Cascaded Pyramid Network for Multi-Person Pose Estimation

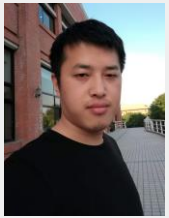
Gang YU
yugang@megvii.com

Megvii (Face++)

Face++ 旷视

Team members:

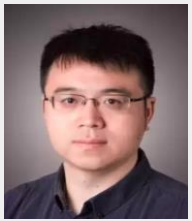

Yilun Chen*


Zhicheng Wang*


Xiangyu Peng


Zhiqiang Zhang


Gang Yu


Jian Sun

(https://arxiv.org/abs/1711.07319)

Code:https://github.com/allenchen9512/tf-cpn

Megvii (Face++)

Results

- COCO 17 Keypoints (test_challenge)

	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

Overview

- Top-down Pipeline
- Network Design
  - Motivation: How human locate keypoints?
  - Our Network Architecture
- Techniques & Experiments
- Conclusion

Overview

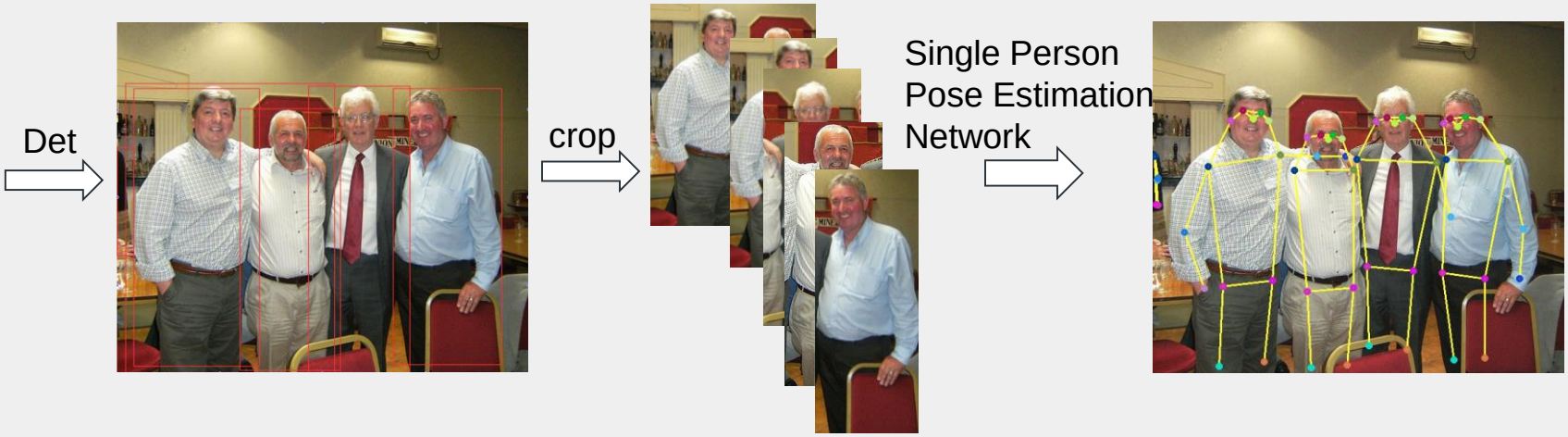
- Top-down Pipeline

Top-Down pipeline

Det

Top-Down pipeline

Top-Down pipeline



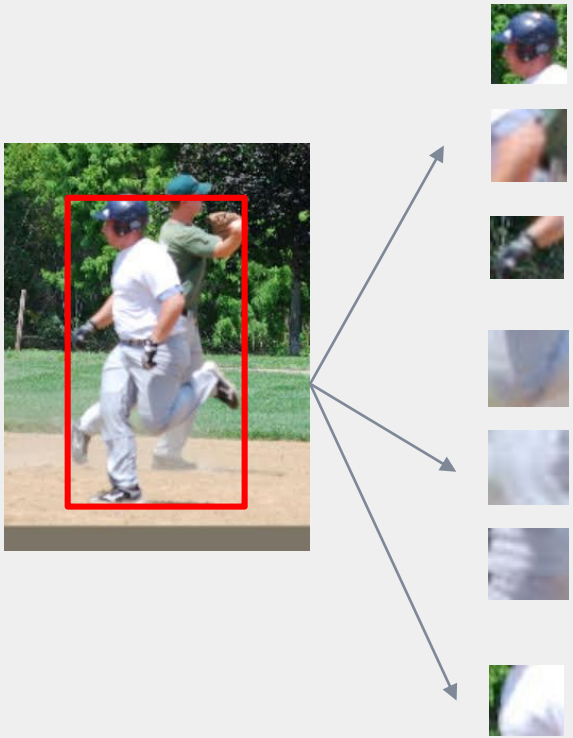
Overview

- Top-down Pipeline
- Network Design

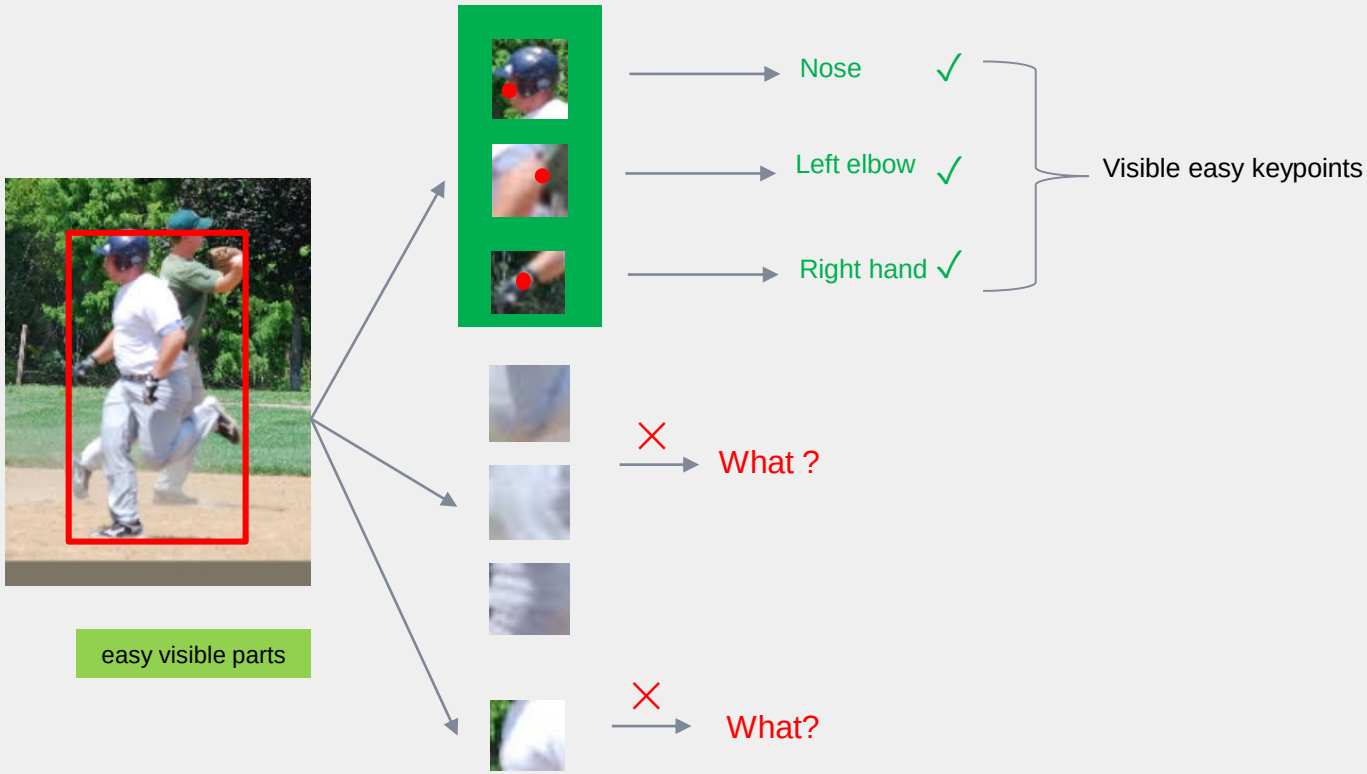
Overview

- Top-down Pipeline
- Network Design
  - Motivation: How human locates keypoints?

