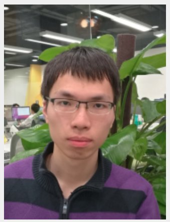
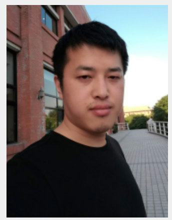
MSCOCO Keypoints Challenge 2017

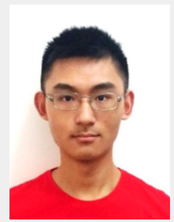
Megvii (Face++)

Face++ 旷视

Team members(Keypoints & Detection):

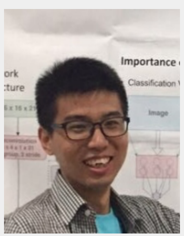

Yilun Chen*


Zhicheng Wang*


Xiangyu Peng

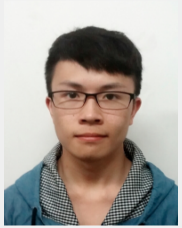

Zhiqiang Zhang


Gang Yu



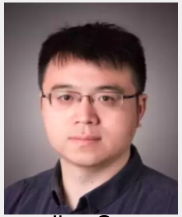
Chao Peng


Tete Xiao


Zeming Li


Xiangyu Zhang


Yuning Jiang


Jian Sun

Megvii (Face++)

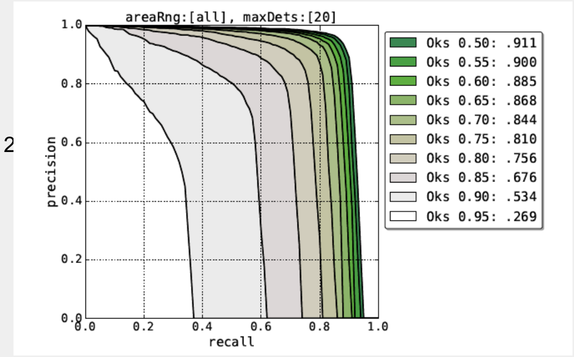
Results

- COCO 17 & 16 Keypoints

	AP	AP50	AP75	APM	APL	AR	AR50	AR75	ARM	ARL
CMU-Pose[1]	0.605	0.834	0.664	0.551	0.681	0.659	0.864	0.713	0.594	0.748
G-RMI[2]	0.598	0.81	0.651	0.567	0.667	0.664	0.865	0.712	0.618	0.726
Ours	**0.726**	**0.905**	**0.791**	**0.684**	**0.788**	**0.788**	**0.943**	**0.846**	**0.746**	**0.846**

[1] Cao, Zhe, et al. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." (2016).
[2] Papandreou, George, et al. "Towards Accurate Multi-person Pose Estimation in the Wild." (2017).

Note: [1] and [2] are evaluated on COCO 2016 test challenge dataset, while ours method is evaluated on COCO 2



Overview

- Top-down Pipeline
- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locate keypoints?
  - Our Network Architecture
- Techniques & Experiments
- Conclusion

