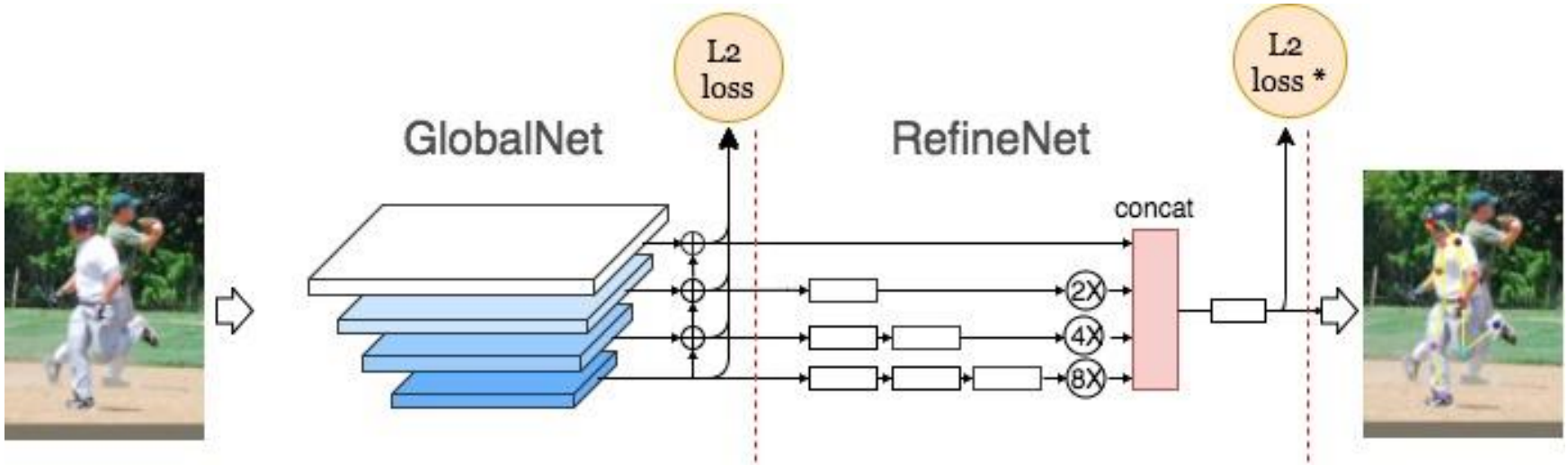
Table 4. Comparison of different hard keypoints number in online hard keypoints mining.

GlobalNet	RefineNet	AP(OKS)
-	L2 loss	68.2
L2 loss	L2 loss	68.6
-	L2 loss*	68.5
L2 loss	L2 loss*	69.4
L2 loss*	L2 loss*	69.1

Table 5. Comparison of models with different losses function. Here "-" denotes that the model applies no loss function in corresponding subnetwork. "L2 loss*" means L2 loss with online hard keypoints mining.

Experiments: Design Choices of GlobalNet & RefineNet

Models	AP(OKS)	FLOPs
GlobalNet only	66.6	3.90G
GlobalNet + Concat	68.5	5.87G
GlobalNet + one bottleneck +Concat	69.2	6.92G
ours (CPN)	69.4	6.20G



Connections	AP(OKS)	FLOPs
C_2	68.3	5.02G
$C_2 \sim C_3$	68.4	5.50G
$C_2 \sim C_4$	69.1	5.88G
$C_2 \sim C_5$	69.4	6.20G

Experiments

MEGVII 旷视

Methods	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR	$AR_{@.5}$	$AR_{@.75}$	AR_m	AR_l
FAIR Mask R-CNN*	68.9	89.2	75.2	63.7	76.8	75.4	93.2	81.2	70.2	82.6
G-RMI*	69.1	85.9	75.2	66.0	74.5	75.1	90.7	80.7	69.7	82.4
bangbangren+*	70.6	88.0	76.5	65.6	79.2	77.4	93.6	83.0	71.8	85.0
oks*	71.4	89.4	78.1	65.9	79.1	77.2	93.6	83.4	71.8	84.5
Ours+ (CPN+)	72.1	90.5	78.9	67.9	78.1	78.7	94.7	84.8	74.3	84.7

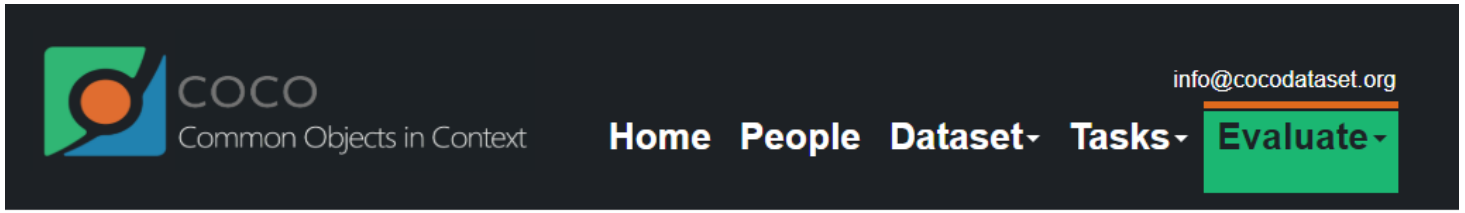
Table 9. Comparisons of final results on COCO test-challenge2017 dataset. "*" means that the method involves extra data for training. Specifically, FAIR Mask R-CNN involves distilling unlabeled data, oks uses AI-Challenger keypoints dataset, bangbangren and G-RMI use their internal data as extra data to enhance performance. "+" indicates results using ensembled models. The human detector of Ours+ is a detector that has an AP of 62.9 of human class on COCO minival dataset. CPN and CPN+ in this table all use the backbone of ResNet-Inception [39] framework.

Methods	AP	$AP_{@.5}$	$AP_{@.75}$	AP_m	AP_l	AR	$AR_{@.5}$	$AR_{@.75}$	AR_m	AR_l
CMU-Pose [6]	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
Mask-RCNN [16]	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
Associative Embedding [29]	65.5	86.8	72.3	60.6	72.6	70.2	89.5	76.0	64.6	78.1
G-RMI [31]	64.9	85.5	71.3	62.3	70.0	69.7	88.7	75.5	64.4	77.1
G-RMI* [31]	68.5	87.1	75.5	65.8	73.3	73.3	90.1	79.5	68.1	80.4
Ours (CPN)	72.1	91.4	80.0	68.7	77.2	78.5	95.1	85.3	74.2	84.3
Ours+ (CPN+)	73.0	91.7	80.9	69.5	78.1	79.0	95.1	85.9	74.8	84.7

Table 10. Comparisons of final results on COCO test-dev dataset. "*" means that the method involves extra data for training. "+" indicates results using ensembled models. The human detectors of Our and Ours+ the same detector that has an AP of 62.9 of human class on COCO minival dataset.CPN and CPN+ in this table all use the backbone of ResNet-Inception [39] framework.

Summary for CPN

- Hard Keypoints with Coarse-to-fine Strategy (context)
- Code: https://github.com/chenyilun95/tf-cpn
- MS COCO2017 Challenge Winner



Keypoint Leaderboard

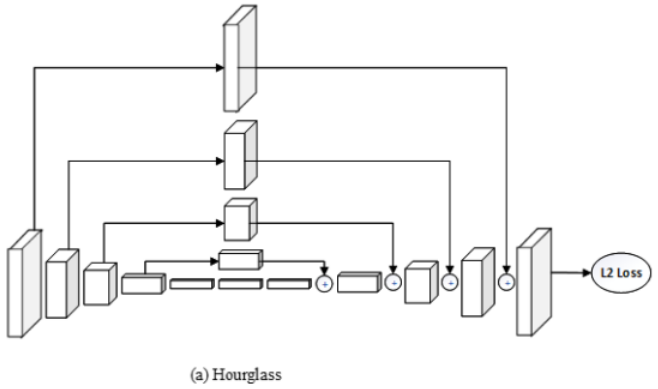
Dev | Challenge16 | Challenge17 | Challenge18

Copy to Clipboard | Export to CSV | Search:

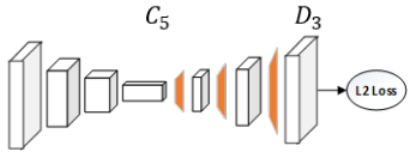
	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}	date
Megvii (Face++)	0.721	0.905	0.789	0.679	0.781	0.787	0.947	0.848	0.743	0.847	2017-10-29
oks	0.714	0.894	0.781	0.659	0.791	0.772	0.936	0.834	0.718	0.845	2017-10-29
bangbangren	0.706	0.880	0.765	0.656	0.792	0.774	0.936	0.830	0.718	0.850	2017-10-29
G-RMI	0.691	0.859	0.752	0.660	0.745	0.751	0.907	0.807	0.697	0.824	2017-10-29
FAIR Mask R-CNN	0.689	0.892	0.752	0.637	0.768	0.754	0.932	0.812	0.702	0.826	2017-10-29
SJTU	0.680	0.867	0.747	0.633	0.750	0.735	0.908	0.795	0.686	0.804	2017-10-29

Top-Down: A Simple Baseline

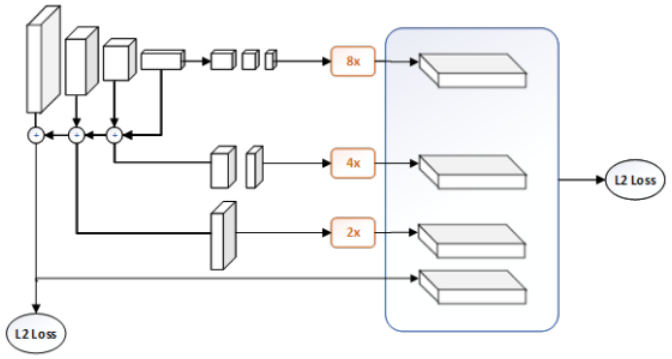
- Motivation
  - Simple Baseline & OKS based tracking
  - Spatial Resolution



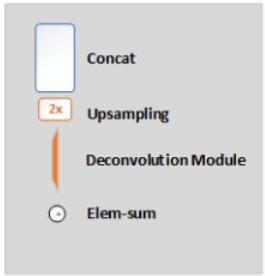
(a) Hourglass

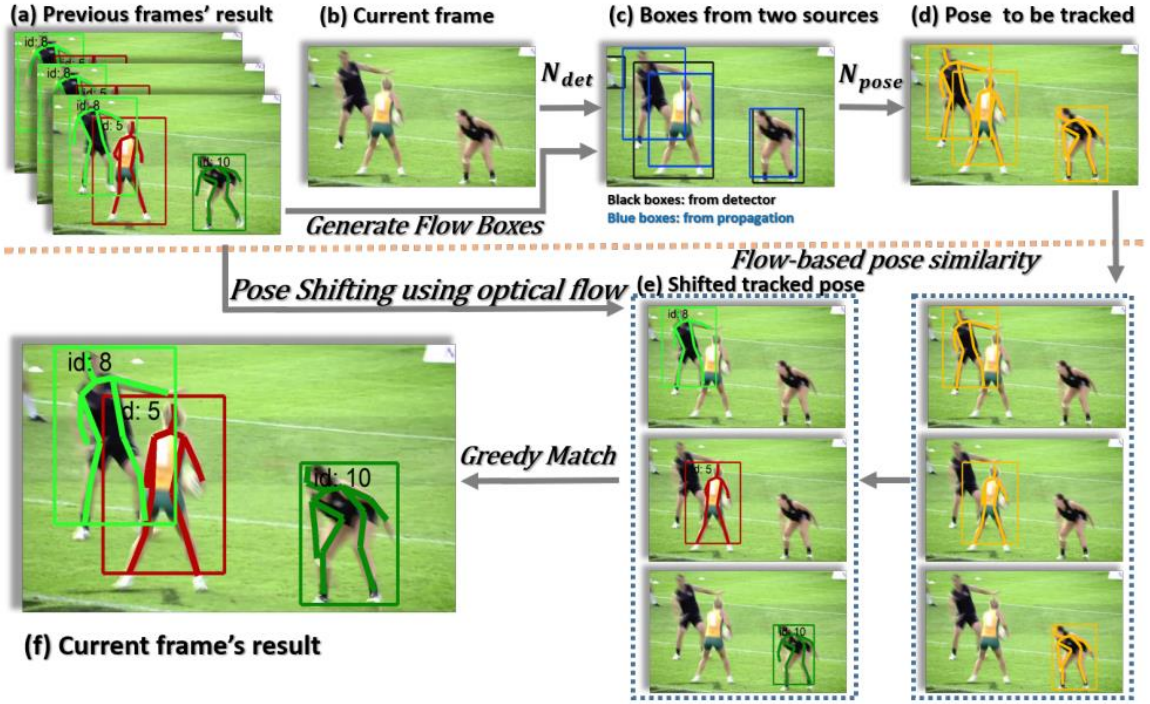


(c) Our Network



(b) CPN





Simple Baselines for Human Pose Estimation and Tracking, Bin Xiao, Haiping Wu, Yichen Wei, ECCV 2018