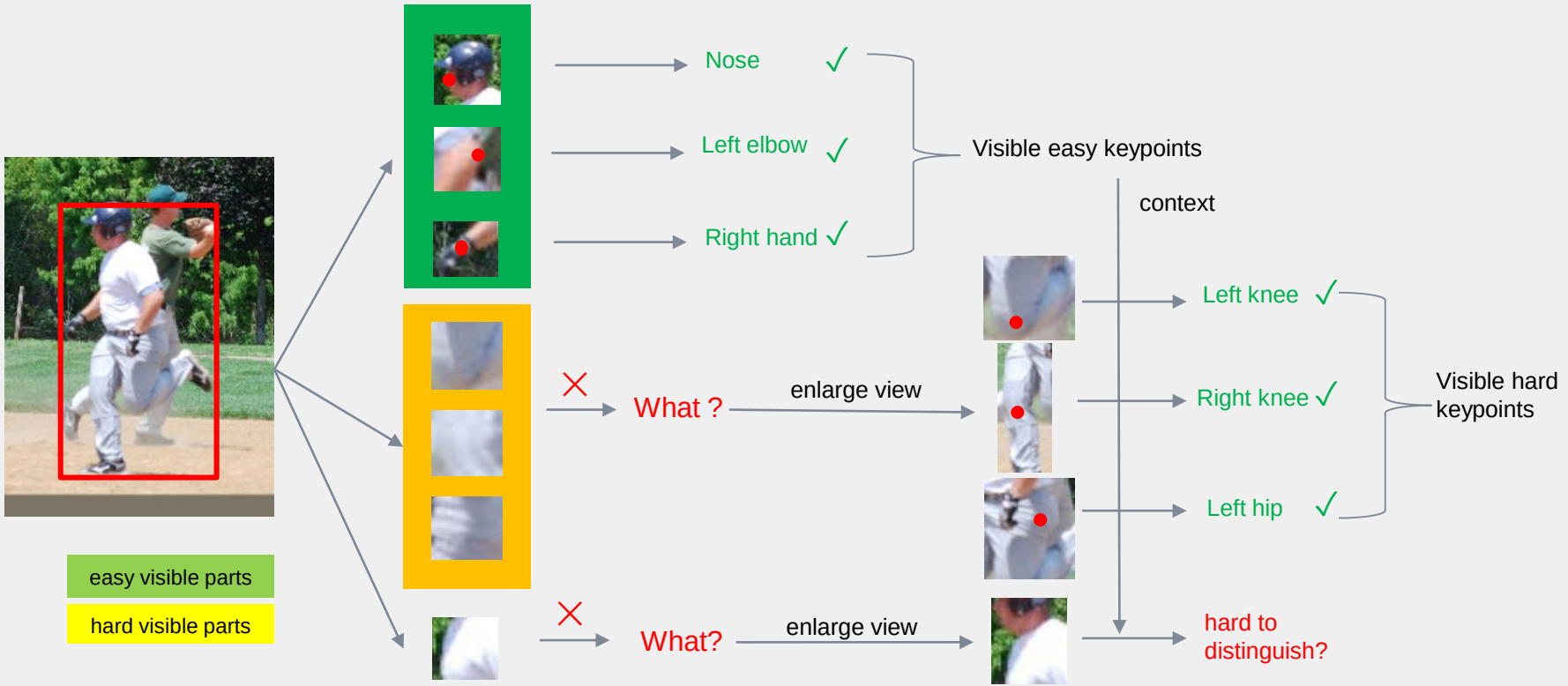
Motivation:
How human locate keypoints?



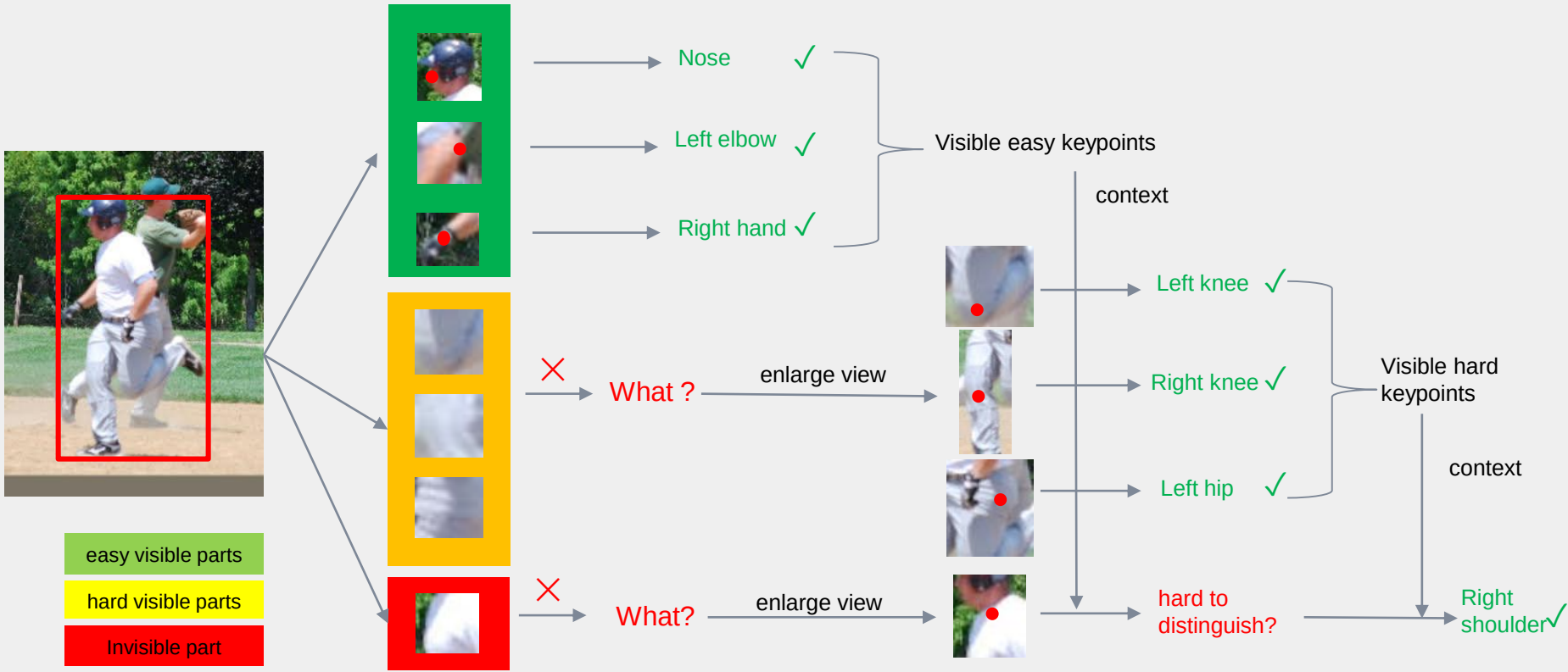
Motivation:
How human locate keypoints?