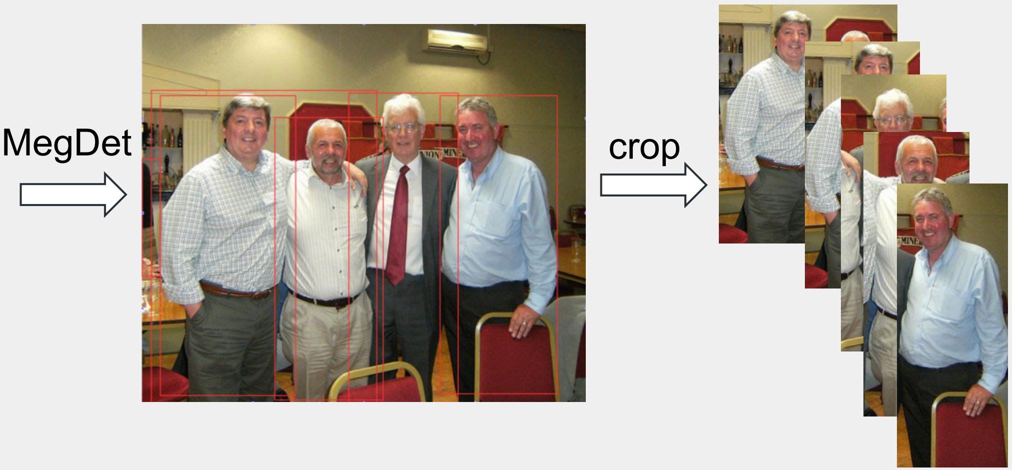
Overview

- Top-down Pipeline

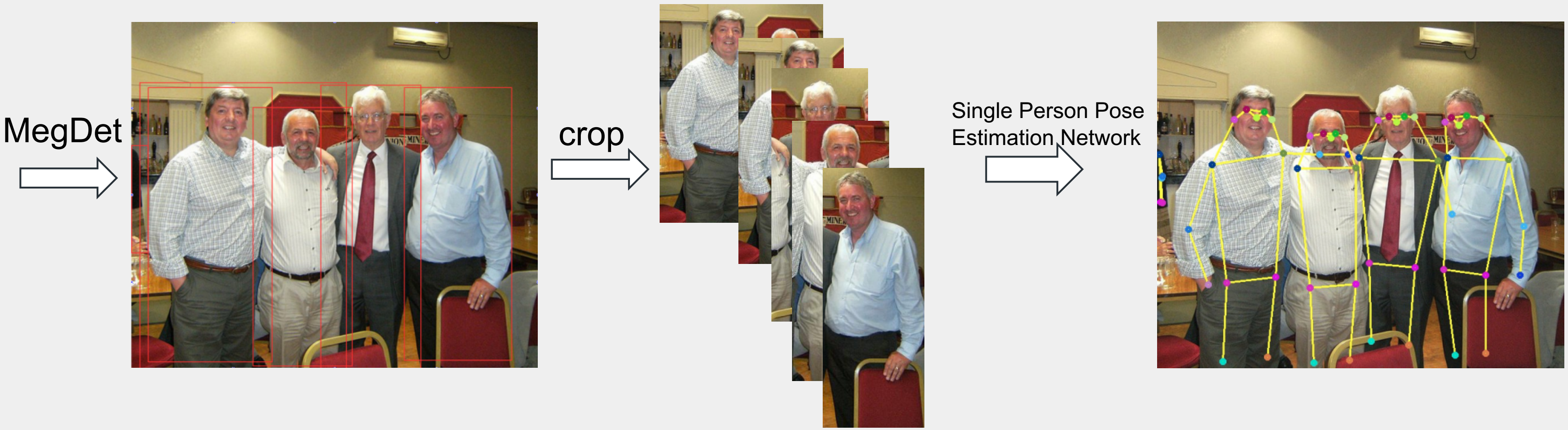
Top-Down pipeline

MegDet

Top-Down pipeline



Top-Down pipeline



Person Detector

- Our person detector is based on MegDet trained on 80-class labeled data, without specific training for person. (Human detection AP is 62.0)

Human AP(area = all)	Human AP(area = medium)	Human AP(area = large)
62.0	69.1	78.5

Overview

- Top-down Pipeline
- Network Design

Overview

- Top-down Pipeline
- Network Design
  - Is Hourglass good for COCO keypoint?

Is Hourglass good for COCO keypoint

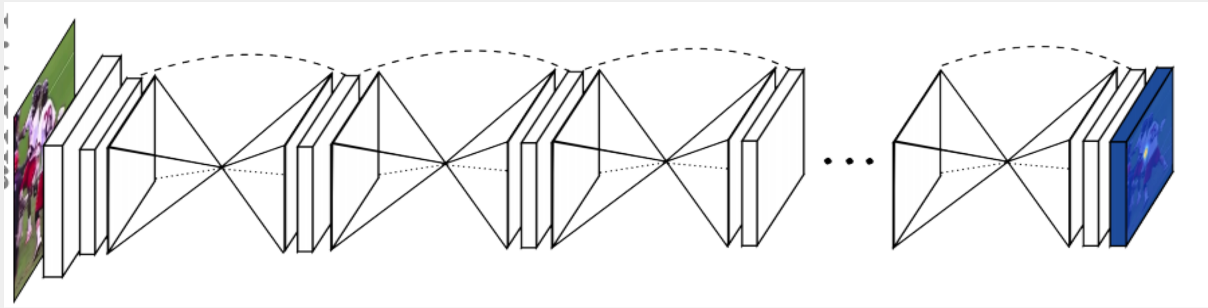
models	input size	FLOPs	param_dim	param_size	depth_conv_fc	AP
Hourglass[2] 1-stage	256x192	3.9G	3M	12MB	38	0.602
ResNet-50-FPN[1]	256x192	3.9G	24M	93MB	51	0.671

- ResNet-FPN-like[1] network works better than hourglass-like[2] network (1-stage） of the same FLOPs.

