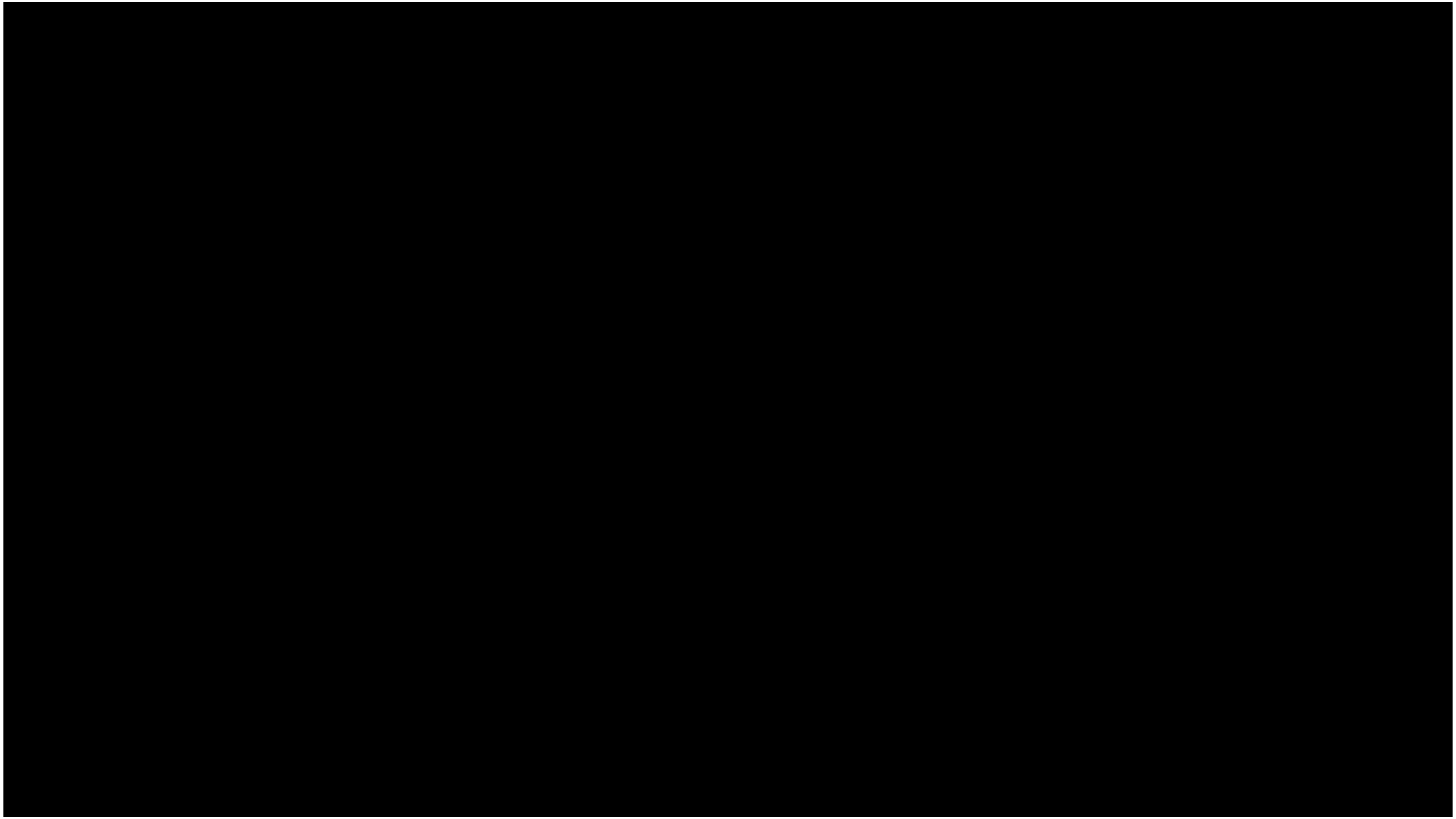
Human Pose Estimation Algorithm and Application

Gang Yu
yugang@megvii.com

Outline

- Introduction to Human Pose Estimation
- Algorithm
  - Cascade Pyramid Network
  - Multi-stage Pose Estimation
- 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):

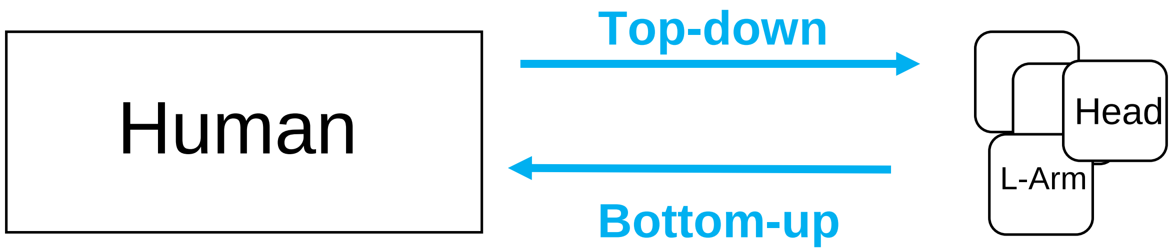
AR	% AR at OKS=.50:.05:.95
$AR^{OKS=.50}$	% AR at OKS=.50
$AR^{OKS=.75}$	% AR at OKS=.75

AR Across Scales:

AR^{medium}	% AR for medium objects: 32^2 < area < 96^2
AR^{large}	% AR for large objects: area > 96^2