Top-Down: A Simple Baseline

- Experiments on COCO and PoseTrack

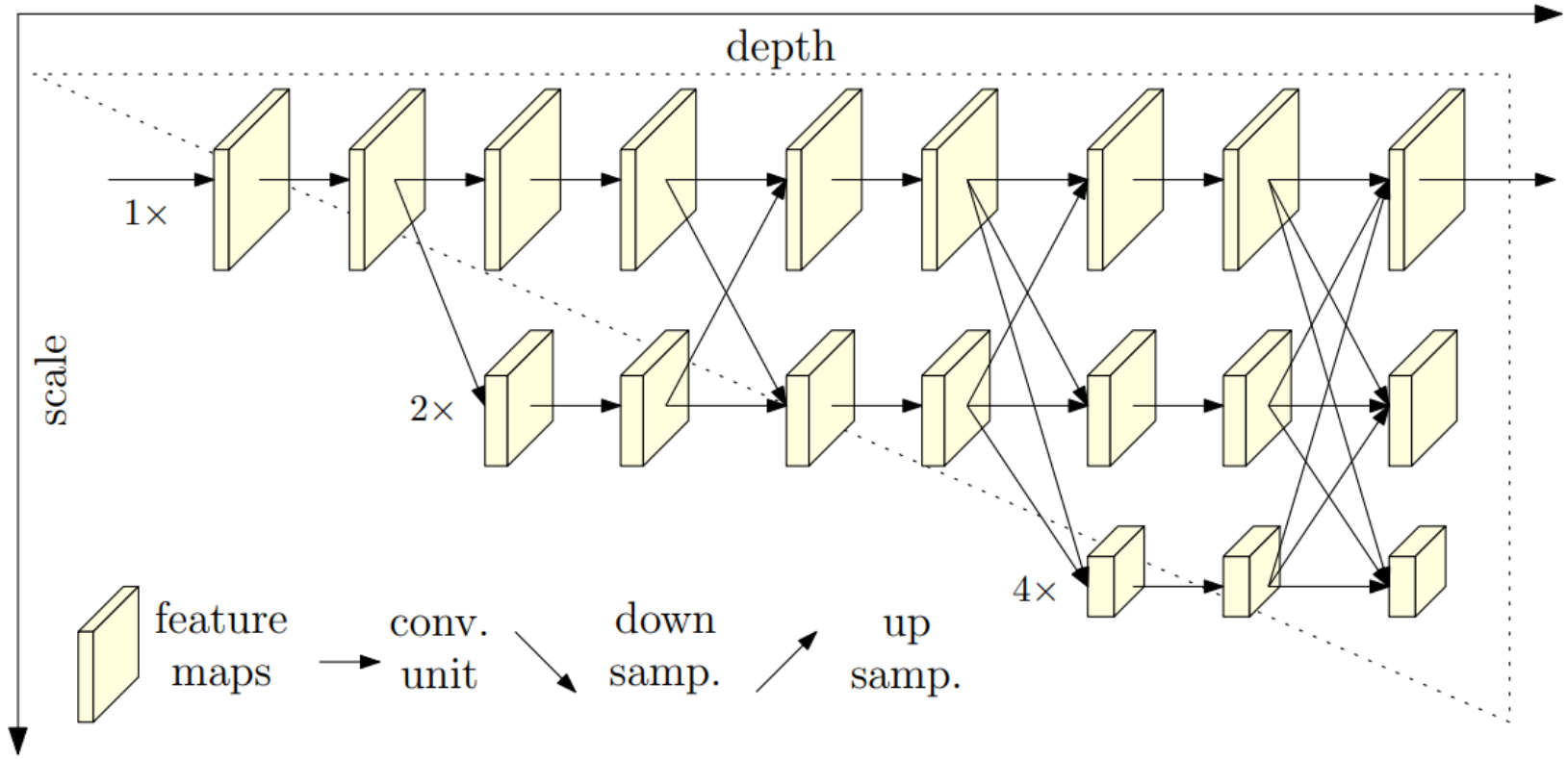
Method	Backbone	Input Size	OHKM	*AP*
8-stage Hourglass	-	256×192	✗	66.9
8-stage Hourglass	-	256×256	✗	67.1
CPN	ResNet-50	256×192	✗	68.6
CPN	ResNet-50	384×288	✗	70.6
CPN	ResNet-50	256×192	✓	69.4
CPN	ResNet-50	384×288	✓	71.6
Ours	ResNet-50	256×192	✗	70.4
Ours	ResNet-50	384×288	✗	72.2

Method	Backbone	Detector	With Joint Propagation	Similarity Metric	mAP Total	MOTA Total
a_1	ResNet-50	R-FCN	✗	S_{Bbox}	66.0	57.6
a_2	ResNet-50	R-FCN	✗	S_{Pose}	66.0	57.7
a_3	ResNet-50	R-FCN	✓	S_{Bbox}	70.3	61.4
a_4	ResNet-50	R-FCN	✓	S_{Pose}	70.3	61.8
a_5	ResNet-50	R-FCN	✓	S_{Flow}	70.3	61.8
a_6	ResNet-50	R-FCN	✓	$S_{Multi-Flow}$	**70.3**	**62.2**
b_1	ResNet-50	FPN-DCN	✗	S_{Bbox}	69.3	59.8
b_2	ResNet-50	FPN-DCN	✗	S_{Pose}	69.3	59.7
b_3	ResNet-50	FPN-DCN	✓	S_{Bbox}	72.4	62.1
b_4	ResNet-50	FPN-DCN	✓	S_{Pose}	72.4	61.8
b_5	ResNet-50	FPN-DCN	✓	S_{Flow}	72.4	62.4
b_6	ResNet-50	FPN-DCN	✓	$S_{Multi-Flow}$	**72.4**	**62.9**
c_1	ResNet-152	FPN-DCN	✗	S_{Bbox}	72.9	62.0
c_2	ResNet-152	FPN-DCN	✗	S_{Pose}	72.9	61.9
c_3	ResNet-152	FPN-DCN	✓	S_{Bbox}	76.7	64.8
c_4	ResNet-152	FPN-DCN	✓	S_{Pose}	76.7	64.9
c_5	ResNet-152	FPN-DCN	✓	S_{Flow}	76.7	65.1
c_6	ResNet-152	FPN-DCN	✓	$S_{Multi-Flow}$	**76.7**	**65.4**

Simple Baselines for Human Pose Estimation and Tracking, Bin Xiao, Haiping Wu, Yichen Wei, ECCV 2018

Top-Down: HRNet

- Motivation
  - High Resolution Feature maps



Deep High-Resolution Representation Learning for Human Pose Estimation, Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang, CVPR2019

Top-Down: HRNet

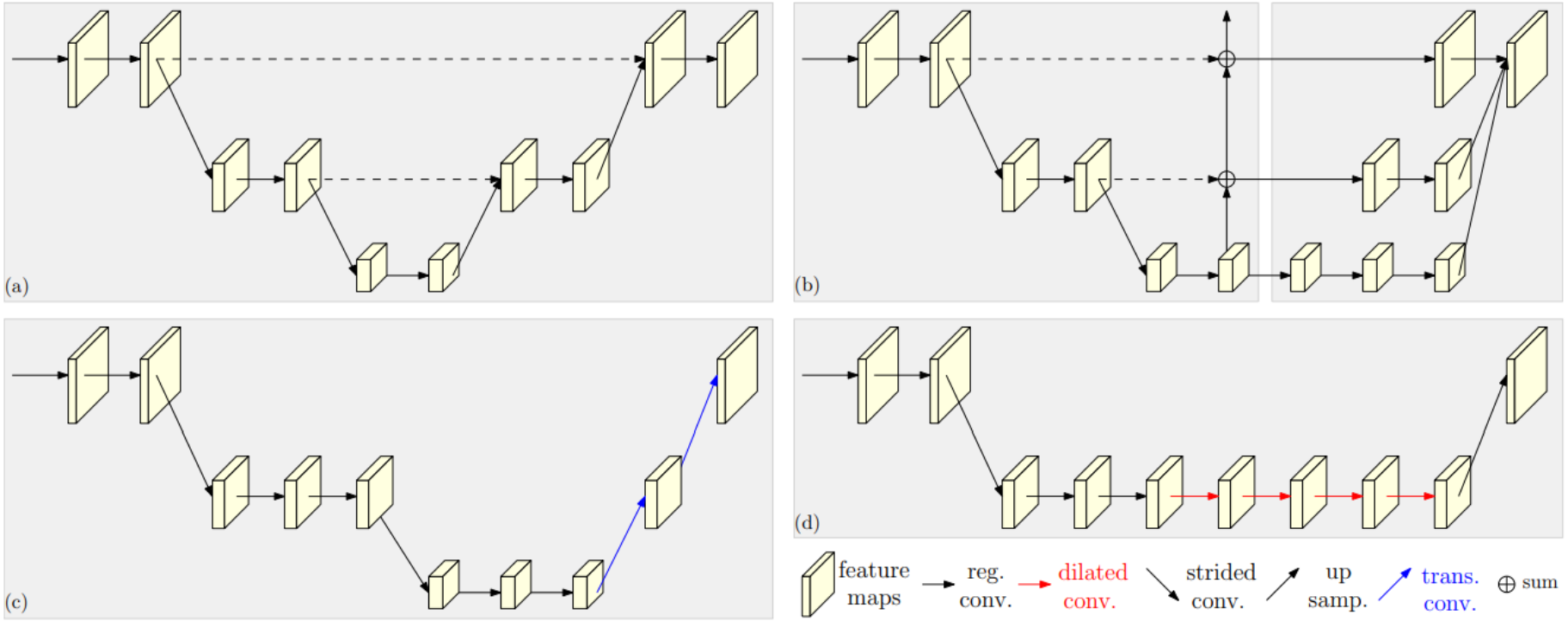


Figure 2. Illustration of representative pose estimation networks that rely on the high-to-low and low-to-high framework. (a) Hourglass [40]. (b) Cascaded pyramid networks [11]. (c) SimpleBaseline [72]: transposed convolutions for low-to-high processing. (d) Combination with dilated convolutions [27]. Bottom-right legend: reg. = regular convolution, dilated = dilated convolution, trans. = transposed convolution, strided = strided convolution, concat. = concatenation. In (a), the high-to-low and low-to-high processes are symmetric. In (b), (c) and (d), the high-to-low process, a part of a classification network (ResNet or VGGNet), is *heavy*, and the low-to-high process is *light*. In (a) and (b), the skip-connections (dashed lines) between the same-resolution layers of the high-to-low and low-to-high processes mainly aim to fuse low-level and high-level features. In (b), the right part, refinenet, combines the low-level and high-level features that are processed through convolutions.

Deep High-Resolution Representation Learning for Human Pose Estimation, Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang, CVPR2019

Top-Down: HRNet

- Experiments

Table 1. Comparisons on the COCO validation set. Pretrain = pretrain the backbone on the ImageNet classification task. OHKM = online hard keypoints mining [11].