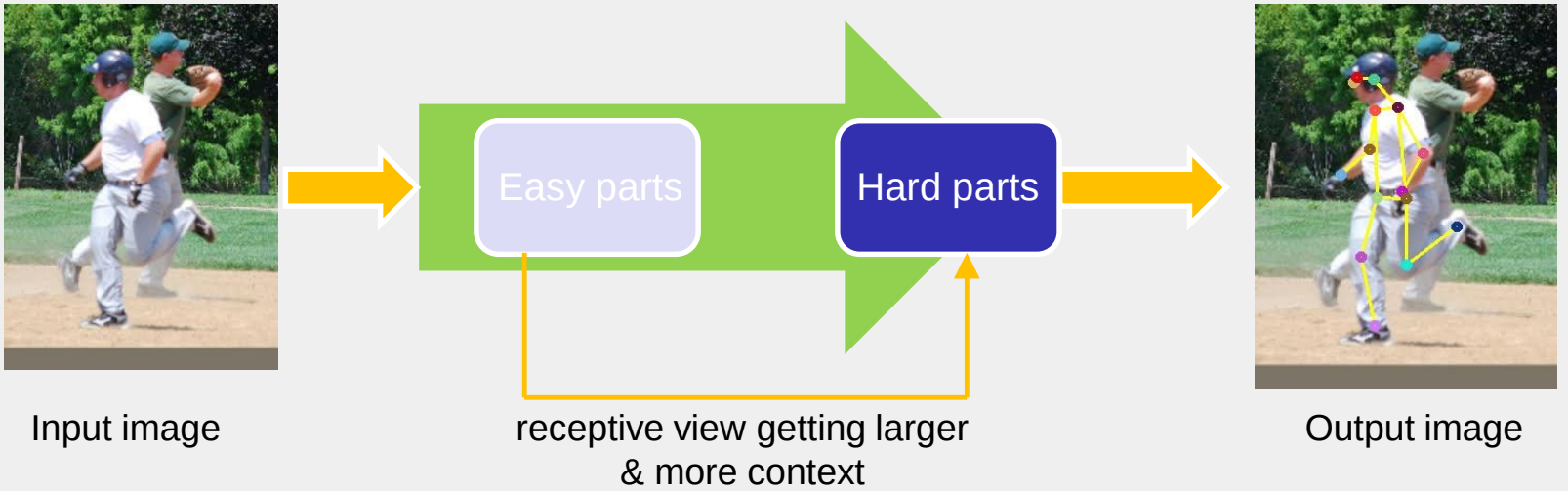
Motivation:
How human locate keypoints?



Motivation:
How human locate keypoints?



Network's Design Goal

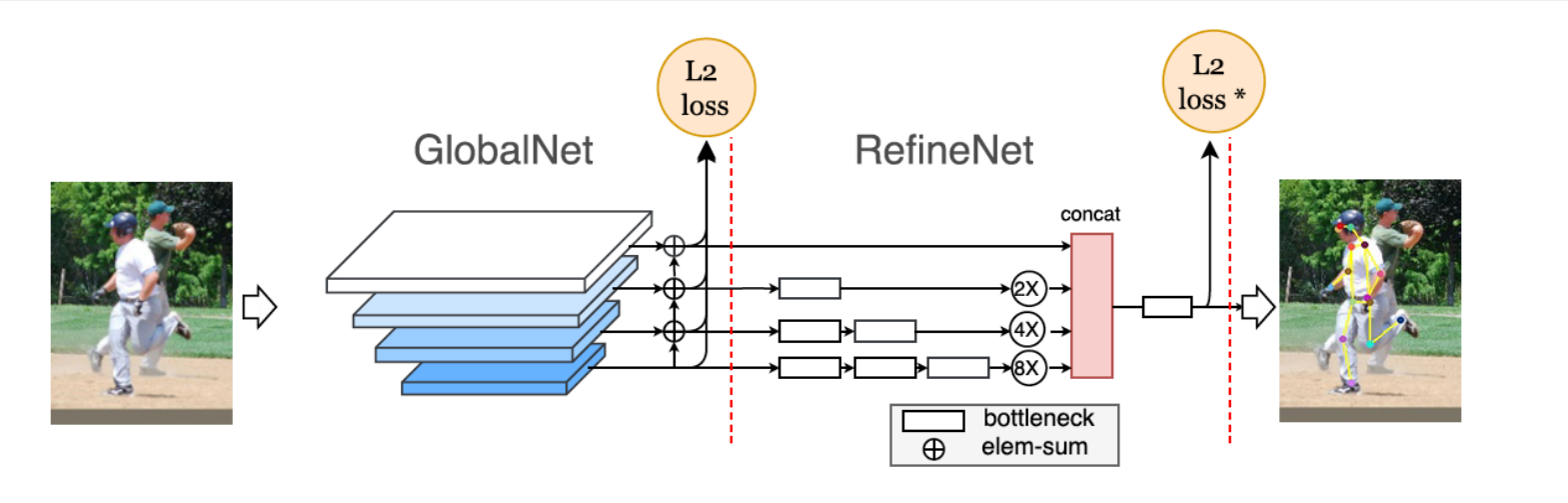


Overview

- Top-down Pipeline

- Network Design
  - Motivation: How human locate keypoints?
  - Our Network Architecture

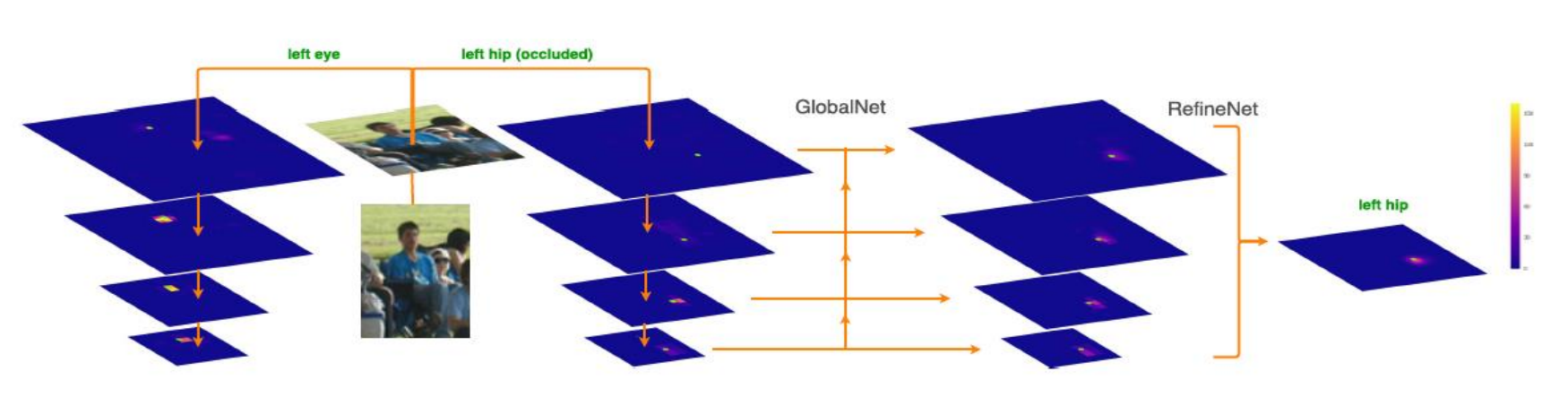
Network Architecture



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)

Network Architecture



The green dots means the groundtruth location of keypoints.

Heatmap view:

- Easy parts like left eye successfully been detected, while hard parts like left hip fail to be detected in GlobalNet.
- Hard parts like left hip successfully been detected in the RefineNet stage.

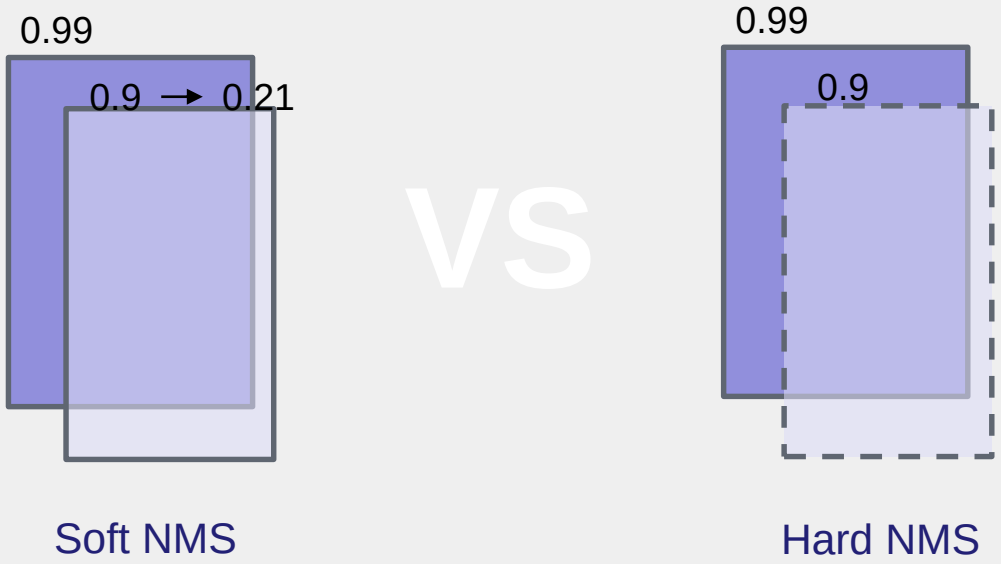
Overview

- Top-down Pipeline
- Network Design
  - Motivation: How human locate keypoint?
  - Our Network Architecture
- Techniques & Experiments