How to Do Pose Estimation: Top-down vs Bottom-up

- 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

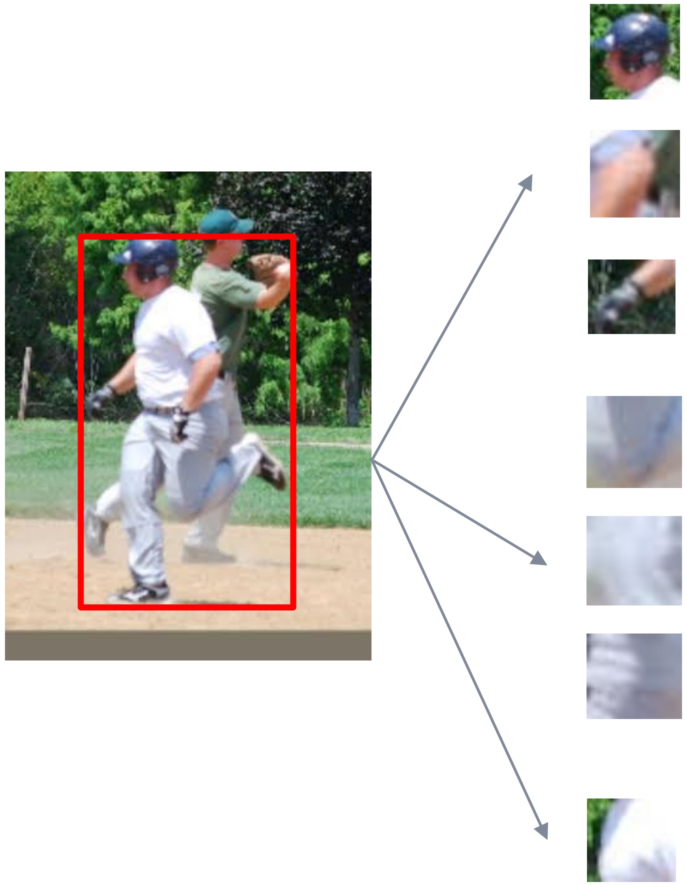
- Ambiguous Appearance
- Crowd Case
- Large Pose
- Inference Speed

Outline

- Introduction to Human Pose Estimation
- Algorithm
  - Cascade Pyramid Network
  - Multi-stage Pose Estimation
- Application
- Conclusion

ALGORITHM: Cascade Pyramid Network

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

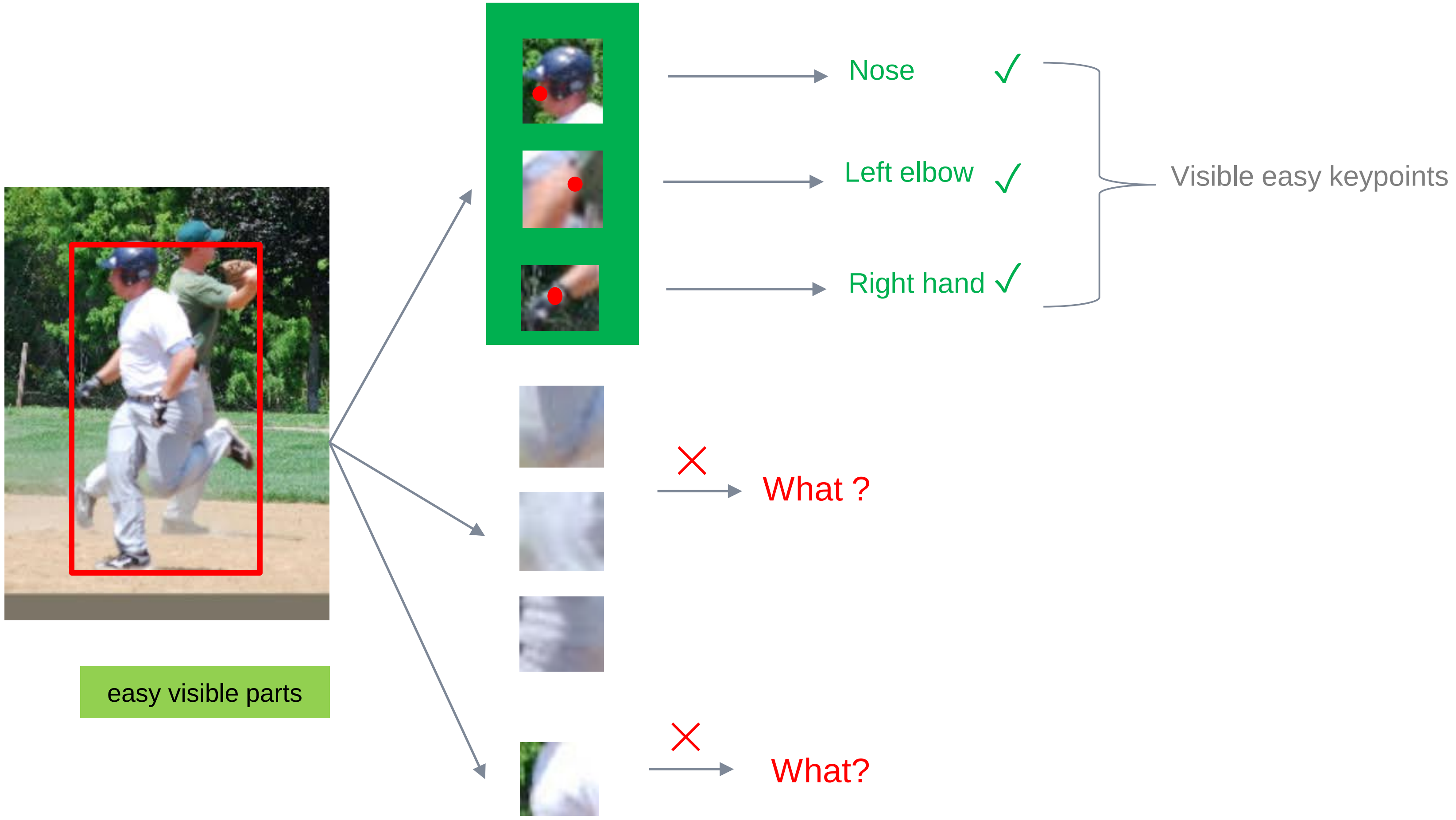


Cascaded Pyramid Network for Multi-Person Pose Estimation, Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, Jian Sun, CVPR 2018