Method	Backbone	Pretrain	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stage Hourglass [40]	8-stage Hourglass	N	256 × 192	25.1M	14.3	66.9	–	–	–	–	–
CPN [11]	ResNet-50	Y	256 × 192	27.0M	6.20	68.6	–	–	–	–	–
CPN + OHKM [11]	ResNet-50	Y	256 × 192	27.0M	6.20	69.4	–	–	–	–	–
SimpleBaseline [72]	ResNet-50	Y	256 × 192	34.0M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [72]	ResNet-101	Y	256 × 192	53.0M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [72]	ResNet-152	Y	256 × 192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32	HRNet-W32	N	256 × 192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32	HRNet-W32	Y	256 × 192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48	HRNet-W48	Y	256 × 192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
SimpleBaseline [72]	ResNet-152	Y	384 × 288	68.6M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W32	HRNet-W32	Y	384 × 288	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNet-W48	HRNet-W48	Y	384 × 288	63.6M	32.9	**76.3**	**90.8**	**82.9**	**72.3**	**83.4**	**81.2**

Deep High-Resolution Representation Learning for Human Pose Estimation, Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang, CVPR2019

Top-Down: Multi-stage Pose Estimation

- Motivation
  - Upperbound
  - Only Two-stages available (limited Context)

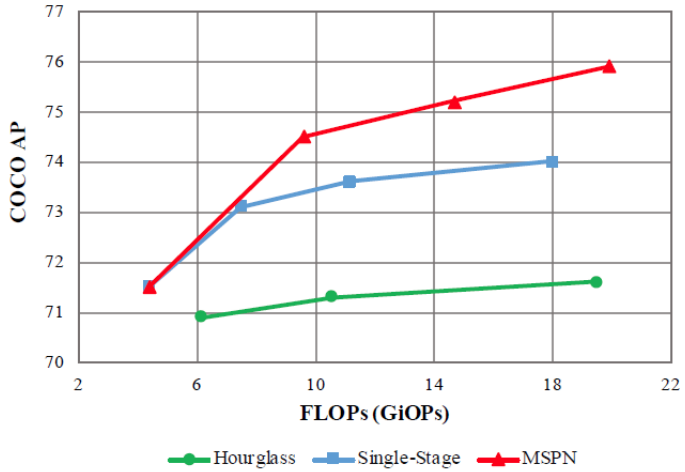


Figure 1. Pose estimation performance on COCO minival dataset of Hourglass [29], a single-stage model using ResNet [17], and our proposed MSPN under different model capacity (measured in FLOPs).

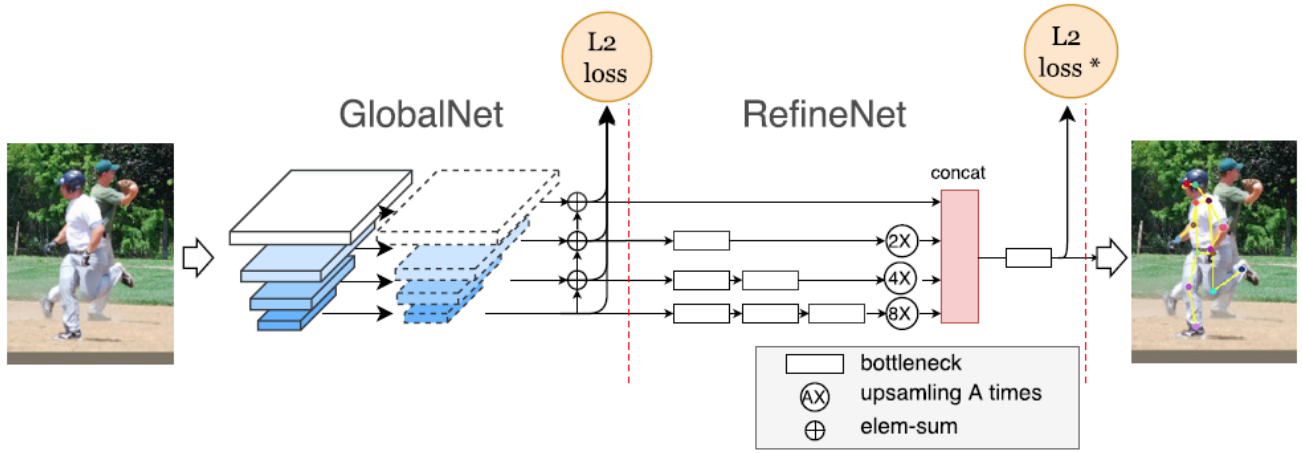
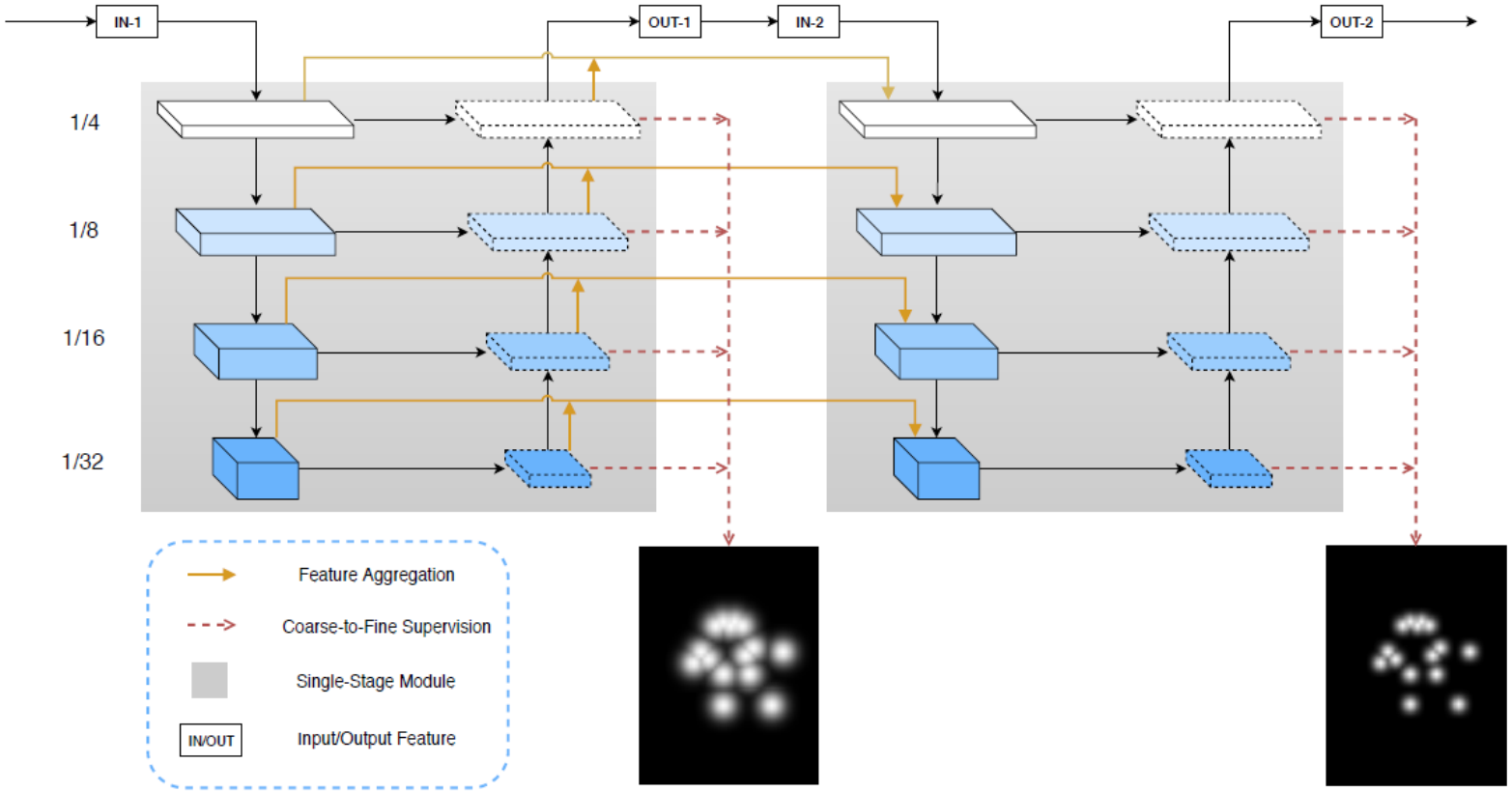


Figure 1. Cascaded Pyramid Network. "L2 loss*" means L2 loss with online hard keypoints mining.

Rethinking on Multi-Stage Networks for Human Pose Estimation, Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, Jian Sun

Top-Down: Multi-stage Pose Estimation

- Method
  - Coarse-to-fine with better information flow
  - Involve more stages



Top-Down: Multi-stage Pose Estimation

- Cross Stage Feature Aggregation
- Coarse-to-fine Supervision

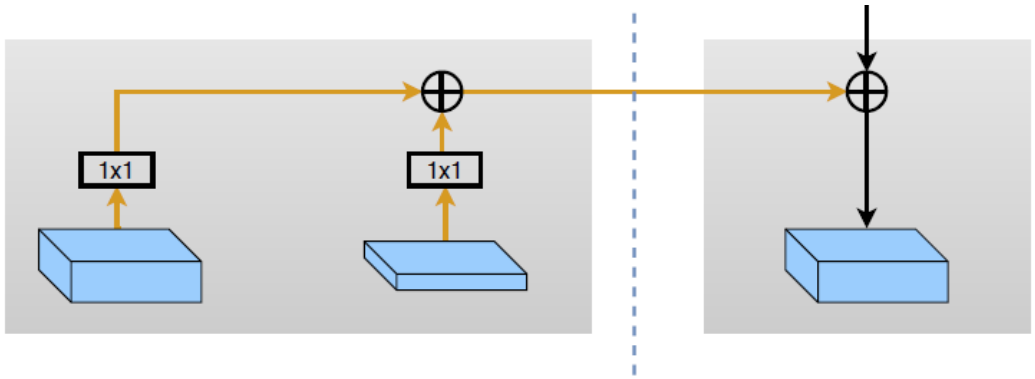


Figure 3. Cross Stage Feature Aggregation on a specific scale. Two 1×1 convolutional operations are applied to the features of previous stage before aggregation. See Figure 2 for the overall network structure.

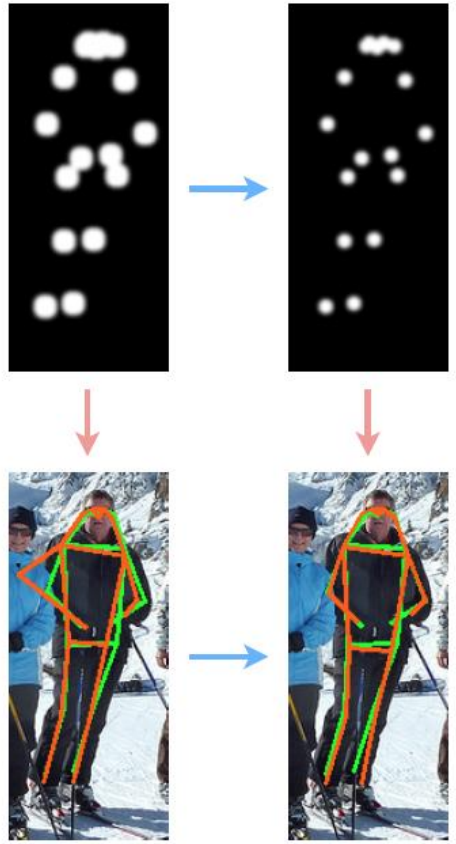


Figure 4. Illustration of coarse-to-fine supervision. The first row shows ground-truth heat maps in different stages and the second row represents corresponding predictions and ground truth annotations. The orange line is the prediction result and the green line indicates ground truth.

Experiments: More Stages

Stages	Hourglass FLOPs(G)	Hourglass AP	Stages	MSPN FLOPs(G)	MSPN AP
1	3.9	65.4	1	4.4	71.5
2	6.2	70.9	2	9.6	74.5
4	10.6	71.3	3	14.7	75.2
8	19.5	71.6	4	19.9	75.9
2†	15.4†	71.7†	-	-	-

Table 2. Results of Hourglass and MSPN with different number of stages on COCO minival dataset. "†" denotes the result of a variant Hourglass [28] as illustrated in Section 3.1. MSPN adopts Res-50 in each single-stage module.

Method	Res-50	2×Res-18	L-XCP	4× S-XCP
AP	71.5	71.6	73.7	74.7
FLOPs	4.4G	4.0G	6.1G	5.7G

Experiments: CTF & CSFA

Components			Hourglass	MSPN
BaseNet	CTF	CSFA		
√			71.3	73.3
√	√		72.5	74.2
√	√	√	73.0	74.5

Table 4. Ablation Study of MSPN on COCO minival dataset. 'BaseNet' represents a 4-stage Hourglass or 2-stage MSPN based on Res-50 with similar complexity, see Table 2. 'CTF' indicates the coarse-to-fine supervision. 'CSFA' means the cross stage feature aggregation.