Techniques & Experiments

Person Detector

Non-Maximum Suppression (NMS) strategies



Techniques & Experiments

Person Detector

Non-Maximum Suppression (NMS) strategies

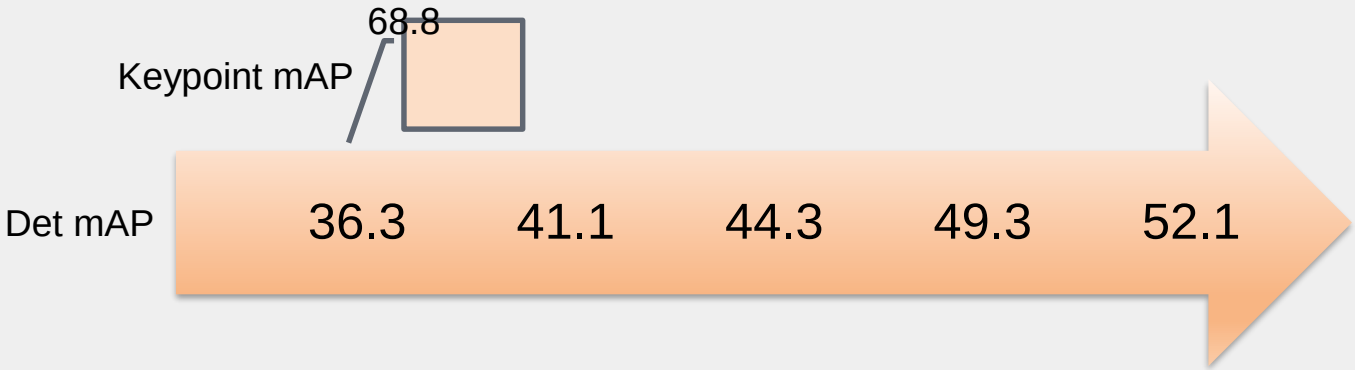
NMS	AP(all)	AP(H)	AR(H)	AP(OKS)
NMS(thr=0.3)	40.1	53.5	60.3	68.2
NMS(thr=0.4)	40.5	54.4	61.7	68.9
NMS(thr=0.5)	40.8	54.9	62.9	69.2
NMS(thr=0.6)	40.8	55.2	64.3	69.2
Soft-NMS [4]	41.1	55.3	67.0	69.4

Table 1. Comparison between different NMS methods and key-points detection performance with the same model. H is short for human.