Algorithm: Cascade Pyramid Network

MEGVII 旷视

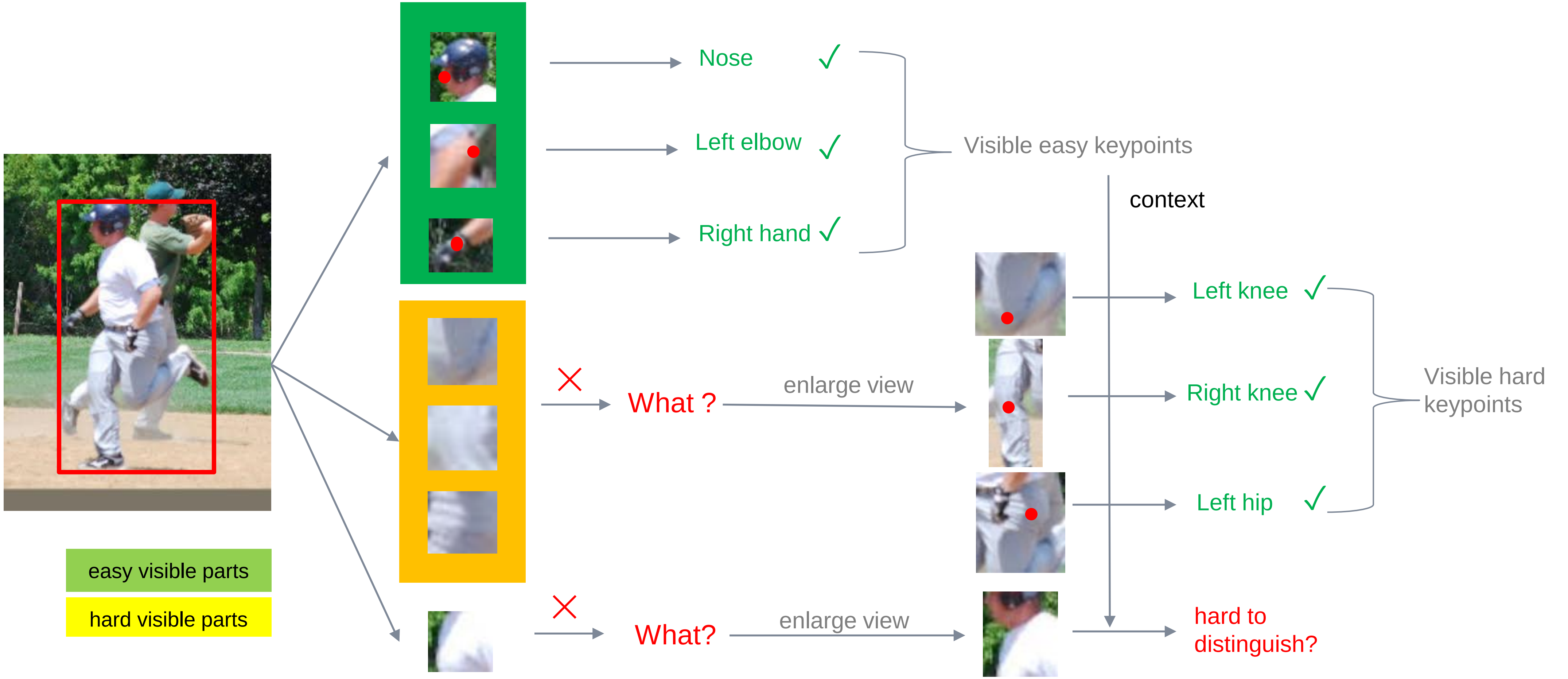
- Motivation: How to locate the “hard” joints
- Human perspective



Algorithm: Cascade Pyramid Network

MEGVII 旷视

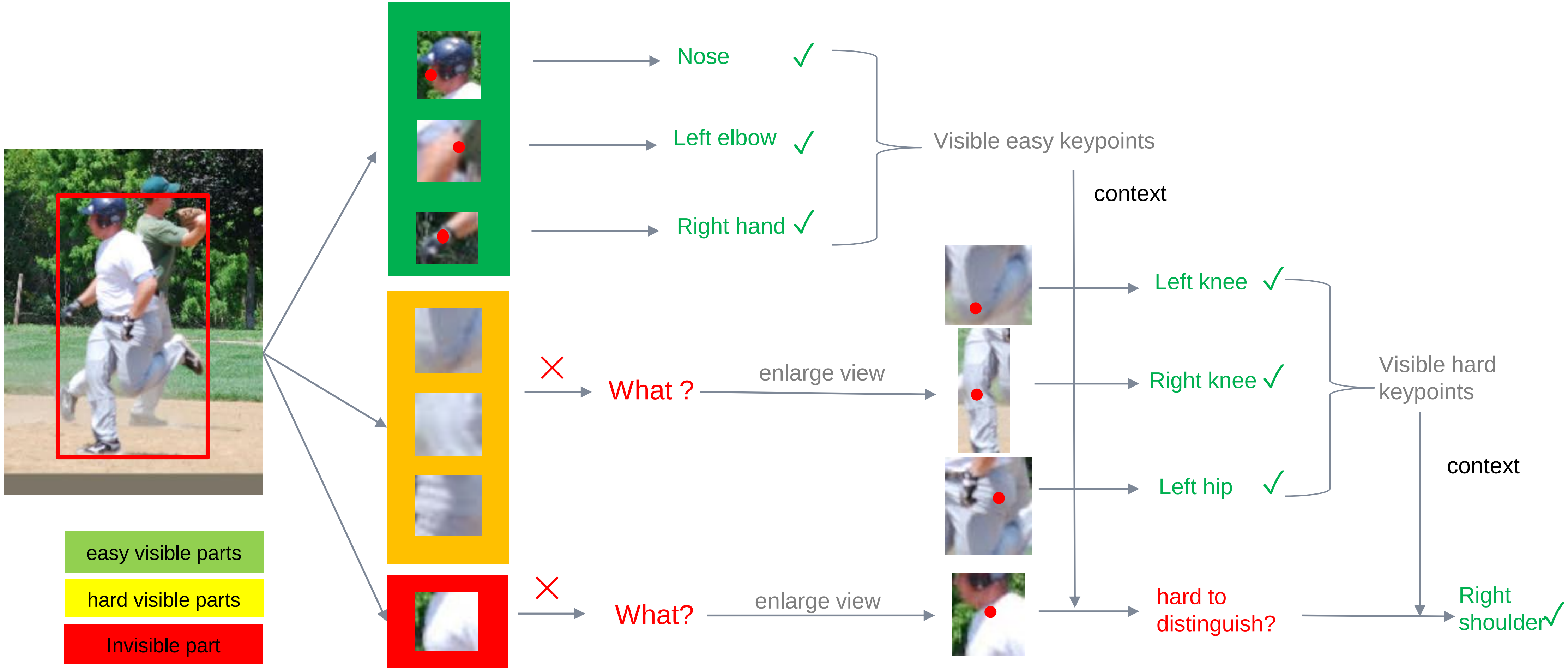
- Motivation: How to locate the “hard” joints
- Human perspective