Experiments: COCO test-dev

MEGVII 旷视

Method	Backbone	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CMU Pose [5]	-	-	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
Mask R-CNN [16]	Res-50-FPN	-	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
G-RMI [31]	Res-152	353×257	64.9	85.5	71.3	62.3	70.0	69.7	88.7	75.5	64.4	77.1
AE [28]	-	512×512	65.5	86.8	72.3	60.6	72.6	70.2	89.5	76.0	64.6	78.1
CPN [9]	Res-Inception	384×288	72.1	91.4	80.0	68.7	77.2	78.5	95.1	85.3	74.2	84.3
Simple Base [46]	Res-152	384×288	73.7	91.9	81.1	70.3	80.0	79.0	-	-	-	-
HRNet [39]	HRNet-W48	384×288	75.5	92.5	83.3	71.9	81.5	80.5	-	-	-	-
Ours (MSPN)	4×Res-50	384×288	**76.1**	**93.4**	**83.8**	**72.3**	**81.5**	**81.6**	**96.3**	**88.1**	**77.5**	**87.1**
CPN+ [9]	Res-Inception	384×288	73.0	91.7	80.9	69.5	78.1	79.0	95.1	85.9	74.8	84.6
Simple Base+* [46]	Res-152	384×288	76.5	92.4	84.0	73.0	82.7	81.5	95.8	88.2	77.4	87.2
HRNet* [39]	HRNet-W48	384×288	77.0	92.7	84.5	73.4	83.1	82.0	-	-	-	-
Ours (MSPN*)	4×Res-50	384×288	77.1	93.8	84.6	73.4	82.3	82.3	96.5	88.9	78.4	87.7
Ours (MSPN+*)	4×Res-50	384×288	**78.1**	**94.1**	**85.9**	**74.5**	**83.3**	**83.1**	**96.7**	**89.8**	**79.3**	**88.2**

Table 7. Comparisons of results on COCO test-dev dataset. ”+” indicates using an ensemble model and ”*” means using external data.

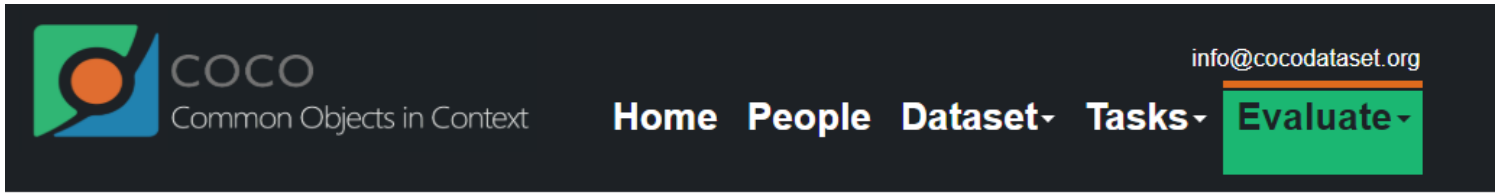
Experiments: COCO test-Challenge

Method	Backbone	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Mask R-CNN* [16]	ResX-101-FPN	-	68.9	89.2	75.2	63.7	76.8	75.4	93.2	81.2	70.2	82.6
G-RMI* [31]	Res-152	353×257	69.1	85.9	75.2	66.0	74.5	75.1	90.7	80.7	69.7	82.4
CPN+ [9]	Res-Inception	384×288	72.1	90.5	78.9	67.9	78.1	78.7	94.7	84.8	74.3	84.7
Sea Monsters+*	-	-	74.1	90.6	80.4	68.5	82.1	79.5	94.4	85.1	74.1	86.8
Simple Base+* [46]	Res-152	384×288	74.5	90.9	80.8	69.5	82.9	80.5	95.1	86.3	75.3	87.5
Ours (MSPN+*)	4×Res-50	384×288	**76.4**	**92.9**	**82.6**	**71.4**	**83.2**	**82.2**	**96.0**	**87.7**	**77.5**	**88.6**

Table 8. Comparisons of results on COCO test-challenge dataset. ”+” means using an ensemble model and ”*” means using external data.

Summary for MSPN

- Refined Coarse-to-fine Strategy
- Code: https://github.com/megvii-detection/MSPN
- MS COCO2018 Challenge Winner



Keypoint Leaderboard

Dev | Challenge16 | Challenge17 | Challenge18

Copy to Clipboard | Export to CSV | Search:

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}	date
Megvii (Face++)	0.764	0.929	0.826	0.714	0.832	0.822	0.960	0.877	0.775	0.886	2018-09-09
MSRA	0.745	0.909	0.808	0.695	0.829	0.805	0.951	0.863	0.753	0.875	2018-09-09
The Sea Monsters	0.741	0.906	0.804	0.685	0.821	0.795	0.944	0.851	0.741	0.868	2018-09-09
DGDBQ	0.738	0.900	0.798	0.687	0.806	0.795	0.944	0.850	0.743	0.866	2018-09-09
KPLab	0.728	0.904	0.794	0.685	0.800	0.796	0.948	0.855	0.747	0.863	2018-09-09
ByteDance-SEU	0.728	0.906	0.794	0.685	0.800	0.796	0.947	0.854	0.747	0.862	2018-09-09

Bottom-Up: DeepCut

- Motivation
  - Part Detector
  - Assemble (Integer Linear Optimization)

(a) (b) (c)

Figure 1. Method overview: (a) initial detections (= part candidates) and pairwise terms (graph) between all detections that (b) are jointly clustered belonging to one person (one colored subgraph = one person) and each part is labeled corresponding to its part class (different colors and symbols correspond to different body parts); (c) shows the predicted pose sticks.

DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, Bernt Schiele, CVPR 2016

Bottom-Up: DeeperCut

- Motivation
  - Deeper Part Detector + Assemble (image-conditioned pairwise terms + incremental optimization)

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP	time [s/frame]
1-stage optimize, 100 det, nms 1x	70.3	61.6	52.1	43.7	50.6	47.0	40.6	52.6	578
1-stage optimize, 100 det, nms 2x	71.3	64.1	55.8	44.1	53.8	48.7	41.3	54.5	596
1-stage optimize, 150 det, nms 2x	74.1	65.6	56.0	44.3	54.4	49.2	39.8	55.1	1041
2-stage optimize	75.9	66.8	58.8	46.1	54.1	48.7	42.4	56.5	483
3-stage optimize	78.3	69.3	58.4	47.5	55.1	49.6	42.5	57.6	271
+ split detections	**78.5**	**70.5**	**59.7**	**48.7**	**55.4**	**50.6**	**44.4**	**58.7**	**270**
DeepCut [10]	50.1	44.1	33.5	26.5	33.0	28.5	14.4	33.3	259220

Table 6. Performance (AP) of different hierarchical versions of *DeeperCut* on MPII Multi-Person Val.

DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model, Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, Bernt Schiele, ECCV2016

Bottom-Up: OpenPose

- Motivation
  - Part Detector (CPM) + Assemble (PAF)

Figure 2. Overall pipeline. Our method takes the entire image 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). We finally assemble them into full body poses for all people in the image (e).

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, CVPR 2017

Bottom-Up: OpenPose

- Motivation
  - Part Detector (CPM) + Assemble (PAF)

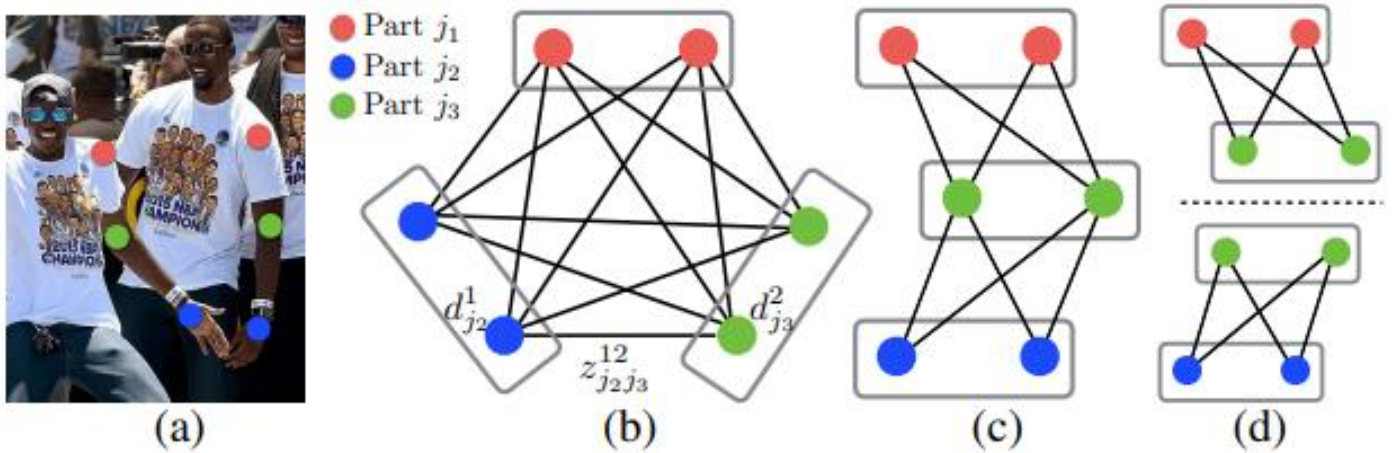


Figure 6. Graph matching. (a) Original image with part detections (b) K-partite graph (c) Tree structure (d) A set of bipartite graphs

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, CVPR 2017

Bottom-Up: OpenPose

- Experiments on MPI and COCO

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	**mAP**	s/image
Subset of 288 images as in [22]									
Deepcut [22]	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
Iqbal et al. [12]	70.0	65.2	56.4	46.1	52.7	47.9	44.5	54.7	10
DeeperCut [11]	87.9	84.0	71.9	63.9	68.8	63.8	58.1	71.2	230
Ours	**93.7**	**91.4**	**81.4**	**72.5**	**77.7**	**73.0**	**68.1**	**79.7**	**0.005**
Full testing set									
DeeperCut [11]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Iqbal et al. [12]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Ours (one scale)	89.0	84.9	74.9	64.2	71.0	65.6	58.1	72.5	0.005
Ours	**91.2**	**87.6**	**77.7**	**66.8**	**75.4**	**68.9**	**61.7**	**75.6**	**0.005**

Table 1. Results on the MPII dataset. Top: Comparison result on the testing subset. Middle: Comparison results on the whole testing set. Testing without scale search is denoted as "(one scale)".

Method	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
GT Bbox + CPM [11]	62.7	86.0	69.3	58.5	70.6
SSD [16] + CPM [11]	52.7	71.1	57.2	47.0	64.2
Ours - 6 stages	58.4	81.5	62.6	54.4	65.1
+ CPM refinement	61.0	84.9	67.5	56.3	69.3

Table 4. Self-comparison experiments on the COCO validation set.

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, CVPR 2017

Bottom-Up: Associative Embedding

- Motivation
  - Part Detector (Hourglass) + Assemble (AE)

Figure 1. Both multi-person pose estimation and instance segmentation are examples of computer vision tasks that require detection of visual elements (joints of the body or pixels belonging to a semantic class) and grouping of these elements (as poses or individual object instances).

Associative Embedding: End-to-End Learning for Joint Detection and Grouping, Alejandro Newell, Zhiao Huang, Jia Deng, NIPS 2017

Bottom-Up: Associative Embedding

- Motivation
  - Part Detector (Hourglass) + Assemble (AE)

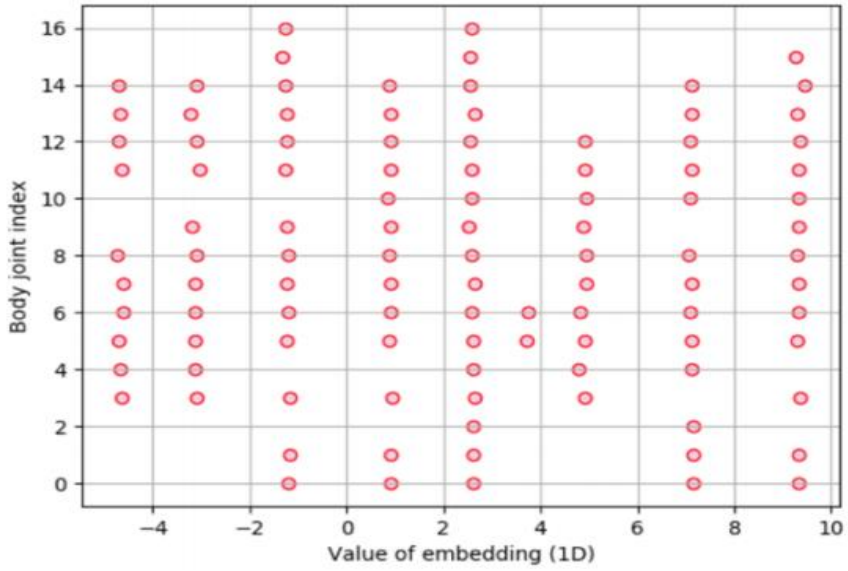


Figure 4. Tags produced by our network on a held-out validation image from the MS-COCO training set. The tag values are already well separated and decoding the groups is straightforward.