Techniques & Experiments

Person Detector

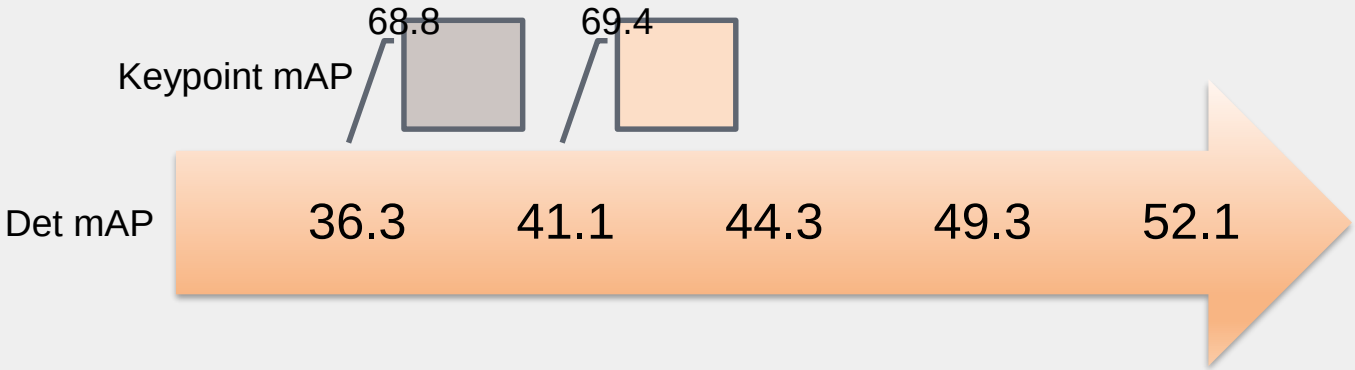
Detection Performance



Techniques & Experiments

Person Detector

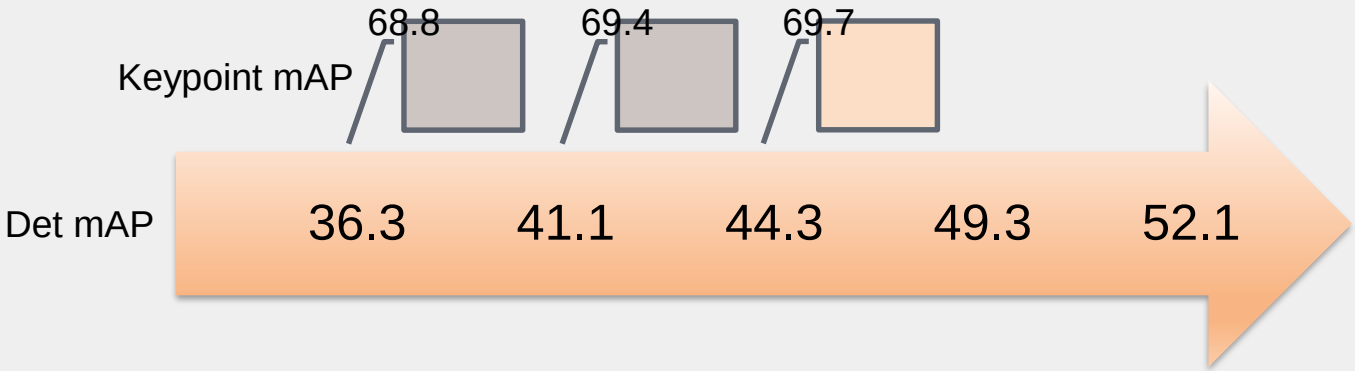
Detection Performance



Techniques & Experiments

Person Detector

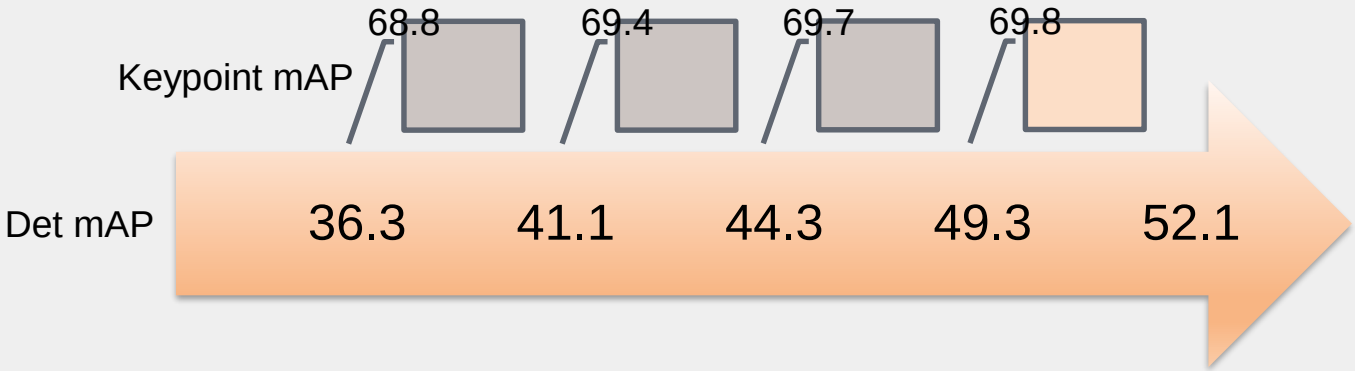
Detection Performance



Techniques & Experiments

Person Detector

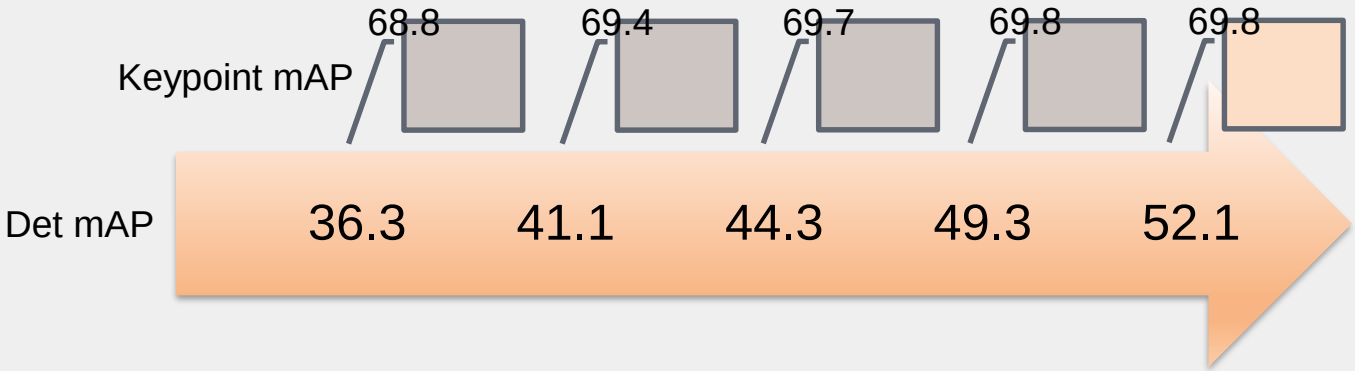
Detection Performance



Techniques & Experiments

Person Detector

Detection Performance



Techniques & Experiments

Person Detector

Detection Performance

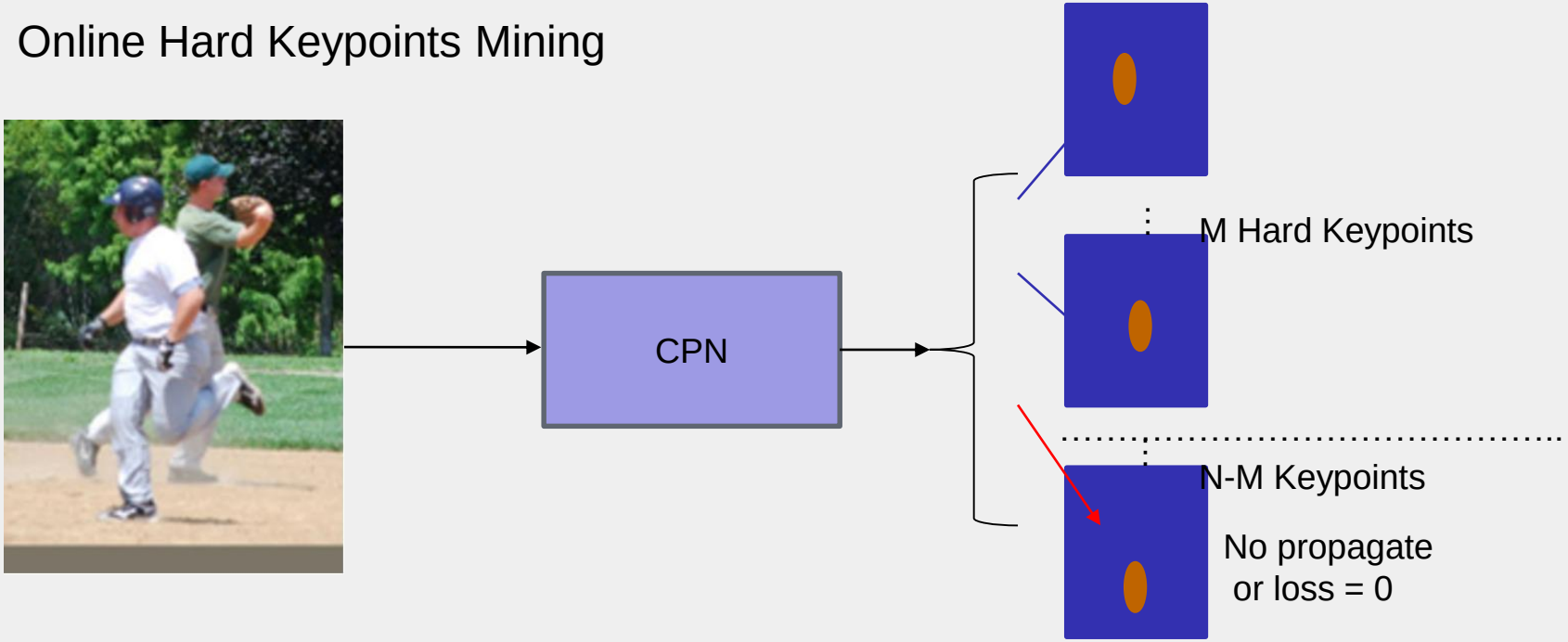
Det Methods	AP(all)	AP(H)	AR(H)	AP(OKS)
FPN-1	36.3	49.6	58.5	68.8
FPN-2	41.1	55.3	67.0	69.4
FPN-3	44.3	58.4	71.3	69.7
ensemble-1	49.3	61.4	71.8	69.8
ensemble-2	52.1	62.9	74.7	69.8

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.

Techniques & Experiments

Cascaded Pyramid Network

Online Hard Keypoints Mining



Techniques & Experiments

Cascaded Pyramid Network

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.

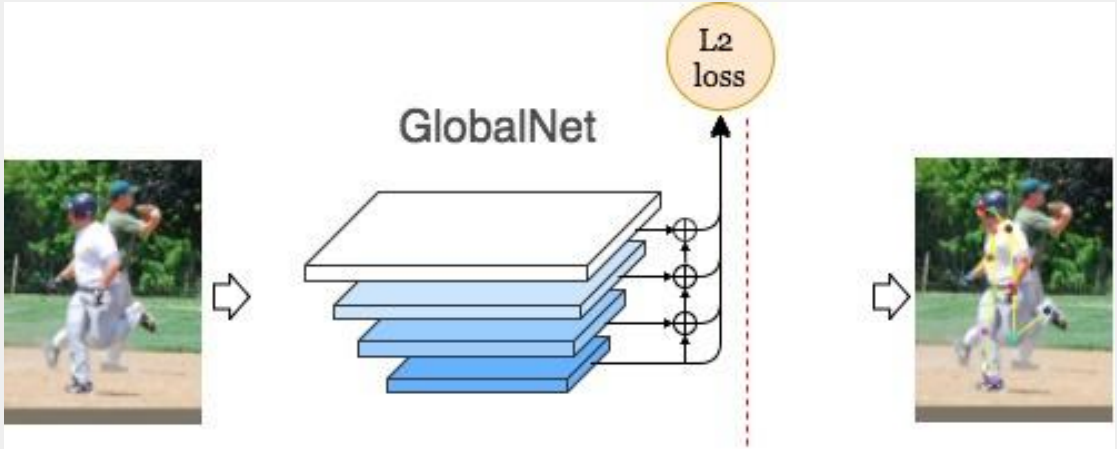
GlobalNet	RefineNet	AP(OKS)
-	L2 loss	68.2
L2 loss	L2 loss	68.6
-	L2 loss*	68.5
L2 loss	L2 loss*	69.4
L2 loss*	L2 loss*	69.1

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.