[1] Lin, Tsung-Yi, et al. "Feature Pyramid Networks for Object Detection." arXiv preprint arXiv:1612.03144 (2016).
[2] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European Conference on Computer Vision. 2016.

Is Hourglass good for COCO keypoint

Model	FLOPs	Pckh-0.5 (MPI val)	AP@OKS0.75 (COCO val)
1-stage hourglass(256*192)	3.9G	0.893	0.663
2-stage hourglass(256*192)	6.1G	0.921	0.755
3-stage hourglass(256*192)	8.3G	0.924	0.754
4-stage hourglass(256*192)	10.5G	0.924	0.752



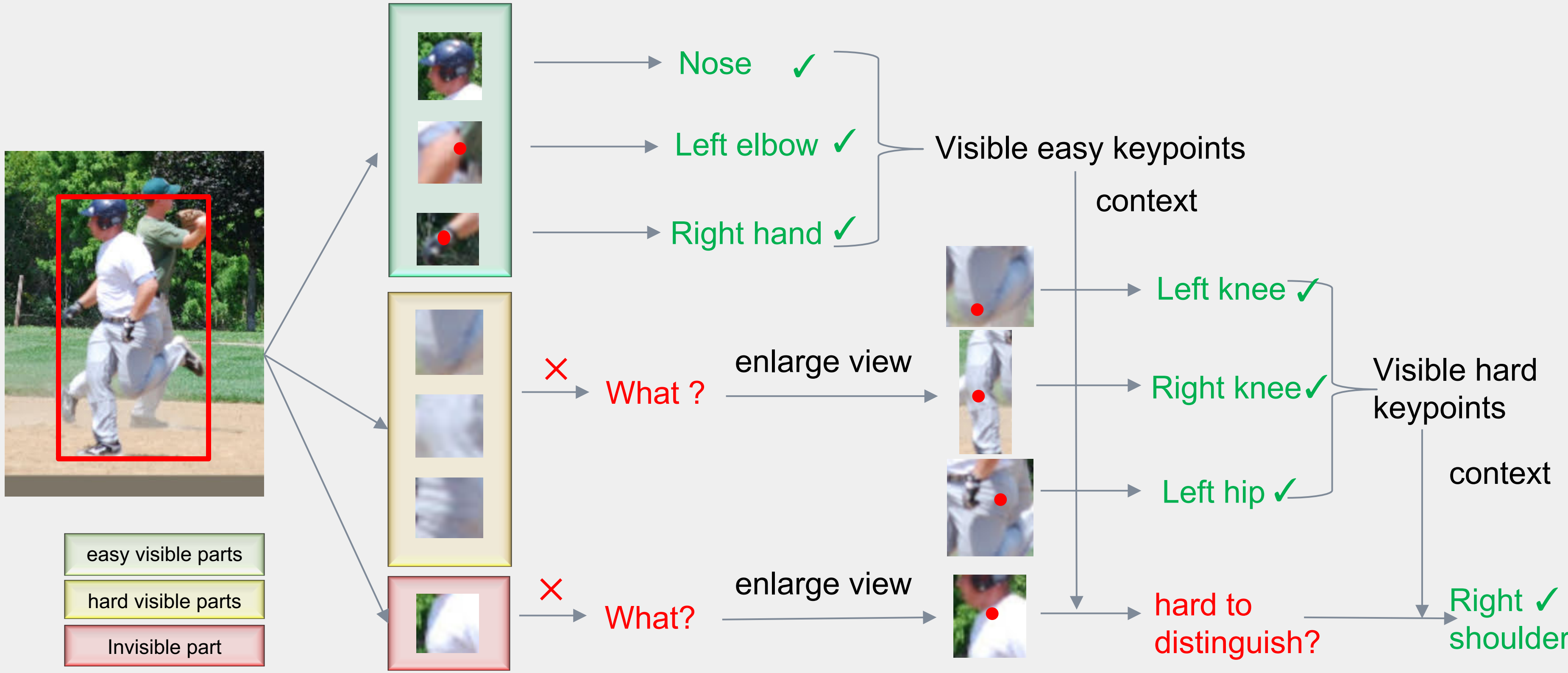
- Two stages are enough for keypoint localization for better trade-off.
- More stages (stages larger than 2) are not good at high-precision localization, for example @0.75 OKS
  - Guess: Hourglass stages harm the spatial resolution.