Algorithm: Cascade Pyramid Network

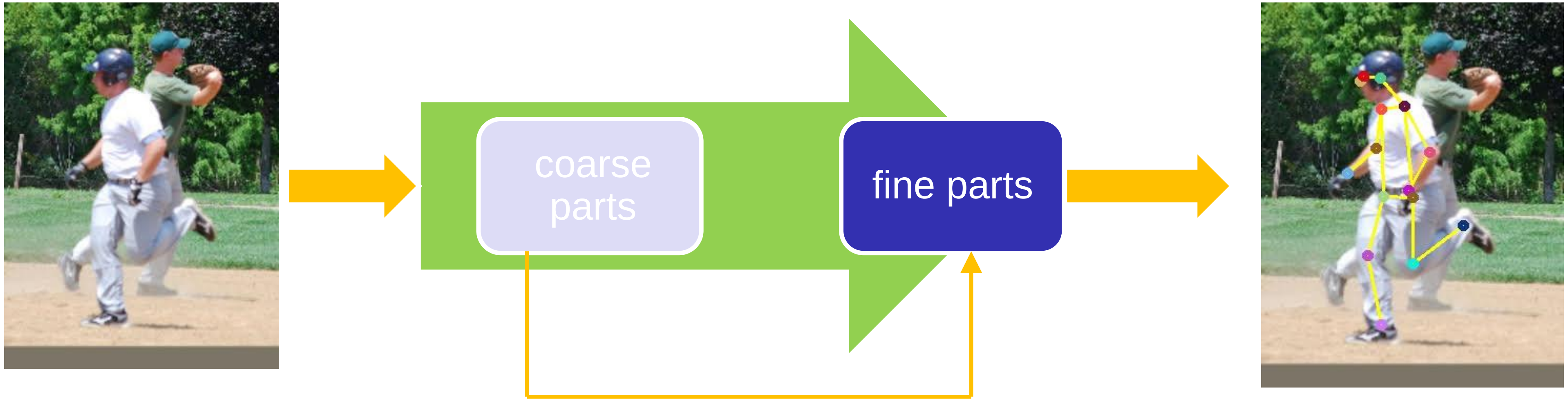
MEGVII 旷视

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



Algorithm: Cascade Pyramid Network

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

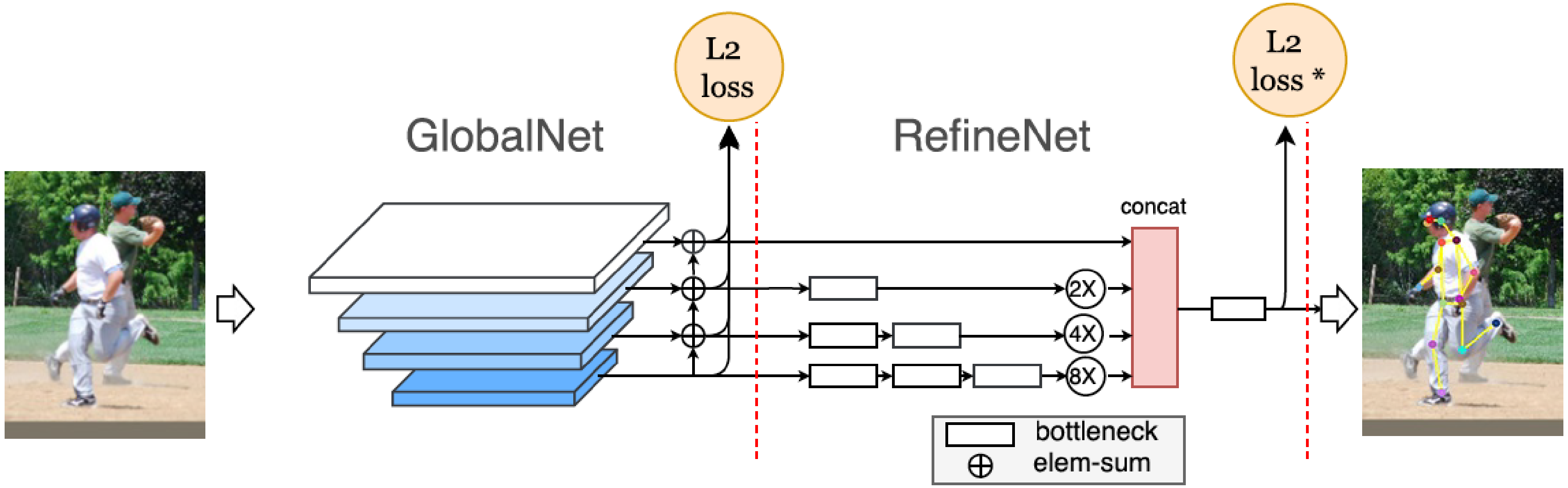


Input image

receptive view getting larger
& more context

Output image

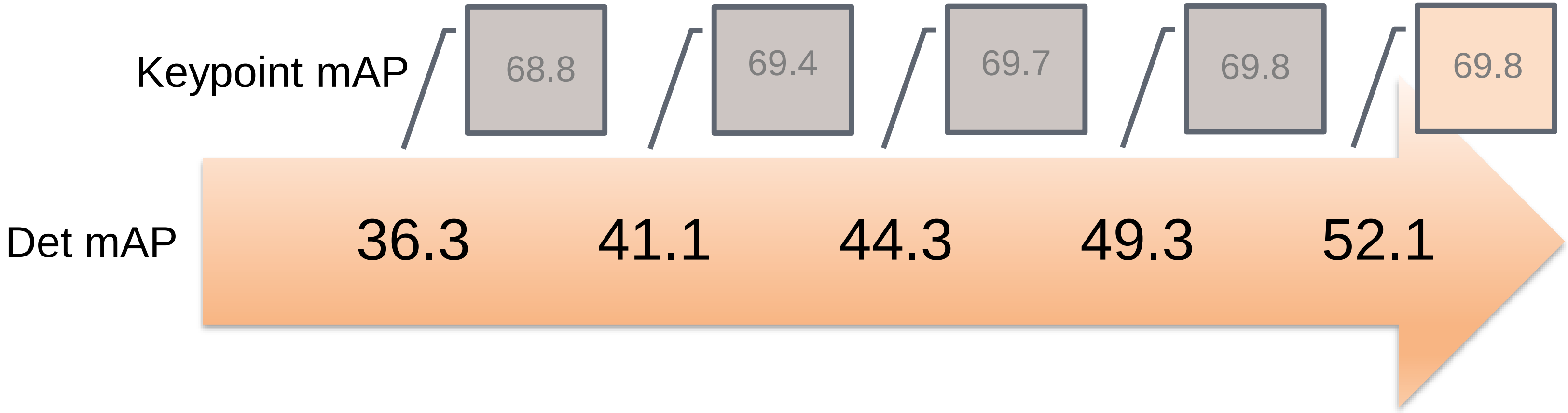
Network Architecture



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 keypoints 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

MEGVII 旷视

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.

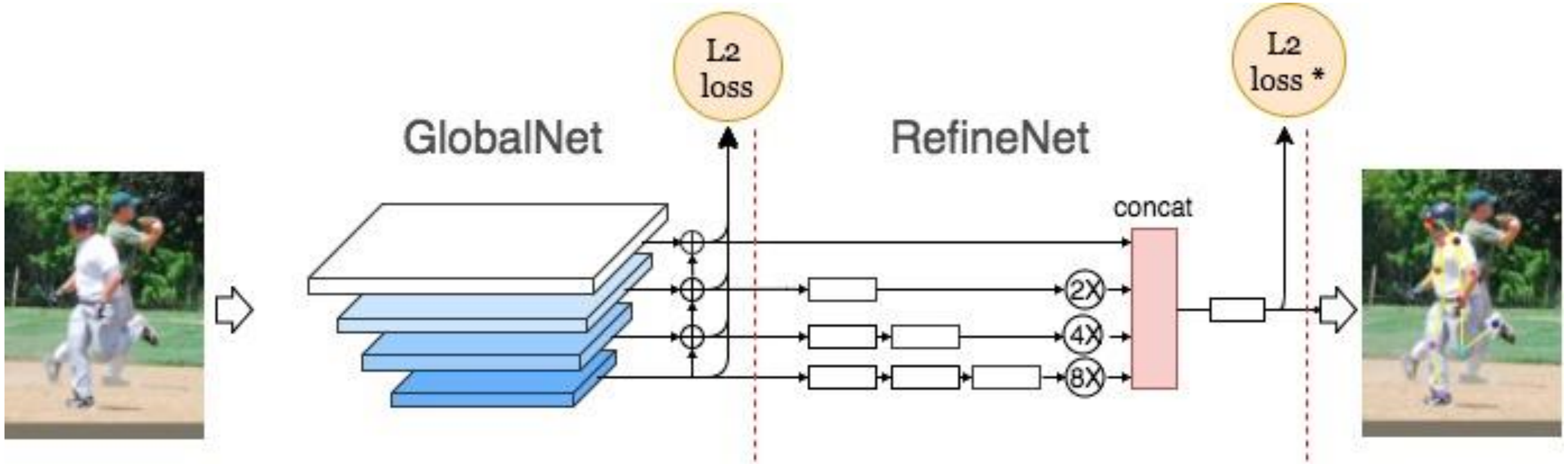
GlobalNet	RefineNet	AP(OKS)
-	L2 loss	68.2
L2 loss	L2 loss	68.6
-	L2 loss*	68.5
L2 loss	L2 loss*	69.4
L2 loss*	L2 loss*	69.1

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