Associative Embedding: End-to-End Learning for Joint Detection and Grouping, Alejandro Newell, Zhiao Huang, Jia Deng, NIPS 2017

Bottom-Up: Associative Embedding

MEGVII 旷视

- Experiments on MPI and COCO

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Iqbal&Gall, ECCV16 [29]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
Insafutdinov et al., ECCV16 [28]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Insafutdinov et al., arXiv16a [45]	89.4	84.5	70.4	59.3	68.9	62.7	54.6	70.0
Levinkov et al., CVPR17 [33]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Insafutdinov et al., CVPR17 [27]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao et al., CVPR17 [6]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Fang et al., arXiv17 [15]	88.4	86.5	78.6	**70.4**	74.4	**73.0**	**65.8**	76.7
Our method	**92.1**	**89.3**	**78.9**	69.8	**76.2**	71.6	64.7	**77.5**

Table 1. Results (AP) on MPII Multi-Person.

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CMU-Pose [6]	0.618	0.849	0.675	0.571	0.682	0.665	0.872	0.718	0.606	0.746
Mask-RCNN [25]	0.627	**0.870**	0.684	0.574	0.711	–	–	–	–	–
G-RMI [42]	0.649	0.855	0.713	**0.623**	0.700	0.697	0.887	0.755	0.644	0.771
Our method	**0.655**	0.868	**0.723**	0.606	**0.726**	**0.702**	**0.895**	**0.760**	**0.646**	**0.781**

Table 3. Results on MS-COCO **test-dev**, excluding systems trained with external data.

Associative Embedding: End-to-End Learning for Joint Detection and Grouping, Alejandro Newell, Zhiao Huang, Jia Deng, NIPS 2017

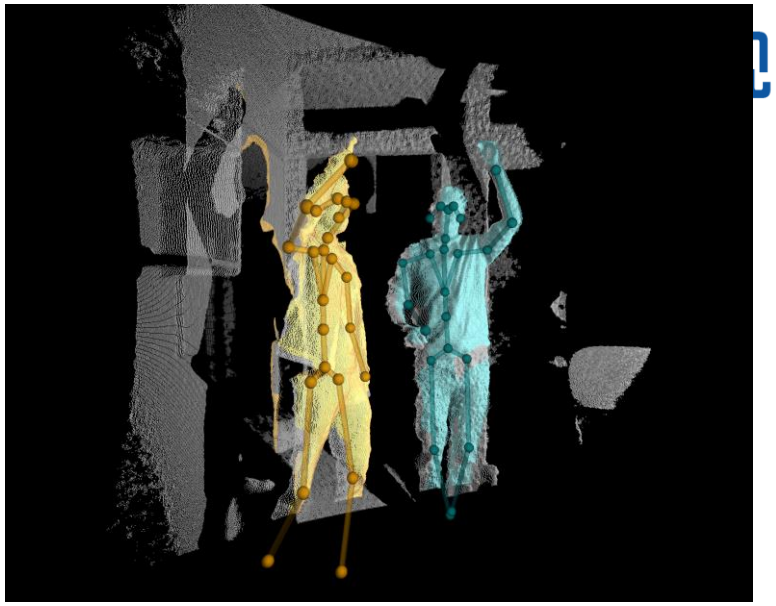
Bottom-Up: Azure Kinect

Azure Kinect DK

Build computer vision and speech models using a developer kit with advanced AI sensors

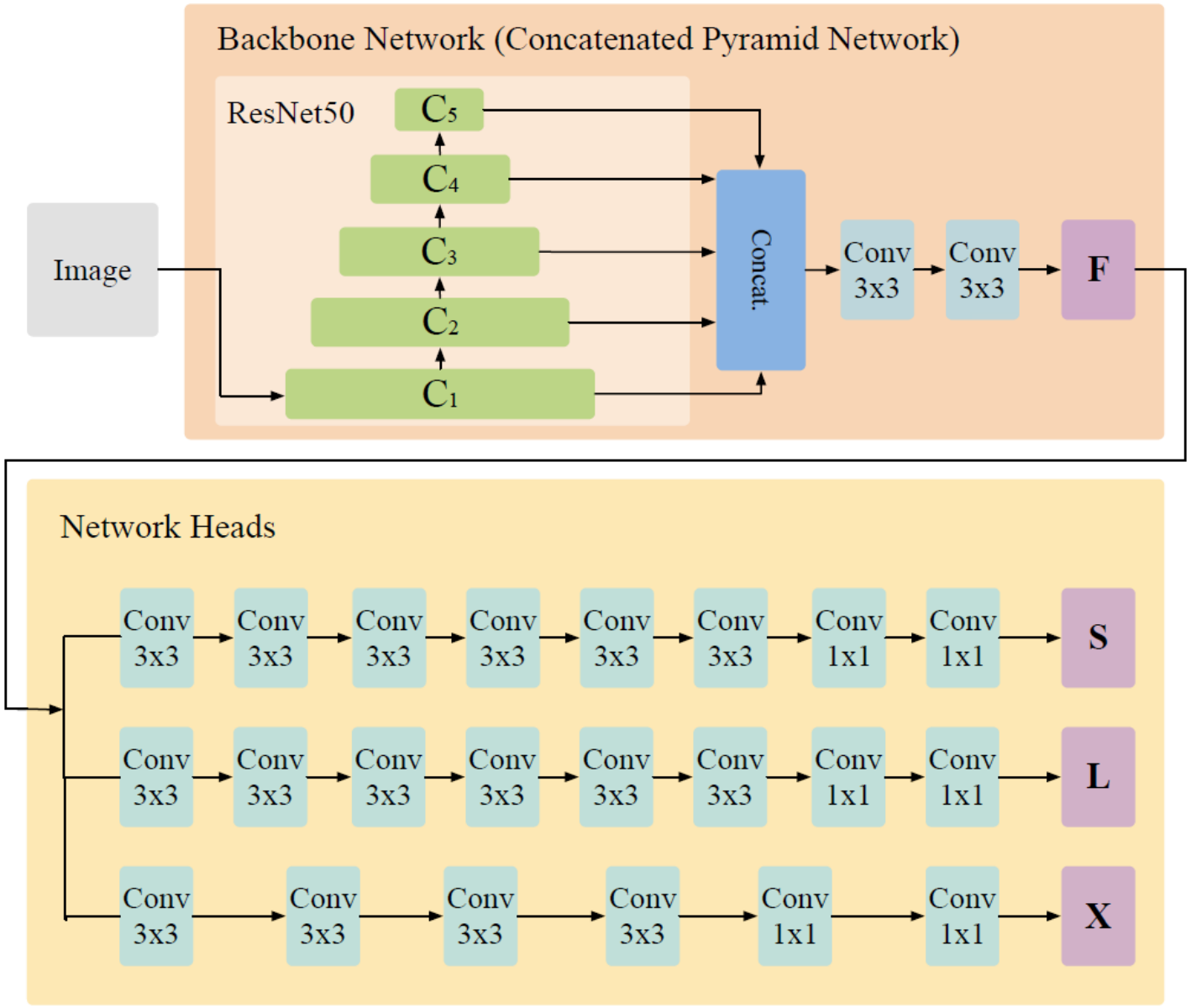
- Get started with a range of SDKs, including an open-source Sensor SDK.
- Experiment with multiple modes and mounting options.
- Add cognitive services and manage connected PCs with easy Azure integration.

Azure Kinect Body Tracking SDK



- Bottom up approach
  - On IR image
    - Insensitive to environment lighting
- DNN outputs
  - Heat map
  - Part Affinity Field
  - Part Segmentation Map
- SDK outputs
  - 3D skeletons
  - Instance segmentation

Neural Network



Microsoft

Summary for 2D Skeleton

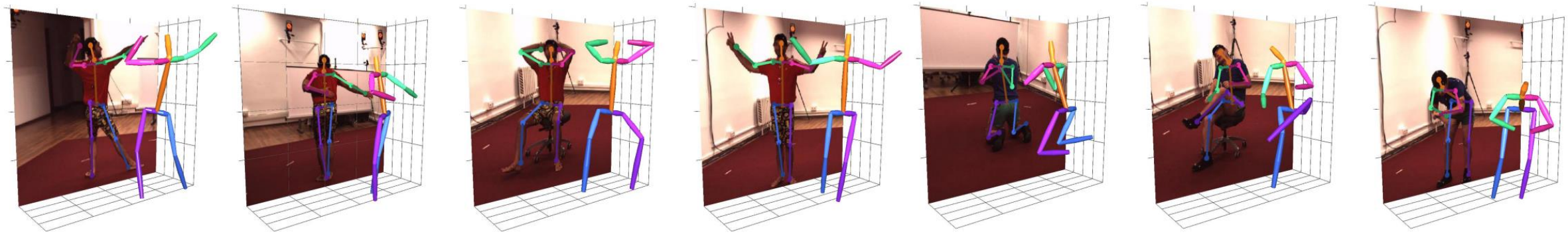
- Top-down vs Bottom-up
- Top-down: Context & spatial resolution
- Bottom-up: Assemble
- Remaining issues
  - Crowd
  - Spatial resolution
  - Speed

Outline

- Introduction to Human Pose Estimation
- 2D Skeleton
  - Top-Down
  - Bottom-Up
- 3D Skeleton
  - 2D -> 3D Skeleton
  - 2D -> 3D shape
- Application
- Conclusion

Benchmark: H3.6M

- Large-scale Constrained 3D Skeleton benchmark
- 3.6M human pose
- Evaluations
  - Protocol 1: Six subjects (S1, S5, S6, S7, S8, S9) are used in training. Evaluation is performed on every 64th frame of Subject 11's videos. Alignment is used.
  - Protocol 2: Five subjects (S1, S5, S6, S7, S8) are used for training. Evaluation is performed on every 64th frame of two subjects (S9, S11)

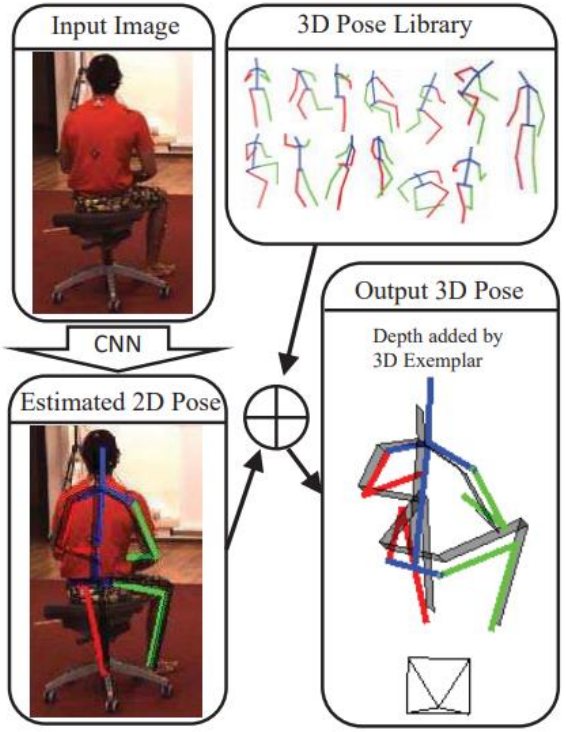


http://vision.imar.ro/human3.6m/description.php

3D Skeleton: 3D Human Pose Estimation = 2D Pose Estimation + Matching

- Motivation
  - 3D = 2D CNN + NN Match

- Split or Joint Training
- 3D structure: 2D Joints



$$p(\mathbf{X}, \mathbf{x}, I) = \underbrace{p(\mathbf{X}|\mathbf{x})}_{\text{NN}} \cdot \underbrace{p(\mathbf{x}|I)}_{\text{CNN}} \cdot p(I)$$

$$P(\mathbf{x}|I) = CNN(I)$$

$$P(\mathbf{X} = \mathbf{X}_i|\mathbf{x}) \propto e^{-\frac{1}{\sigma^2}||M_i(\mathbf{X}_i) - \mathbf{x}||^2}$$

https://zpascal.net/cvpr2017/Chen_3D_Human_Pose_CVPR_2017_paper.pdf

3D Human Pose Estimation = 2D Pose Estimation + Matching, Ching-Hang Chen Deva Ramanan, CVPR2017