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

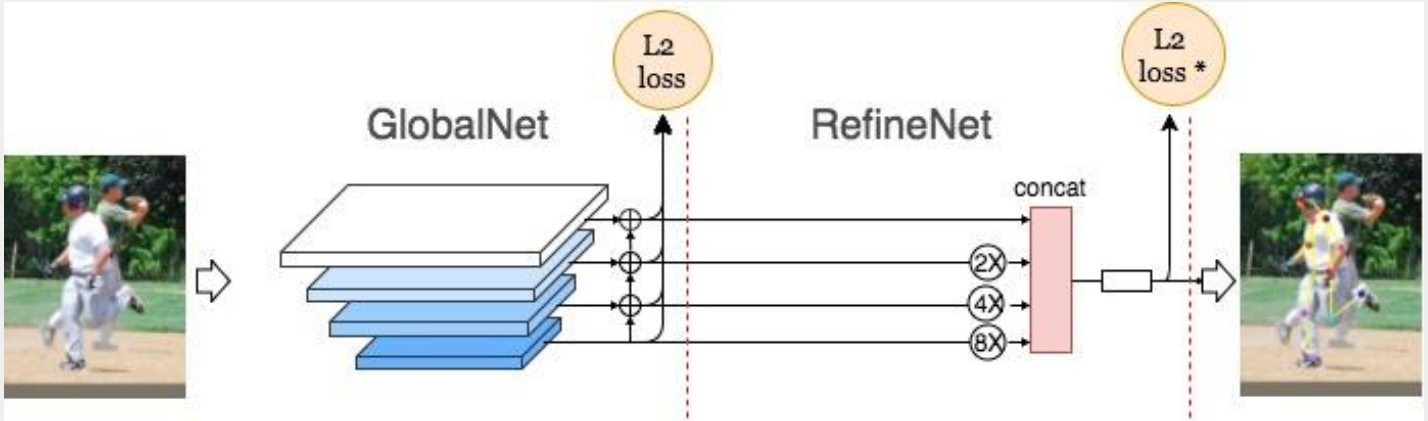


Models	AP(OKS)	FLOPs
GlobalNet only	66.6	3.90G

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

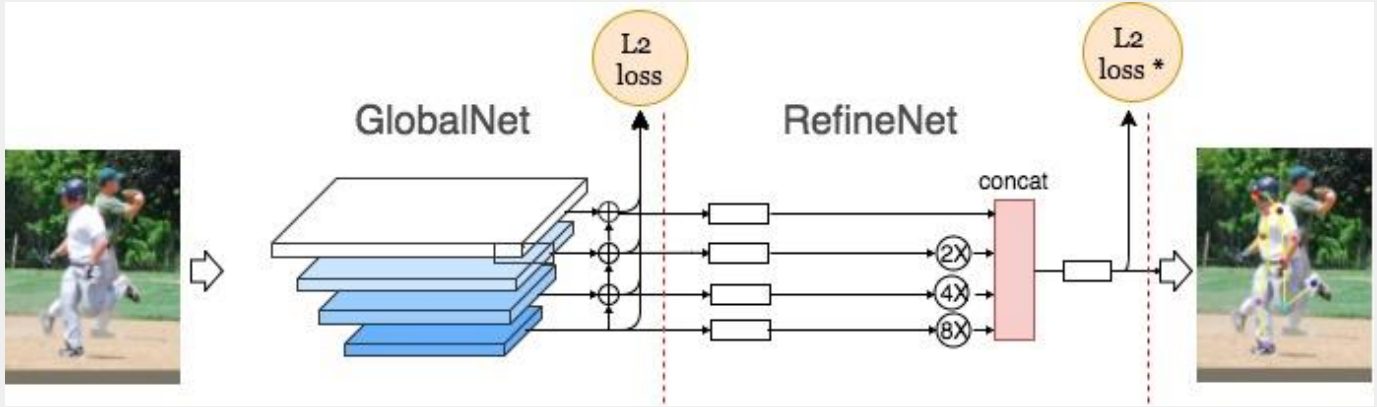


Models	AP(OKS)	FLOPs
GlobalNet only	66.6	3.90G
GlobalNet + Concat	68.5	5.87G

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

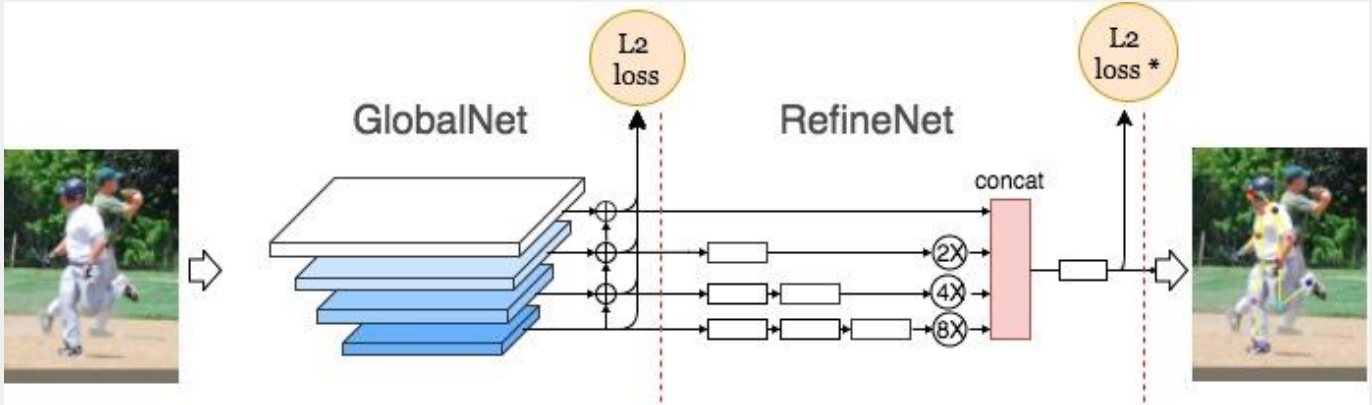


Models	AP(OKS)	FLOPs
GlobalNet only	66.6	3.90G
GlobalNet + Concat	68.5	5.87G
GlobalNet + one bottleneck +Concat	69.2	6.92G

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of 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

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

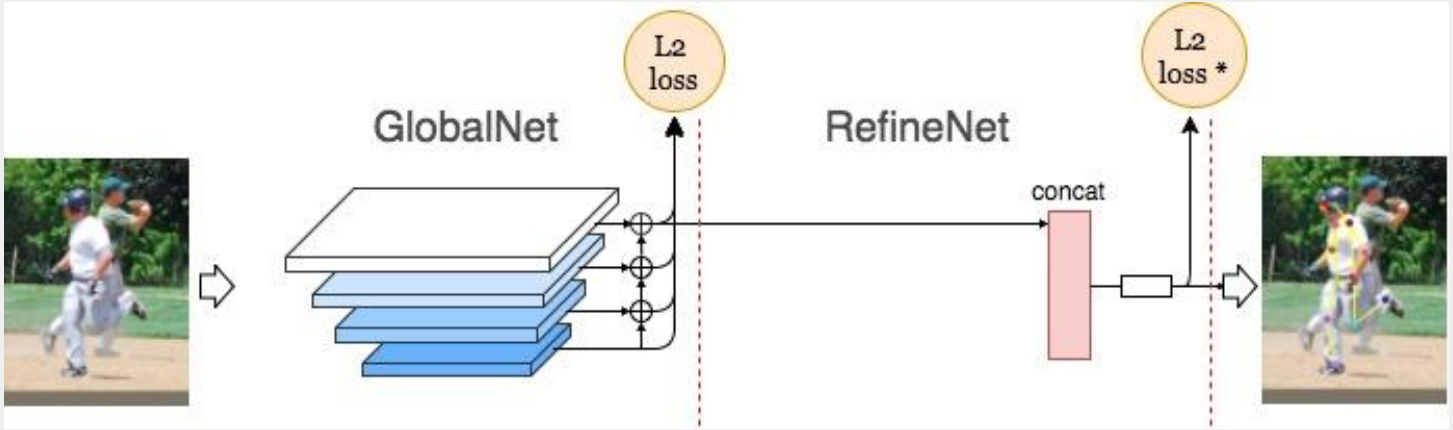
Models	AP(OKS)	FLOPs
GlobalNet only	66.6	3.90G
GlobalNet + Concat	68.5	5.87G
GlobalNet + one bottleneck +Concat	69.2	6.92G
ours (CPN)	69.4	6.20G

Table 6. Comparison of models of different design choices of RefineNet.