MEGVII 旷视

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

MEGVII 旷视

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

COCO
Common Objects in Context

info@cocodataset.org

Home People Dataset Tasks Evaluate

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

Outline

- Introduction to Human Pose Estimation
- Algorithm
  - Cascade Pyramid Network
  - Multi-stage Pose Estimation
- Application
- Conclusion

Algorithm: Multi-stage Pose Estimation

MEGVII 旷视

- Motivation
  - Upperbound
  - Only Two-stages available

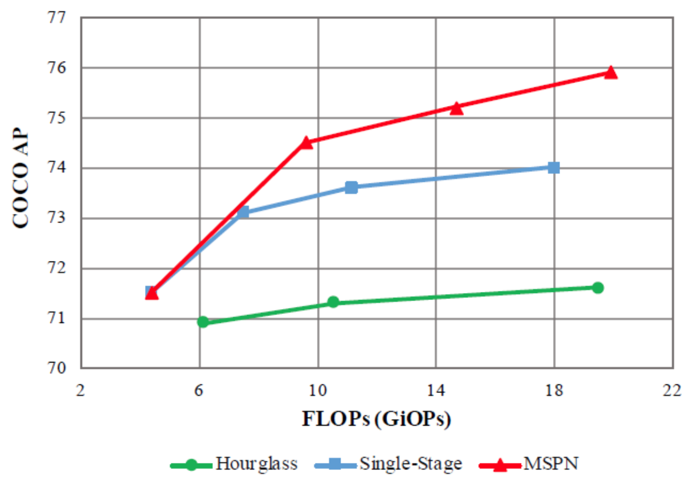


Figure 1. Pose estimation performance on COCO minival dataset of Hourglass [29], a single-stage model using ResNet [17], and our proposed MSPN under different model capacity (measured in FLOPs).