3D Skeleton: 3D Human Pose Estimation = 2D Pose Estimation + Matching

- Experiments

Mean Per Joint Position Error (MPJPE), in mm

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Yasin [35]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0	**170.8**
Rogez [25]	-	-	-	-	-	-	-	-	-
Ours	**71.63**	**66.60**	**74.74**	**79.09**	**70.05**	**67.56**	**89.30**	**90.74**	195.62
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.	Median	-
Yasin [35]	108.2	142.5	86.9	92.1	165.7	102.0	108.3	-	-
Rogez [25]	-	-	-	-	-	-	88.1	-	-
Ours	**83.46**	**93.26**	**71.15**	**55.74**	**85.86**	**62.51**	**82.72**	**69.05**	-

Table 1. Comparison to [35] by **Protocol 1**. Our results are clearly state-of-the-art. Please see text for more details.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Yasin [35]	60.0	54.7	71.6	67.5	63.8	61.9	55.7	73.9	110.8
$\mathbf{X}^*\|gt$ (Ours)	**53.27**	**46.75**	**58.63**	**61.21**	**55.98**	**58.13**	**48.85**	**55.60**	**73.41**
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.	Median	-
Yasin [35]	78.9	96.9	67.9	47.5	89.3	53.4	70.5	-	-
Ours	**60.25**	**76.05**	**62.19**	**35.76**	**61.93**	**51.08**	**57.50**	**51.93**	-

Table 2. Comparison to [35] by **Protocol 1** given 2D ground truth. Our approach is clearly state-of-the-art, indicating the effectiveness of our simple approach to NN matching and warping. Table 7 shows that even simple NN matching produces an average accuracy of 70.93, rivaling prior art.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit	SitDown
Zhou [37]	**87.36**	109.31	**87.05**	**103.16**	116.18	106.88	**99.78**	**124.52**	**199.23**
Tekin [30]	102.41	147.72	88.83	125.38	118.02	112.38	129.17	138.89	224.9
Ours	89.87	**97.57**	89.98	107.87	**107.31**	**93.56**	136.09	133.14	240.12
Method	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg.	Median	-
Zhou [37]	107.42	139.46	118.09	79.39	114.23	97.70	**113.01**	-	-
Tekin [30]	118.42	182.73	138.75	**55.07**	126.29	**65.76**	124.97	-	-
Ours	**106.65**	**139.17**	**106.21**	87.03	**114.05**	90.55	114.18	93.05	-

Table 4. Comparison to [37] and [30] by **Protocol 2**. Our results are close to state-of-the-art.

https://zpascal.net/cvpr2017/Chen_3D_Human_Pose_CVPR_2017_paper.pdf

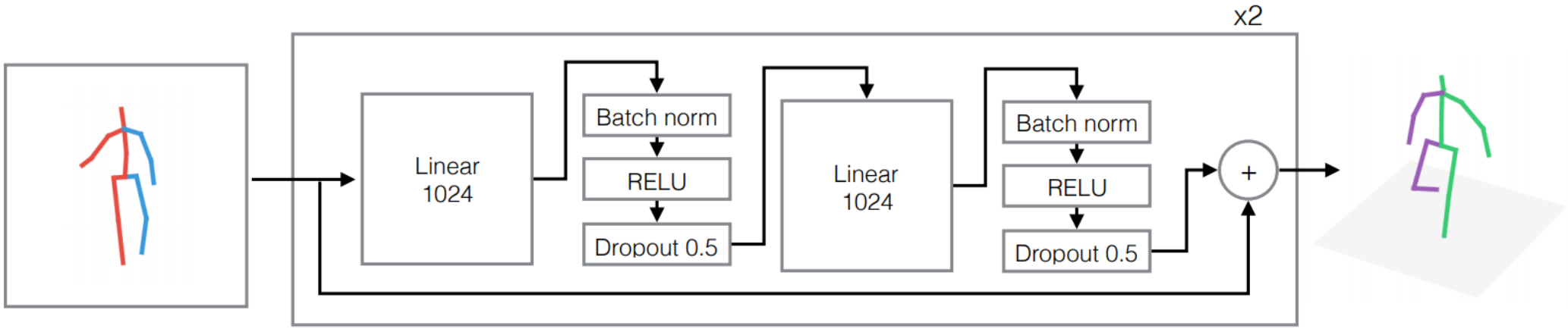
3D Human Pose Estimation = 2D Pose Estimation + Matching, Ching-Hang Chen Deva Ramanan, CVPR2017

3D Skeleton: A simple yet effective baseline for 3d human pose estimation

- Motivation
  - 3D = 2D CNN + Mapping

- Split or Joint Training
- 3D structure: 2D Joints

$$f^* = \min_f \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f(\mathbf{x}_i) - \mathbf{y}_i\right).$$



http://openaccess.thecvf.com/content_ICCV_2017/papers/Martinez_A_Simple_yet_ICCV_2017_paper.pdf

A simple yet effective baseline for 3d human pose estimation, Deva Ramanan, Julieta Martinez, Rayat Hossain, Javier Romero, James J. Little, ICCV2018

3D Skeleton: A simple yet effective baseline for 3d human pose estimation

- Experiments

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
LinKDE [19] (SA)	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Li *et al.* [24] (MA)	–	136.9	96.9	124.7	–	168.7	–	–	–	–	–	–	132.2	70.0	–	–
Tekin *et al.* [46] (SA)	102.4	147.2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Zhou *et al.* [56] (MA)	87.4	109.3	87.1	103.2	116.2	143.3	106.9	99.8	124.5	199.2	107.4	118.1	114.2	79.4	97.7	113.0
Tekin *et al.* [45] (SA)	–	129.1	91.4	121.7	–	162.2	–	–	–	–	–	–	130.5	65.8	–	–
Ghezelghieh *et al.* [13] (SA)	80.3	80.4	78.1	89.7	–	–	–	–	–	–	–	–	–	95.1	82.2	–
Du *et al.* [11] (SA)	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Park *et al.* [32] (SA)	100.3	116.2	90.0	116.5	115.3	149.5	117.6	106.9	137.2	190.8	105.8	125.1	131.9	62.6	96.2	117.3
Zhou *et al.* [54] (MA)	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132.2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Pavlakos *et al.* [33] (MA)	67.4	71.9	66.7	69.1	72.0	**77.0**	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Ours (SH detections) (SA)	61.6	73.4	63.3	58.3	91.8	93.6	66.3	62.0	91.7	109.4	75.7	86.5	67.2	51.2	52.3	73.6
Ours (SH detections) (MA)	53.3	60.8	62.9	62.7	86.4	82.4	57.8	58.7	81.9	99.8	69.1	63.9	67.1	50.9	54.8	67.5
Ours (SH detections FT) (MA)	**51.8**	**56.2**	**58.1**	**59.0**	**69.5**	78.4	**55.2**	**58.1**	**74.0**	**94.6**	**62.3**	**59.1**	**65.1**	**49.5**	**52.4**	**62.9**
Ours (GT detections) (MA)	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5

Protocol #2	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Akhter & Black [2]* (MA) 14j	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
Ramakrishna *et al.* [36]* (MA) 14j	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Zhou *et al.* [55]* (MA) 14j	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
Bogo *et al.* [7] (MA) 14j	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
Moreno-Noguer [27] (MA) 14j	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Pavlakos *et al.* [33] (MA) 17j	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.9
Ours (SH detections) (SA) 17j	50.1	59.5	51.3	56.9	68.5	67.5	51.0	47.2	68.5	85.6	61.2	67.0	55.1	41.1	45.5	58.5
Ours (SH detections) (MA) 17j	42.2	48.0	49.8	50.8	61.7	60.7	44.2	43.6	64.3	76.5	55.8	49.1	53.6	40.8	46.4	52.5
Ours (SH detections FT) (MA) 17j	**39.5**	**43.2**	**46.4**	**47.0**	**51.0**	**56.0**	**41.4**	**40.6**	**56.5**	**69.4**	**49.2**	**45.0**	**49.5**	**38.0**	**43.1**	**47.7**
Ours (SH detections) (SA) 14j	44.8	52.0	44.4	50.5	61.7	59.4	45.1	41.9	66.3	77.6	54.0	58.8	49.0	35.9	40.7	52.1

ed and
I3.6M.
dicates

Table 3. Detailed results on Human3.6M [19] under protocol #2 (rigid alignment in post-processing). The 14j (17j) annotation indicates that the body model considers 14 (17) body joints. The results of all approaches are obtained from the original papers, except for (*), which were obtained from [7].

http://openaccess.thecvf.com/content_ICCV_2017/papers/Martinez_A_Simple_yet_ICCV_2017_paper.pdf

A simple yet effective baseline for 3d human pose estimation, Deva Ramanan, Julieta Martinez, Rayat Hossain, Javier Romero, James J. Little, ICCV2018

3D Skeleton: Compositional Human Pose Regression

- Motivation
  - Bone Representation + 2D & 3D Joint training

- Split or Joint Training
- 3D structure: 2D Joints + bone

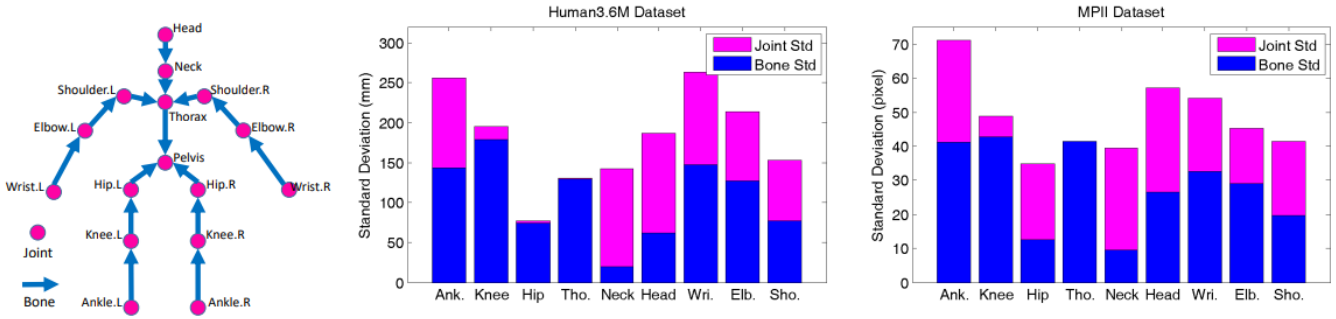


Figure 1. Left: a human pose is represented as either joints $\mathcal{J}$ or bones $\mathcal{B}$. Middle/Right: standard deviations of bones and joints for the 3D Human3.6M dataset [20] and 2D MPII dataset [3].

$$L(\mathcal{B},\mathcal{P}) = L_{xy}(\mathcal{B},\mathcal{P}) + L_z(\mathcal{B},\mathcal{P}).$$

$$L(\mathcal{B},\mathcal{P}) = \sum_{(u,v)\in\mathcal{P}} ||\tilde{\Delta\mathbf{J}}_{u,v} - \tilde{\Delta\mathbf{J}}^{gt}_{u,v}||_1. \qquad (8)$$

$$\Delta\mathbf{J}_{u,v} = \sum_{m=1}^{M-1} \mathbf{J}_{I(m+1)} - \mathbf{J}_{I(m)}$$
$$= \sum_{m=1}^{M-1} sgn(parent(I(m)), I(m+1)) \cdot N^{-1}(\tilde{\mathbf{B}}_{I(m)}).$$

http://openaccess.thecvf.com/content_ICCV_2017/papers/Sun_Compositional_Human_Pose_ICCV_2017_paper.pdf

Compositional Human Pose Regression, Xiao Sun, Jiaxiang Shang, Shuang Liang, Yichen Wei, ICCV2017