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

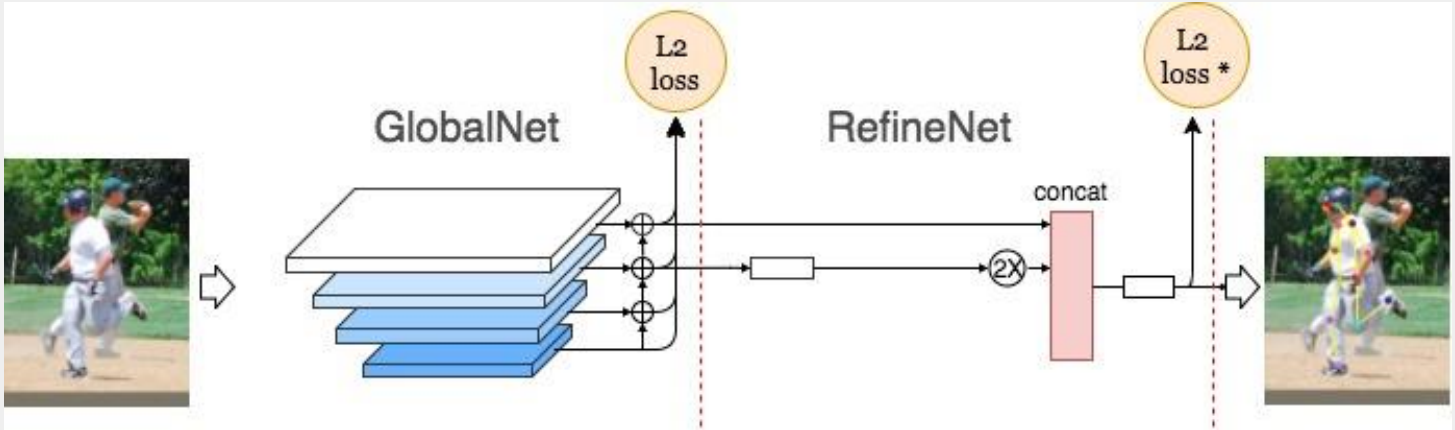


Connections	AP(OKS)	FLOPs
C_2	68.3	5.02G
$C_2 \sim C_3$	68.4	5.50G
$C_2 \sim C_4$	69.1	5.88G
$C_2 \sim C_5$	69.4	6.20G

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

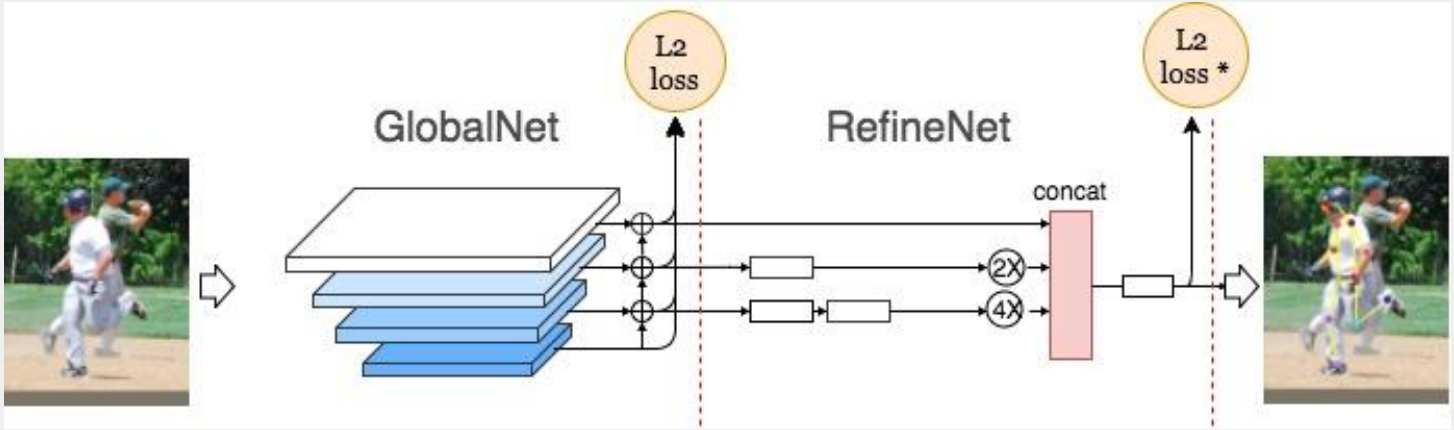


Connections	AP(OKS)	FLOPs
C_2	68.3	5.02G
$C_2 \sim C_3$	68.4	5.50G
$C_2 \sim C_4$	69.1	5.88G
$C_2 \sim C_5$	69.4	6.20G

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

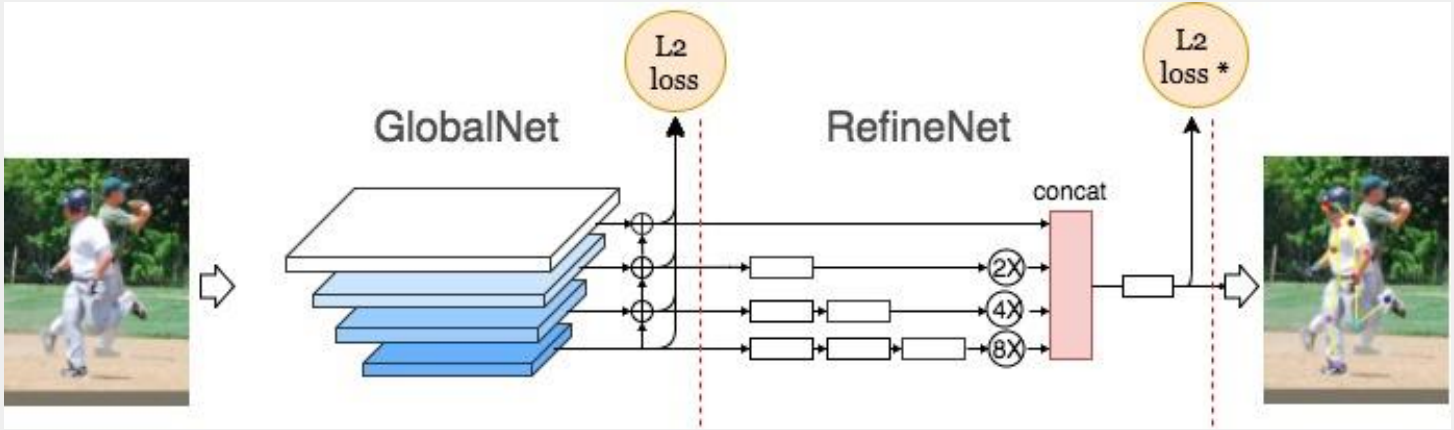


Connections	AP(OKS)	FLOPs
C_2	68.3	5.02G
$C_2 \sim C_3$	68.4	5.50G
$C_2 \sim C_4$	69.1	5.88G
$C_2 \sim C_5$	69.4	6.20G

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet



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

Techniques & Experiments

Cascaded Pyramid Network

Design Choices of RefineNet

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

Table 7. Effectiveness of intermediate connections of CPN.

Techniques & Experiments

Data Pre-processing

Models	Input Size	FLOPs	AP(OKS)
8-stage Hourglass	256 × 192	19.5G	66.9
8-stage Hourglass	256 × 256	25.9G	67.1
CPN* (ResNet-50)	256 × 192	6.2G	68.6
CPN (ResNet-50)	256 × 192	6.2G	69.4
CPN* (ResNet-50)	384 × 288	13.9G	70.6
CPN (ResNet-50)	384 × 288	13.9G	71.6

Table 8. Comparison of models of different input size. CPN* indicates CPN without online hard keypoints mining.

Techniques & Experiments

Data Augmentation (+0.4 AP)

Crop augmentation

Random scales(0.7~ 1.35)
Rotation(-45º ~ 45º)

Techniques & Experiments

Data Augmentation (+0.4 AP)

Crop augmentation

Random scales(0.7~ 1.35)
Rotation(-45º ~ 45º)

Large Batch (+0.4~0.7AP)

Techniques & Experiments

Data Augmentation (+0.4 AP)

Crop augmentation

Random scales(0.7~ 1.35)
Rotation(-45º ~ 45º)

Large Batch (+0.4~0.7AP)

Ensemble(+1.1~1.5AP in minival)

Heatmap merge

	AP% (COCO minival)	AP% (COCO test_challenge)	AP% (COCO test_dev, single_model)
Our network with all techniques	74.7	72.6	72.1

Results on MS COCO

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.

Results on MS COCO

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.