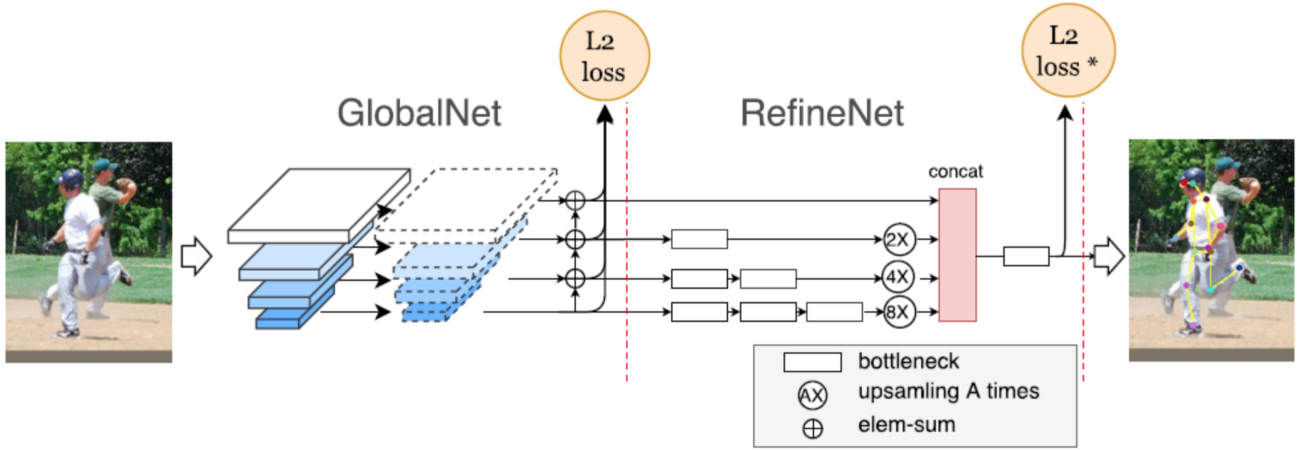
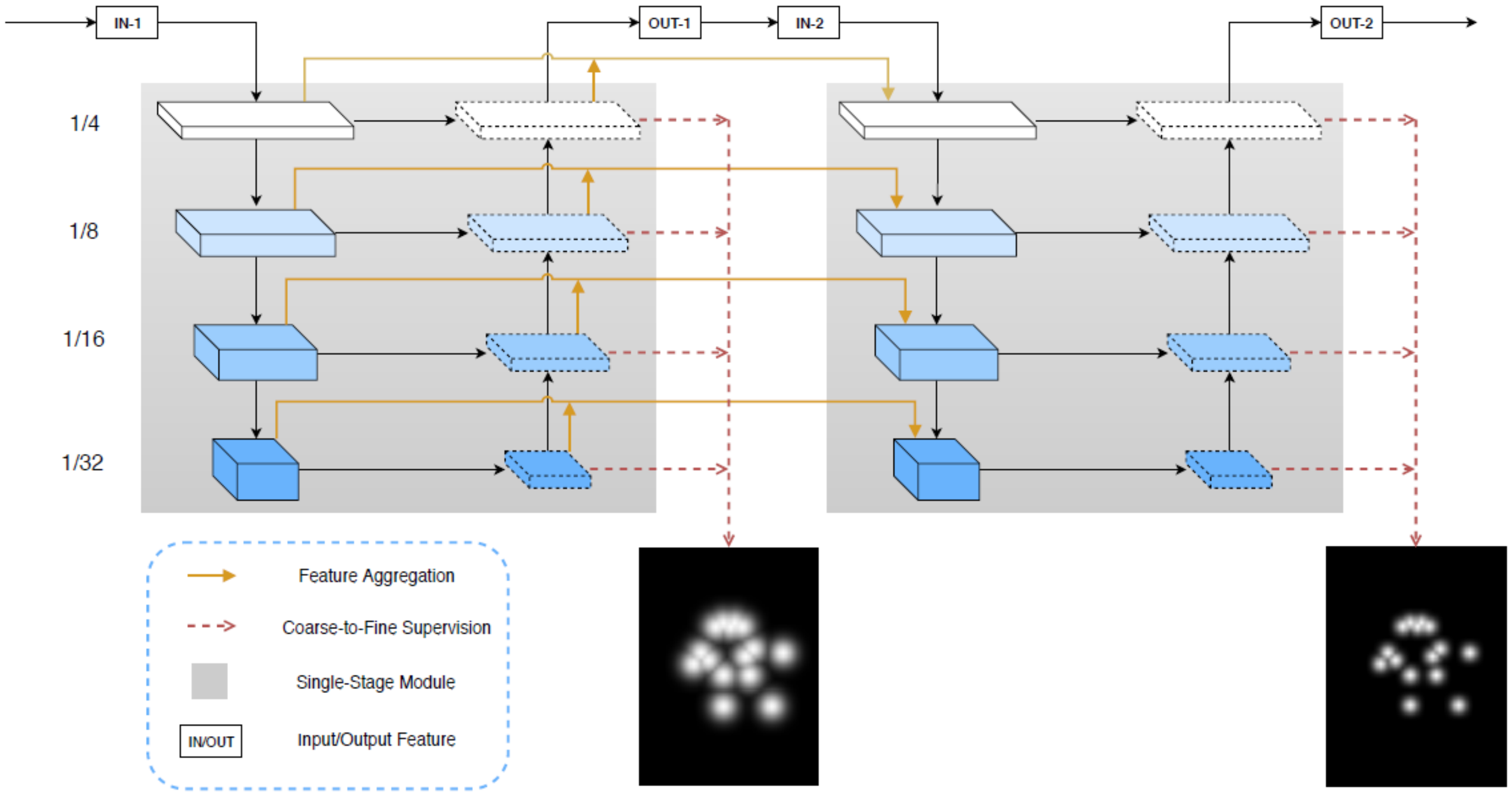


Figure 1. Cascaded Pyramid Network. “L2 loss*” means L2 loss with online hard keypoints mining.