3D Skeleton: Compositional Human Pose Regression

MEGVII 旷视

- Experiments

Training Data	Metric	Baseline	Ours (joint)	Ours(bone)	Ours (both)	Ours (all)
Human3.6M	Joint Error	102.2	$103.3_{\uparrow 1.1}$	$104.6_{\uparrow 2.4}$	$95.2_{\downarrow 7.0}$	$\mathbf{92.4}_{\downarrow 9.8}$
	PA Joint Error	75.0	$74.3_{\downarrow 0.7}$	$75.0_{\downarrow 0.0}$	$68.1_{\downarrow 6.9}$	$\mathbf{67.5}_{\downarrow 7.5}$
	Bone Error	65.5	$63.5_{\downarrow 2.0}$	$62.3_{\downarrow 3.2}$	$59.1_{\downarrow 6.4}$	$\mathbf{58.4}_{\downarrow 7.1}$
	Bone Std	26.4	$23.9_{\downarrow 2.5}$	$21.9_{\downarrow 4.5}$	$22.3_{\downarrow 4.1}$	$\mathbf{21.7}_{\downarrow 4.7}$
	Illegal Angle	3.7%	$3.2\%_{\downarrow 0.5}$	$3.3\%_{\downarrow 0.4}$	$2.6\%_{\downarrow 1.1}$	$\mathbf{2.5\%}_{\downarrow 1.2}$
Human3.6M + MPII	Joint Error	64.2	$62.9_{\downarrow 1.3}$	$63.8_{\downarrow 0.4}$	$60.7_{\downarrow 3.5}$	$\mathbf{59.1}_{\downarrow 5.1}$
	PA Joint Error	51.4	$50.6_{\downarrow 0.8}$	$50.4_{\downarrow 1.0}$	$48.8_{\downarrow 2.6}$	$\mathbf{48.3}_{\downarrow 3.1}$
	Bone Error	49.5	$49.3_{\downarrow 0.2}$	$47.4_{\downarrow 2.1}$	$47.2_{\downarrow 2.3}$	$\mathbf{47.1}_{\downarrow 2.4}$
	Bone Std	19.9	$19.3_{\downarrow 0.6}$	$\mathbf{17.5}_{\downarrow 2.4}$	$17.6_{\downarrow 2.3}$	$18.0_{\downarrow 1.9}$

Table 2. Results of all methods under all evaluation metrics (the lower the better), with or without using MPII data in training. Note that the performance gain of all *Ours* methods relative to the *Baseline* method is shown in the subscript. The *Illegal Angle* metric for "Human3.6M+MPII" setting is not included because it is very good ($< 1\%$) for all methods.

Yasin [52]	Rogez [40]	Chen [7]	Moreno [30]	Zhou [57]	Baseline	Ours (all)
108.3	88.1	82.7	76.5	55.3	51.4	**48.3**

Table 4. Comparison with previous work on Human3.6M. Protocol 1 is used. Evaluation metric is averaged *PA Joint Error*. Extra 2D training data is used in all the methods. *Baseline* and *Ours (all)* use MPII data in the training. *Ours (all)* is the best and also wins in all the 15 activity categories.

http://openaccess.thecvf.com/content_ICCV_2017/papers/Sun_Compositional_Human_Pose_ICCV_2017_paper.pdf

Compositional Human Pose Regression, Xiao Sun, Jiaxiang Shang, Shuang Liang, Yichen Wei, ICCV2017

3D Skeleton: Integral Human Pose Regression

- Motivation
  - Heatmap vs Regression
    - Heatmap: non-differentiable, quantization error
    - Regression: miss spatial structure
  - Integral loss

- Split or Joint Training
- 3D structure: 3D Heatmaps

$$\mathbf{J}_k = \arg\max_{\mathbf{p}} \mathbf{H}_k(\mathbf{p}).$$

$$\mathbf{J}_k = \int_{\mathbf{p}\in\Omega} \mathbf{p} \cdot \tilde{\mathbf{H}}_k(\mathbf{p}).$$

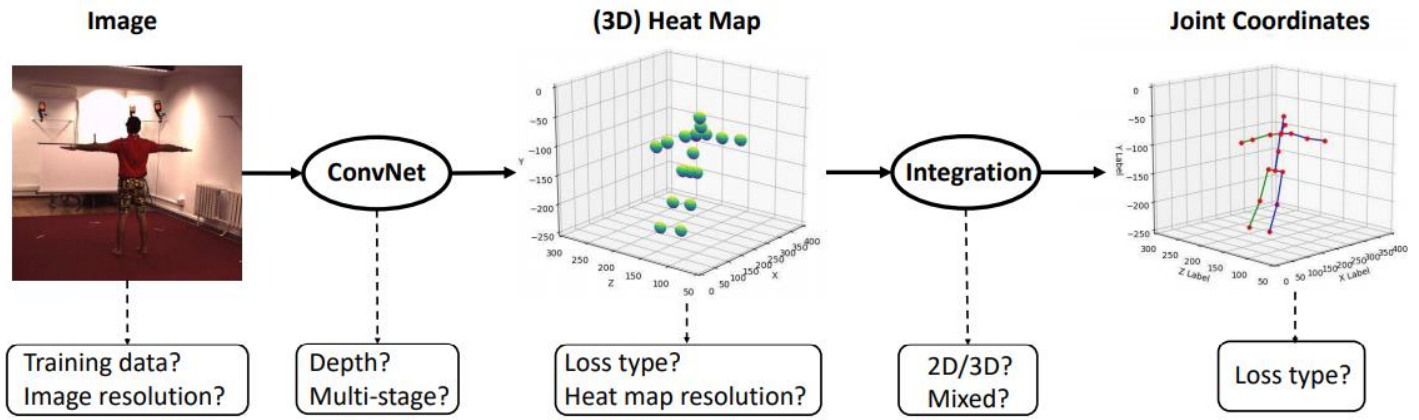


Fig. 1. Overview of pose estimation pipeline and all our ablation experiment settings.

http://openaccess.thecvf.com/content_ICCV_2017/papers/Sun_Compositional_Human_Pose_ICCV_2017_paper.pdf

Integral Human Pose Regression, Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, ECCV2018

3D Skeleton: Integral Human Pose Regression

- Experiments

Table 7. Comparison between methods using heat maps, direct regression, and integral regression. Protocol 2 is used. Two training strategies are investigated. Backbone network is ResNet-50. The relative performance gain is shown in the subscript

Training Data Strategy	R1	H1	H2	I*	I1	I2
Strategy1	106.6	99.5	80.4	$100.2_{\downarrow 6.0\%}$	$86.4_{\downarrow 13.2\%}$	$\mathbf{66.2}_{\downarrow 17.7\%}$
Strategy2	56.2	63.6	59.3	$\mathbf{49.6}_{\downarrow 11.7\%}$	$52.7_{\downarrow 17.1\%}$	$52.4_{\downarrow 11.6\%}$

Method (A, Pro. 1)	Hossain [21]*	Dabral [14]*	Yasin [51]	Rogez [41]	Chen [8]	Moreno [32]	Zhou [54]	Martinez [30]	Kanazawa [26]	Sun [42]	Fang [17]	Ours
PA MPJPE	42.0	36.3	108.3	88.1	82.7	76.5	55.3	47.7	56.8	48.3	45.7	**40.6**

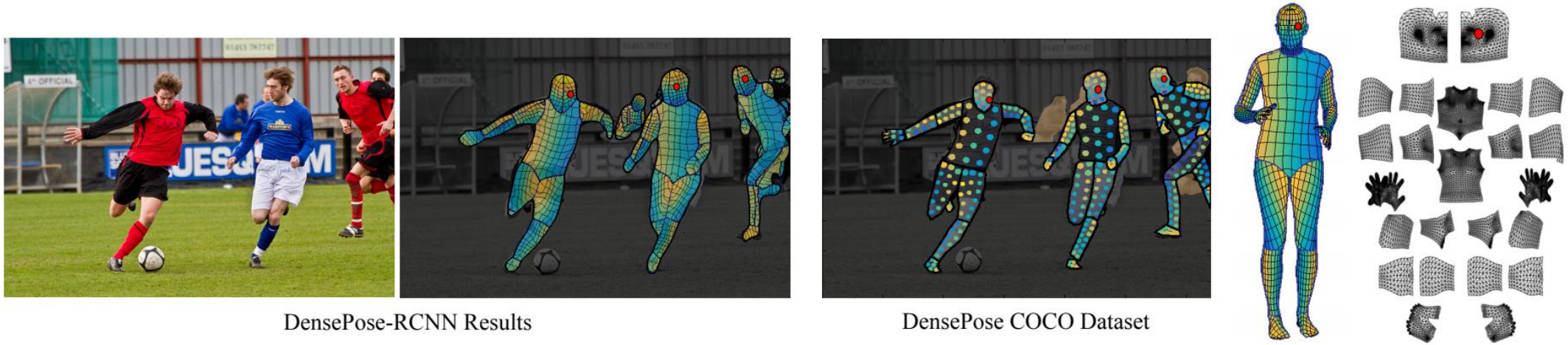
Method (B, Pro. 2)	Hossain [21]*	Dabral [14]*	Chen [8]	Tome [46]	Moreno [32]	Zhou [54]	Jahangiri [25]	Mehta [31]	Martinez [30]	Kanazawa [26]	Fang [17]	Sun [42]	Ours
MPJPE	51.9	52.1	114.2	88.4	87.3	79.9	77.6	72.9	62.9	88.0	60.4	59.1	**49.6**

http://openaccess.thecvf.com/content_ICCV_2017/papers/Sun_Compositional_Human_Pose_ICCV_2017_paper.pdf

Integral Human Pose Regression, Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, ECCV2018

3D Shape: DensePose

- Motivation
  - Dense Correspondence



DensePose-RCNN Results

DensePose COCO Dataset

Figure 1: Dense pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body. We introduce DensePose-COCO, a large-scale ground-truth dataset with image-to-surface correspondences manually annotated on 50K COCO images and train DensePose-RCNN, to densely regress part-specific UV coordinates within every human region at multiple frames per second. *Left:* The image and the regressed correspondence by DensePose-RCNN, *Middle:* DensePose COCO Dataset annotations, *Right:* Partitioning and UV parametrization of the body surface.

DensePose: Dense Human Pose Estimation In The Wild, Rıza Alp Güler, Natalia Neverova, Iasonas Kokkinos, CVPR2018

3D Shape: DensePose

- Dataset
  - DensePose-COCO Dataset

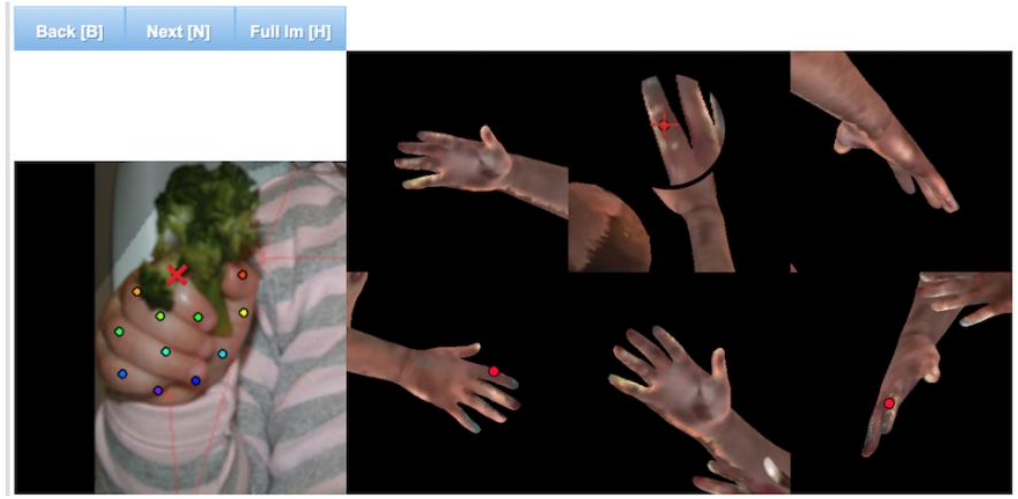


Figure 3: The user interface for collecting per-part correspondence annotations: We provide the annotators six pre-rendered views of a body part such that the whole part-surface is visible. Once the target point is annotated, the point is displayed on all rendered images simultaneously.