Results on MS COCO

Methods	AP - minival	AP - dev	AP - challenge
Ours (CPN)	72.7	72.1	-
Ours (CPN+)	74.5	73.0	72.1

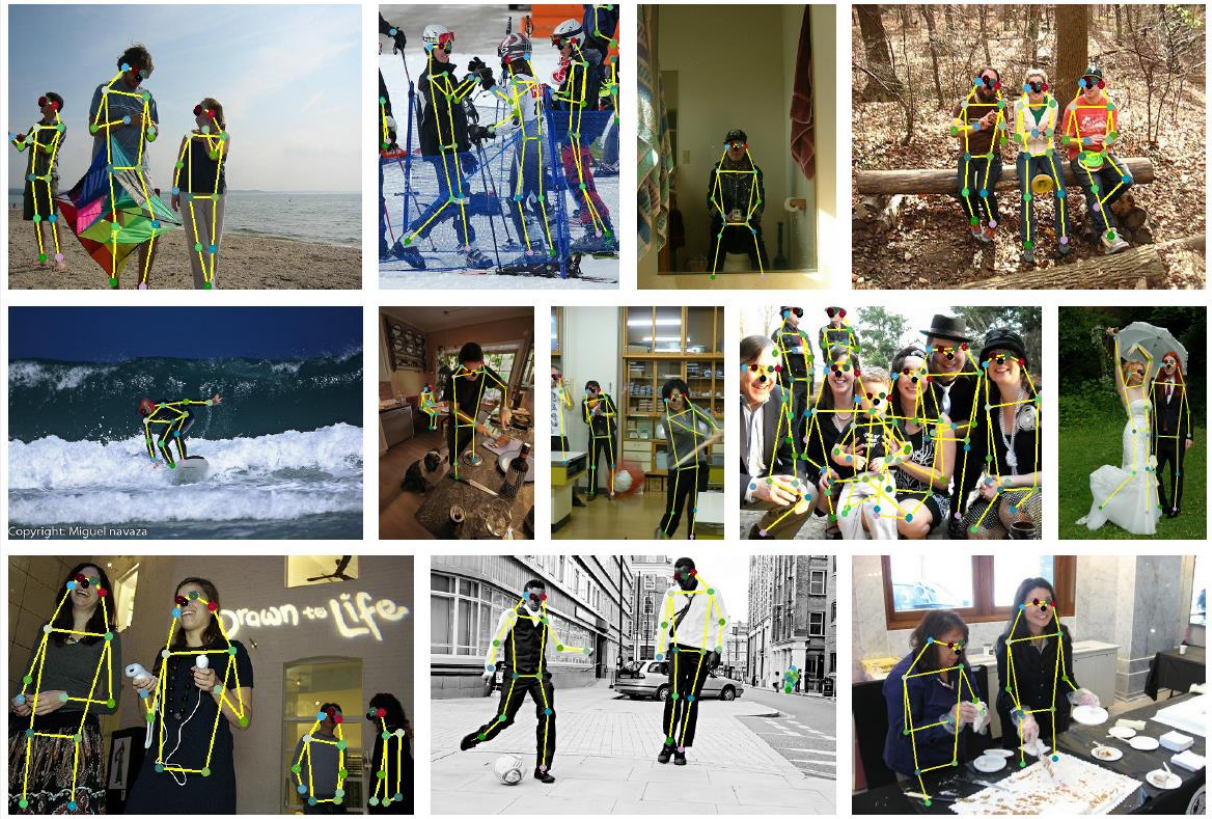
Table 11. Comparison of results on the minvival dataset and the corresponding results on test-dev or test-challenge of the COCO dataset. "+" indicates ensembled model. CPN and CPN+ in this table all use the backbone of ResNet-Inception [39] framework.

Results on PoseTrack

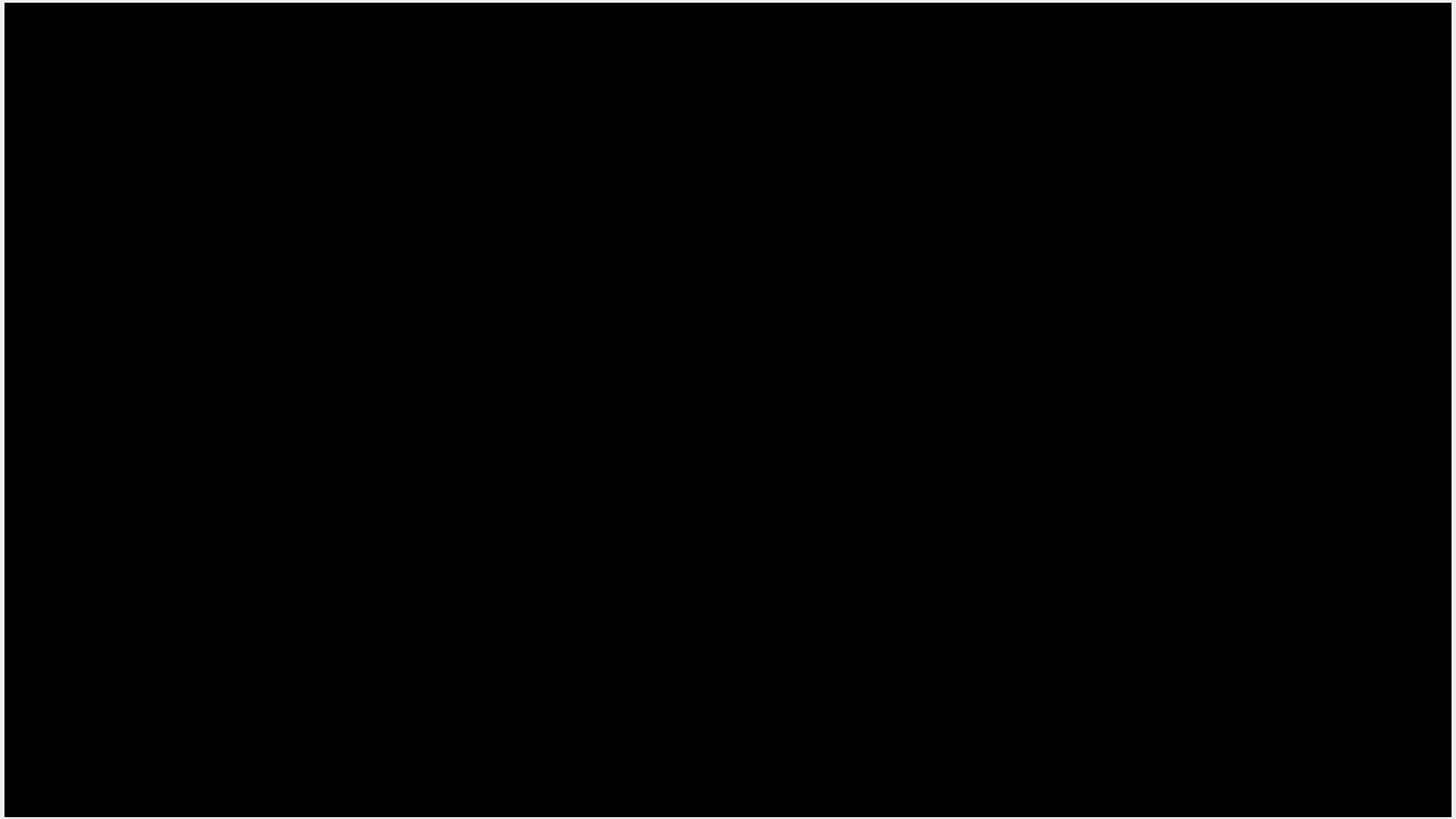
Method	AP
Our	75.5
AlphaPose	66.7
ML_Lab	70.3

Leaderboard: https://posetrack.net/leaderboard.php

Illustrative results of our method



Illustrative results of our method



Conclusion

- The two-stage network design is crucial.
  - GlobalNet: learns the overall keypoints and mainly locates the easy parts of the keypoints.
  - RefineNet: explicitly learns the hard keypoints with online hard keypoints mining.
- Intermediate supervision is important to the utility of resnet in human pose estimation.
- Large batch technique is not only applicable in object detection, but also in keypoint.

Thanks & Questions

yugang@megvii.com