Overview

- Top-down Pipeline

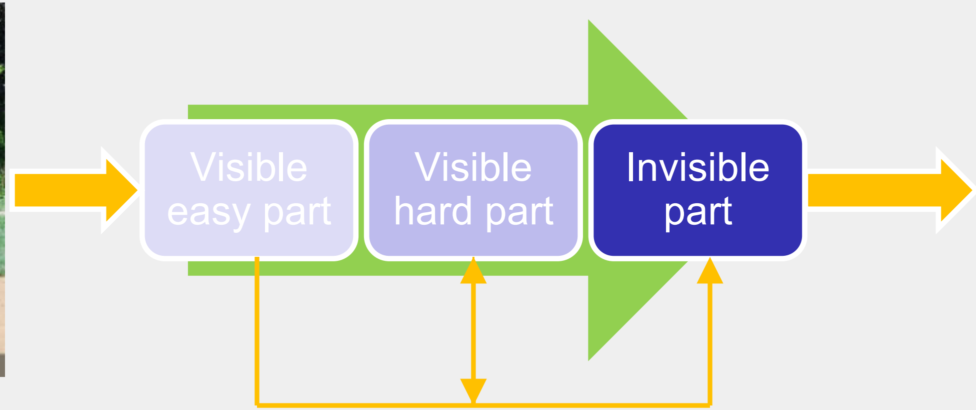
- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locates keypoints?

Motivation:
How human locate keypoints?



Network's Design Goal

Input image



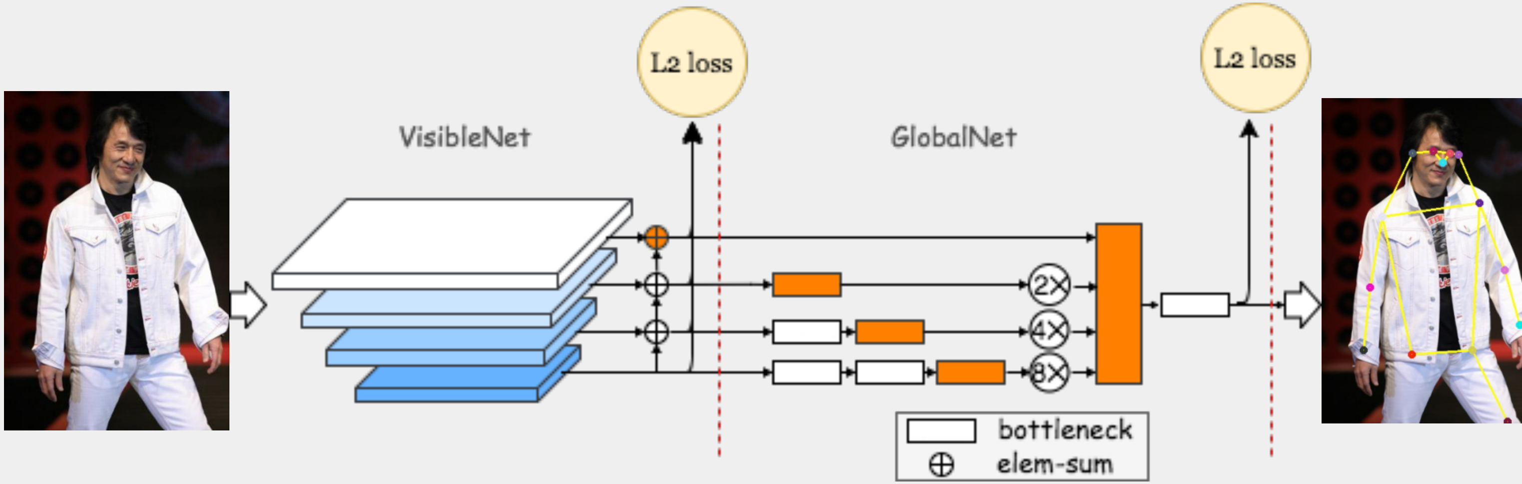
receptive view getting larger
& more context