50K Images, 5M correspondences
24 UV Parts

DensePose: Dense Human Pose Estimation In The Wild, Rıza Alp Güler, Natalia Neverova, Iasonas Kokkinos, CVPR2018

3D Shape: DensePose

- Method

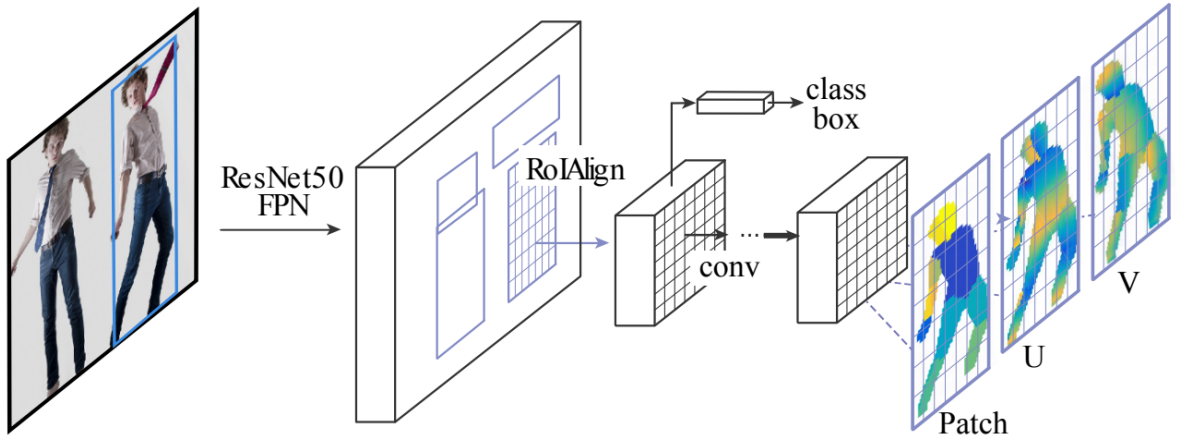


Figure 7: DensePose-RCNN architecture: we use a cascade of region proposal generation and feature pooling, followed by a fully-convolutional network that densely predicts discrete part labels and continuous surface coordinates.

DensePose: Dense Human Pose Estimation In The Wild, Rıza Alp Güler, Natalia Neverova, Iasonas Kokkinos, CVPR2018

3D Shape: DensePose

- Experiments

Method	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_M$	**AP**$_L$	**AR**	**AR**$_{50}$	**AP**$_{75}$	**AR**$_M$	**AR**$_L$
DensePose (ResNet-50)	51.0	83.5	54.2	39.4	53.1	60.1	88.5	64.5	42.0	61.3
DensePose (ResNet-101)	51.8	83.7	56.3	42.2	53.8	61.1	88.9	66.4	45.3	62.1
Multi-task learning										
DensePose + masks	51.9	85.5	54.7	39.4	53.9	61.1	89.7	65.5	42.0	62.4
DensePose + keypoints	52.8	85.6	56.2	42.2	54.7	62.6	89.8	67.7	45.4	63.7
Multi-task learning with cascading										
DensePose-cascade	51.6	83.9	55.2	41.9	53.4	60.4	88.9	65.3	43.3	61.6
DensePose + masks	52.8	85.5	56.1	40.3	54.6	62.0	89.7	67.0	42.4	63.3
DensePose + keypoints	55.8	87.5	61.2	48.4	57.1	63.9	91.0	69.7	50.3	64.8

Table 1: Per-instance evaluation of DensePose-RCNN performance on COCO `minival` subset. All multi-task experiments are based on ResNet-50 architecture. DensePose-cascade corresponds to the base architecture with an iterative refinement module with no input from other tasks.

DensePose: Dense Human Pose Estimation In The Wild, Rıza Alp Güler, Natalia Neverova, Iasonas Kokkinos, CVPR2018

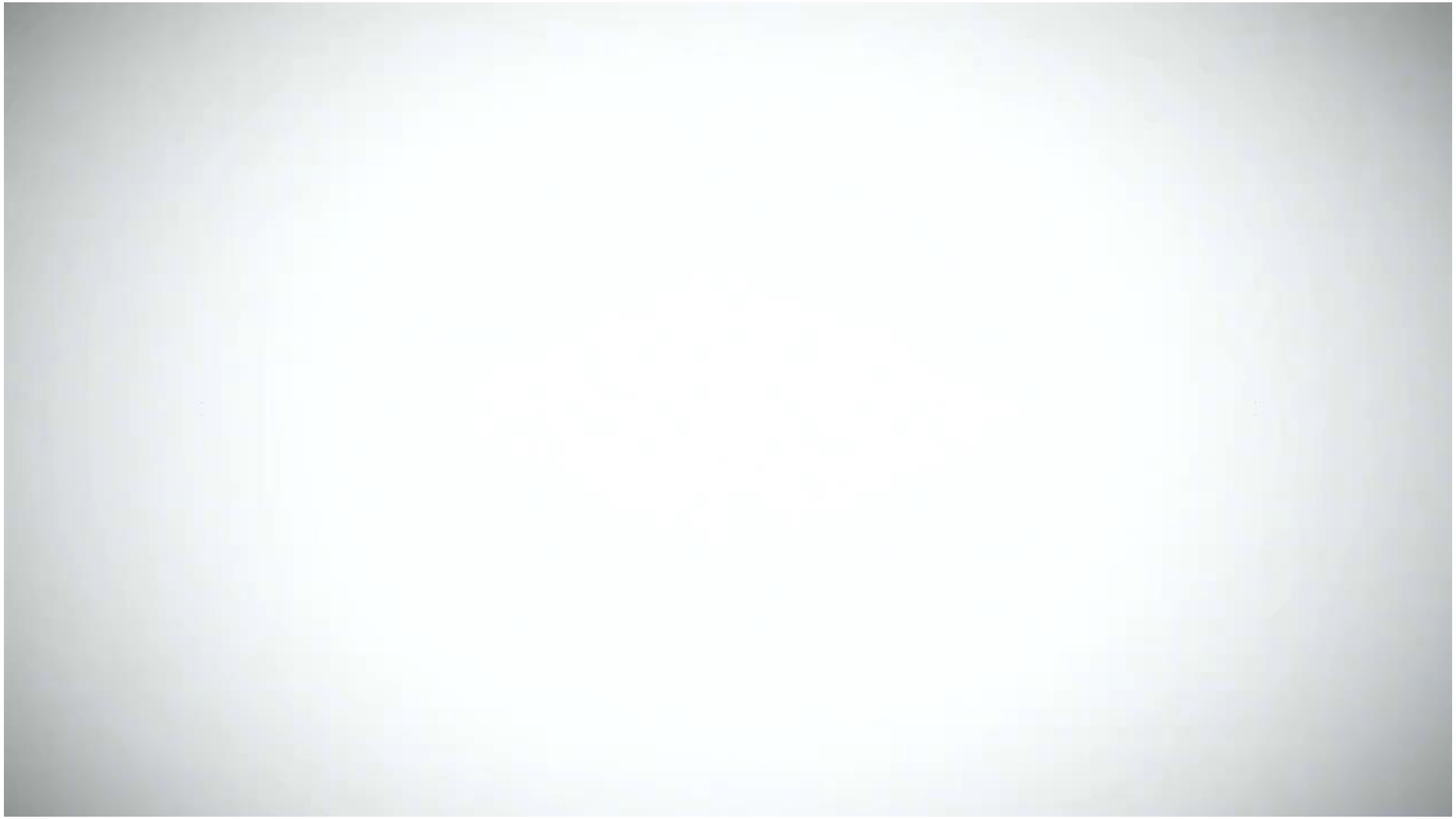
Summary for 3D Skeleton

- 3D Representation: 3D Skeleton vs 3D Shape
- 2D -> 3D Joint -> 3D Shape
- Remaining issues
  - Unconstrained (in the wild) benchmark
  - Ambiguous poses
  - Joint training of both 2D and 3D skeleton data

Outline

- Introduction to Human Pose Estimation
- 2D Skeleton
  - Top-Down
  - Bottom-Up
- 3D Skeleton
  - 2D -> 3D Skeleton
  - 2D -> 3D Shape
- Application
- Conclusion

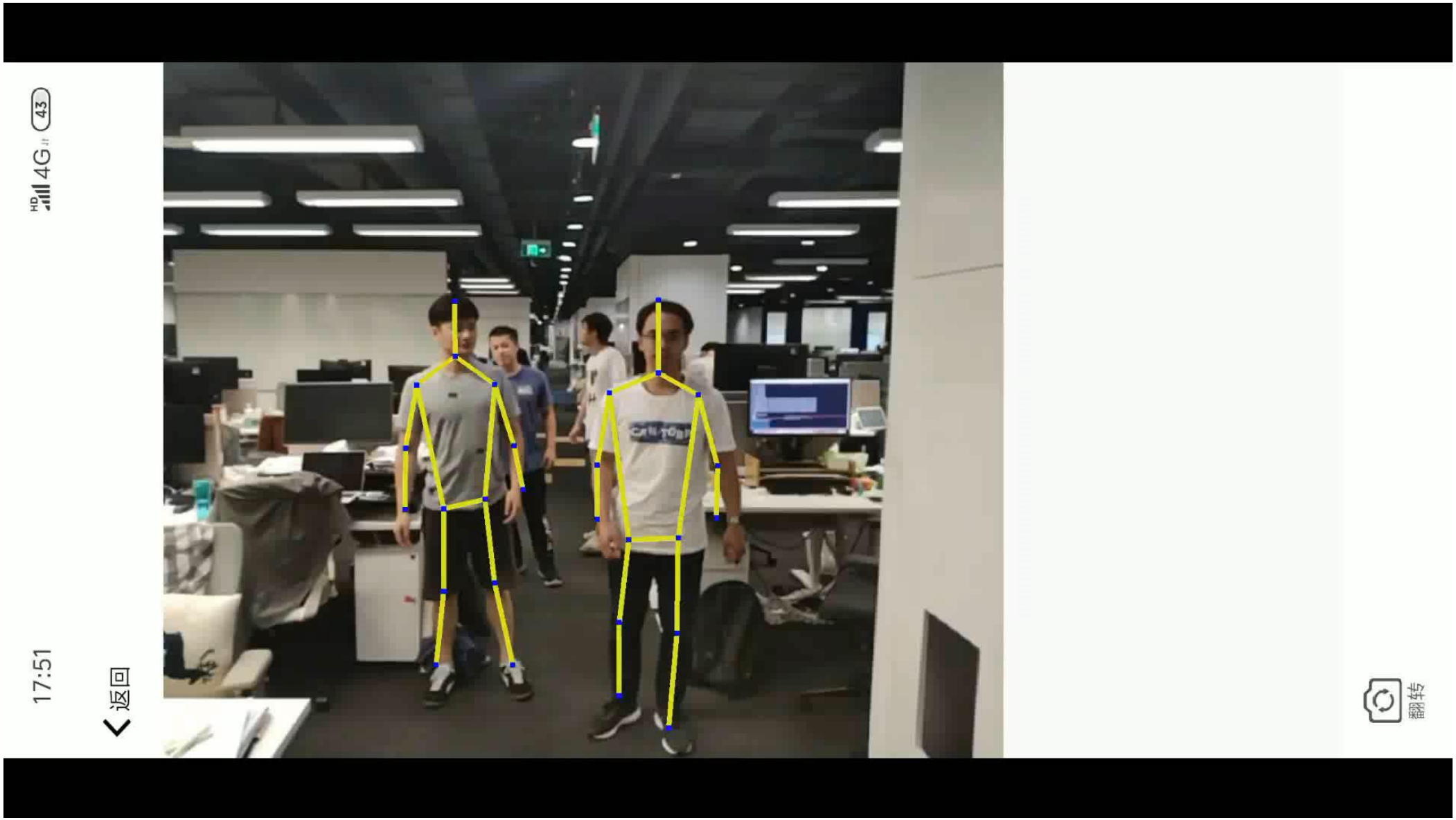
Application: Action Recognition



Application: Robotics

Application: Human-Computer Interaction

Application: Mobile Applications



Outline

- Introduction to Human Pose Estimation
- 2D Skeleton
  - Top-Down
  - Bottom-Up
- 3D Skeleton
  - 2D -> 3D Skeleton
  - 2D -> 3D Shape
- Application
- Conclusion

Conclusion

- 2D Skeleton (context, resolution) -> 3D Skeleton (regression) -> 3D shape (Representation)
- A lot of potential applications based on Skeleton
  - Action, Interaction, Game
- An improvement of skeleton is a large step for the industry

MEGVII 旷视