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

Multi-stage Pose Estimation

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



Multi-stage Pose Estimation

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

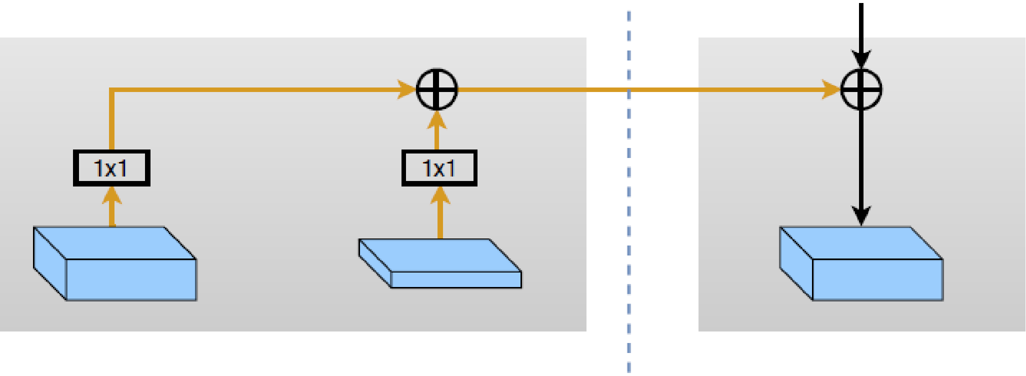


Figure 3. Cross Stage Feature Aggregation on a specific scale. Two 1×1 convolutional operations are applied to the features of previous stage before aggregation. See Figure 2 for the overall network structure.

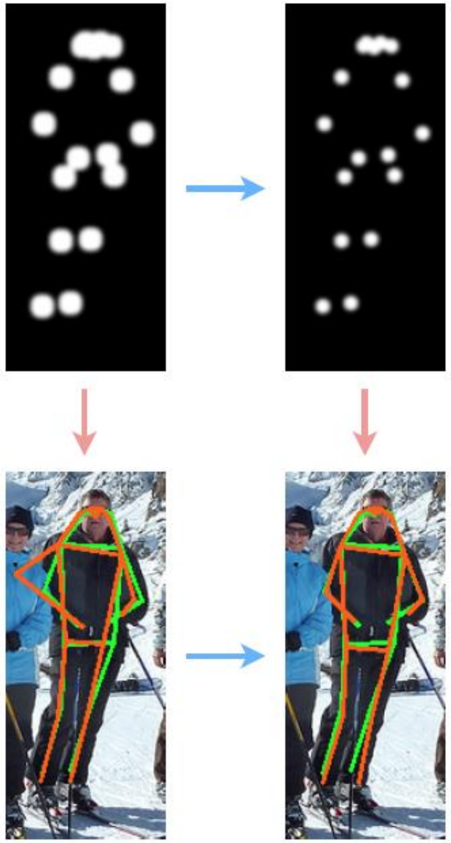


Figure 4. Illustration of coarse-to-fine supervision. The first row shows ground-truth heat maps in different stages and the second row represents corresponding predictions and ground truth annotations. The orange line is the prediction result and the green line indicates ground truth.

Experiments: More Stages

MEGVII 旷视

Stages	Hourglass		Stages	MSPN	
	FLOPs(G)	AP		FLOPs(G)	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

MEGVII 旷视

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

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

MEGVII 旷视

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

COCO Common Objects in Context

info@cocodataset.org

Home People Dataset Tasks Evaluate

Keypoint Leaderboard

Dev Challenge16 Challenge17 Challenge18

Copy to Clipboard Export to CSV

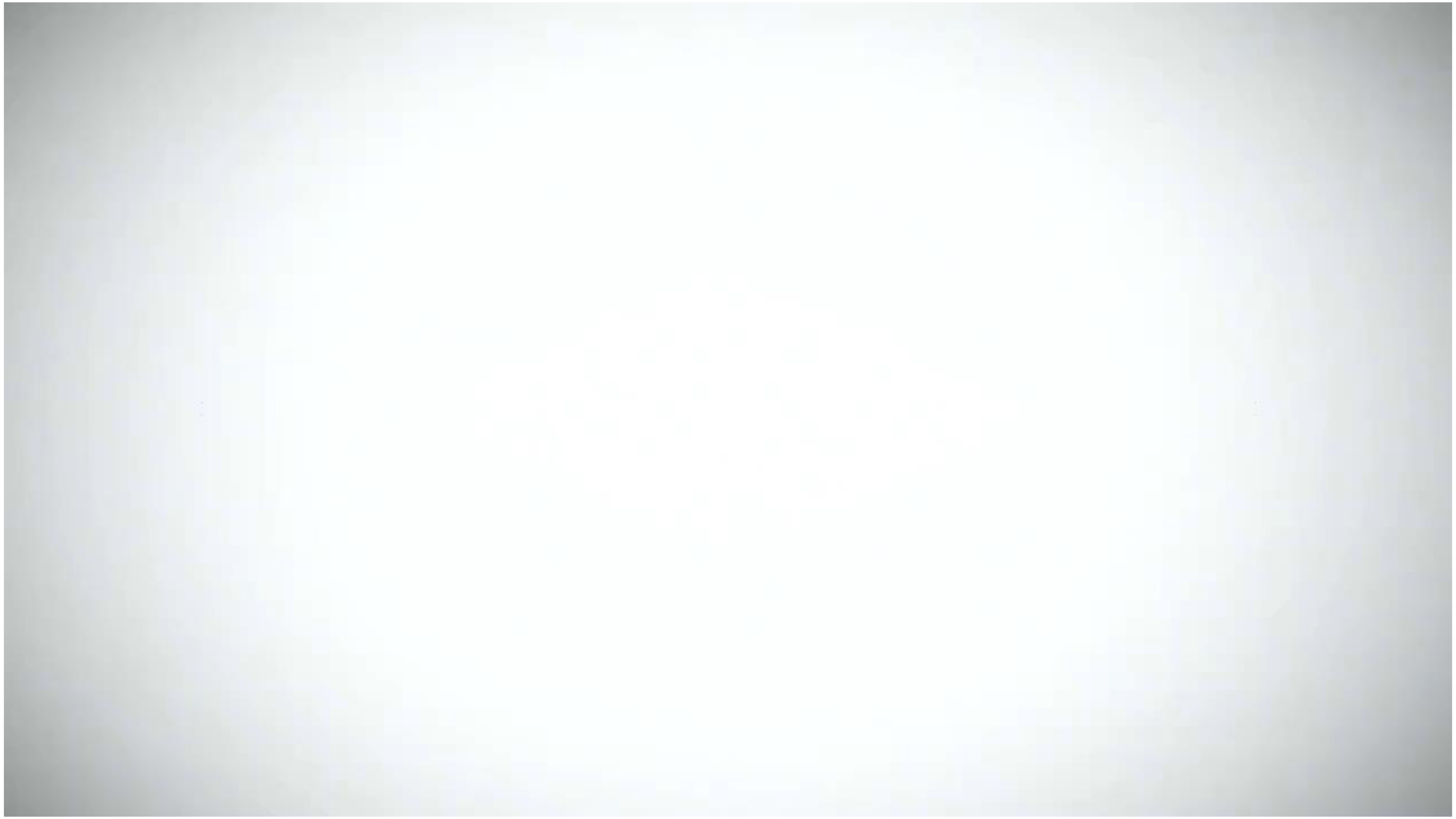
Search:

	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

Outline

- Introduction to Human Pose Estimation
- Algorithm
  - Cascade Pyramid Network
  - Multi-stage Pose Estimation
- Application
- Conclusion

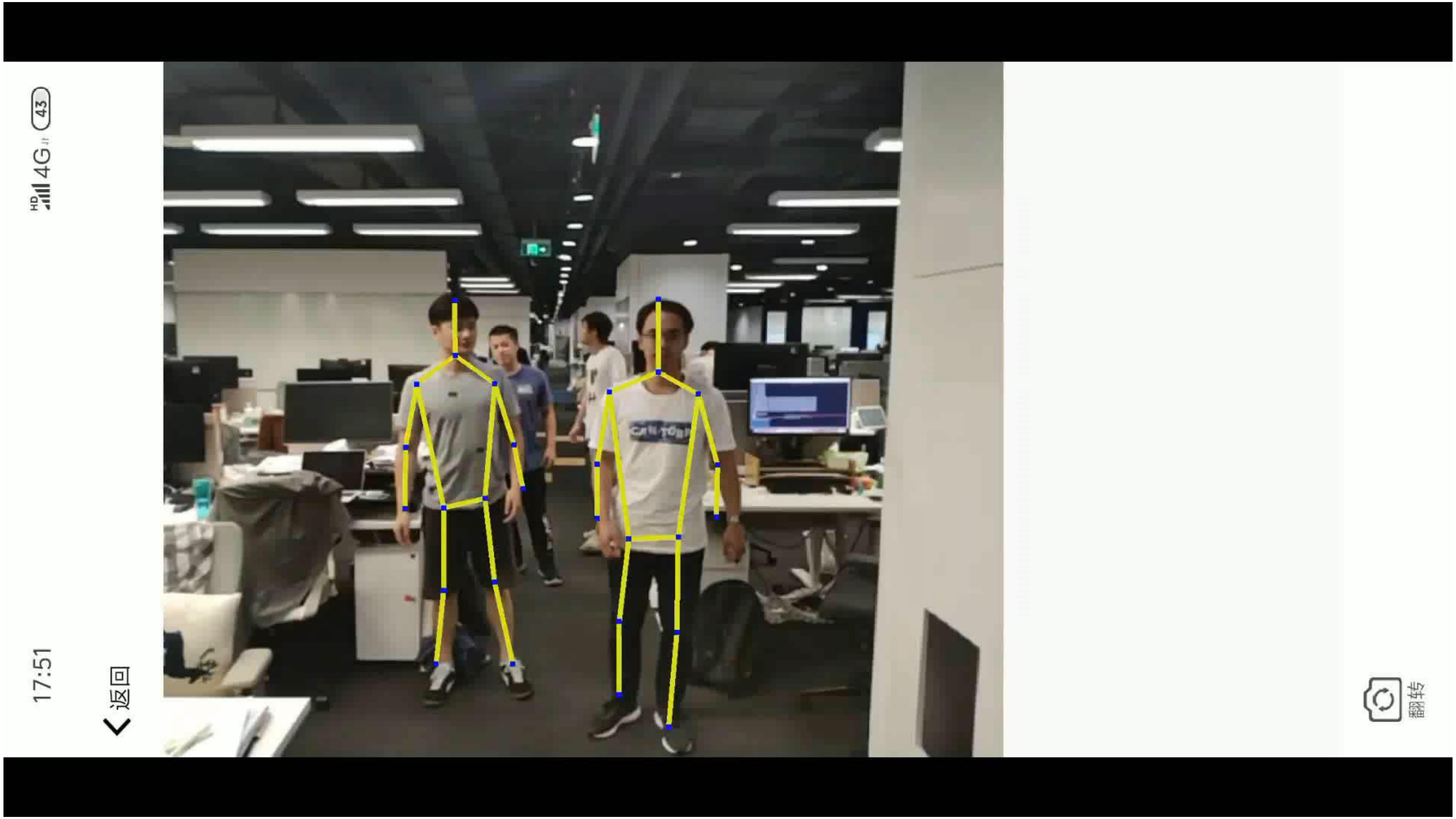
Application: Action Recognition



Application: Robotics

Application: Human-Computer Interaction

Application: Mobile Applications



Outline

- Introduction to Human Pose Estimation
- Algorithm
  - Cascade Pyramid Network
  - Multi-stage Pose Estimation
- Application
- Conclusion

Conclusion

- Context is important for Skeleton
  - Coarse to fine Strategy
- A lot of potential applications based on Skeleton
- An improvement of skeleton is a large step for the industry

广告部分

- Megvii Detection 知乎专栏

MEGVII 旷视