Output image

Overview

- Top-down Pipeline
- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locate keypoints?
  - Our Network Architecture

Network Architecture



Network Design Principles:

- Follow the human perspective
  - locate visible easy parts => locate visible hard parts => locate invisible parts
- Two stages
  - VisibleNet: to locate the both the easy parts (earlier layers) and visible hard parts (deep layers)
  - GlobalNet: to locate hard parts as well

Overview

- Top-down Pipeline
- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locate keypoint?
  - Our Network Architecture
- Techniques & Experiments

Techniques & Experiments

	AP% (COCO minival)
Baseline (ResNet-50-FPN) (256x192)	67.1
Our network (ResNet-50) (256x192)	69.0
Our network (ResNet-50) (384x288)	71.0

	AP% (COCO minival)
Our network (Inception-ResNet) (384x288)	72.3
+ Large Batch	73.0

More ablation experiments on our network will come soon in our CVPR submission.

Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales(0.7~ 1.35)
  - Rotation(-45º ~ 45º)

Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales(0.7~ 1.35)
  - Rotation(-45º ~ 45º)
- Large Batch (+0.4~0.7AP)

Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales(0.7~ 1.35)
  - Rotation(-45º ~ 45º)
- Large Batch (+0.4~0.7AP)
- Segmentation supervision(+0.2~0.6AP)
  - Enhance the network's ability to distinguish the detected person from crowded scene.

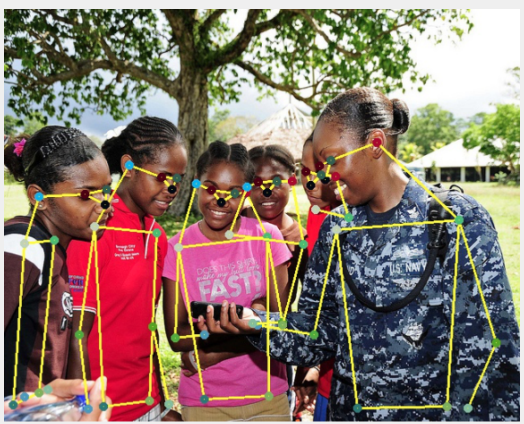
Techniques & Experiments

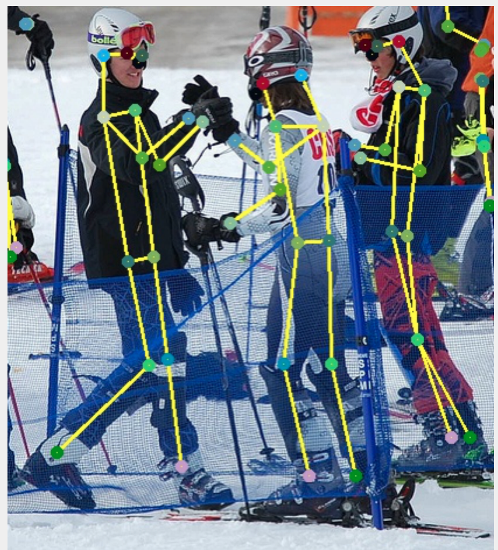
- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales(0.7~ 1.35)
  - Rotation(-45º ~ 45º)
- Large Batch (+0.4~0.7AP)
- Segmentation supervision(+0.2~0.6AP)
  - Enhance the network's ability to distinguish the detected person from crowded scene.
- Ensemble(+1.1~1.5AP)
  - Heatmap merge

	AP% (COCO minival)	AP% (COCO challenge)
Our network with all techniques	74.7	72.6

Face++ 旷视

Illustrative results of our method





Conclusion

- The two-stage network design is crucial.
  - VisibleNet: locates both the visible easy parts (earlier layers) and visible hard parts (deep layers)
  - GlobalNet: locates invisible parts

- Data augmentation is the key to enhance robustness of network, especially in CNN.

- Large batch technique is not only applicable in object detection, but also in keypoint.

- Segmentation supervision is also an universe skill in training CNN.

We are hiring!

@Beijing, @Nanjing, @Seattle

career@megvii.com

Thanks & Questions

wangzhicheng@megvii.com