International Conference on Pattern Recognition

Simple Multi-Resolution Representation Learning for Human Pose Estimation

Trung Q. Tran, Giang V. Nguyen, Daeyoung Kim

Korea Advanced Institute of Science and Technology
trungtq2019@kaist.ac.kr

January, 2021

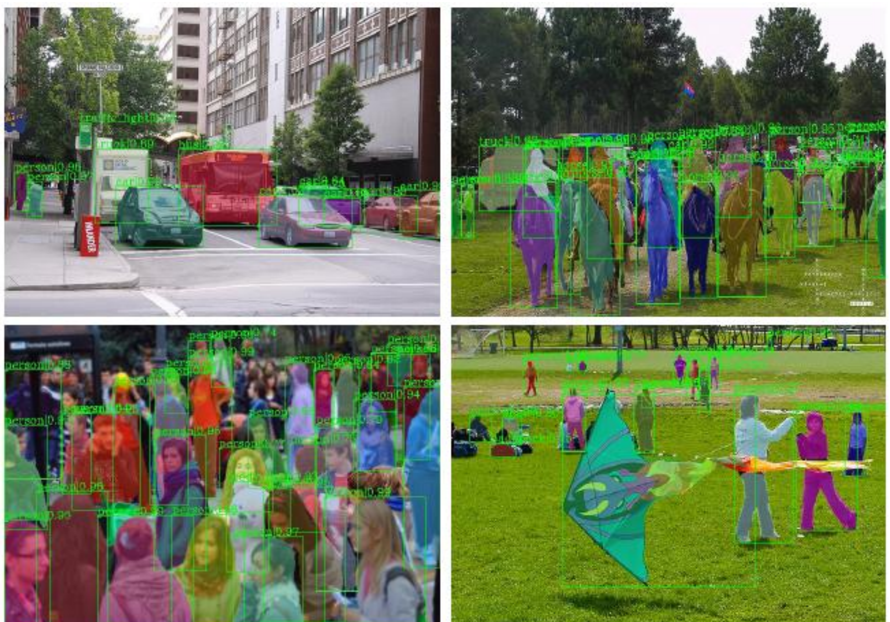
Table of Contents

Table of Contents

Computer Vision Tasks

Cat: 88%
Dog: 10%
Lion: 1%
Tiger: 0.5%
Bird: 0.5%

- Image classification
- Object detection
- Semantic segmentation
- Human pose estimation
- Etc.



Human Pose Estimation

- Important task in computer vision
- Recognizing human keypoints in given images
- Wide range of applications: movement diagnostics, self-driving vehicle, etc.

Source: https://www.danioved.com/portfolio/posenet/

Source: https://www.homecourt.ai/

Simple Pipeline for Human Pose Estimation using Heatmaps

1. **Image understanding**: generating feature maps using feature extractor
2. **Heatmap generation**: generating heatmaps using upsampling layers
3. **Human pose inference**:
   - Predicting keypoint's location using generated heatmaps
   - Connecting predicted keypoints using a pre-defined skeleton

Heatmaps: location confidence of keypoints

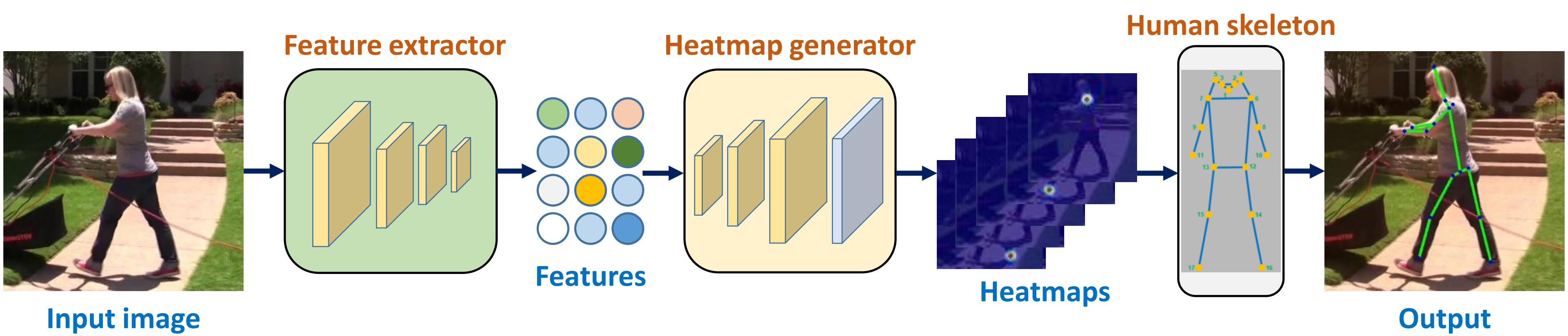
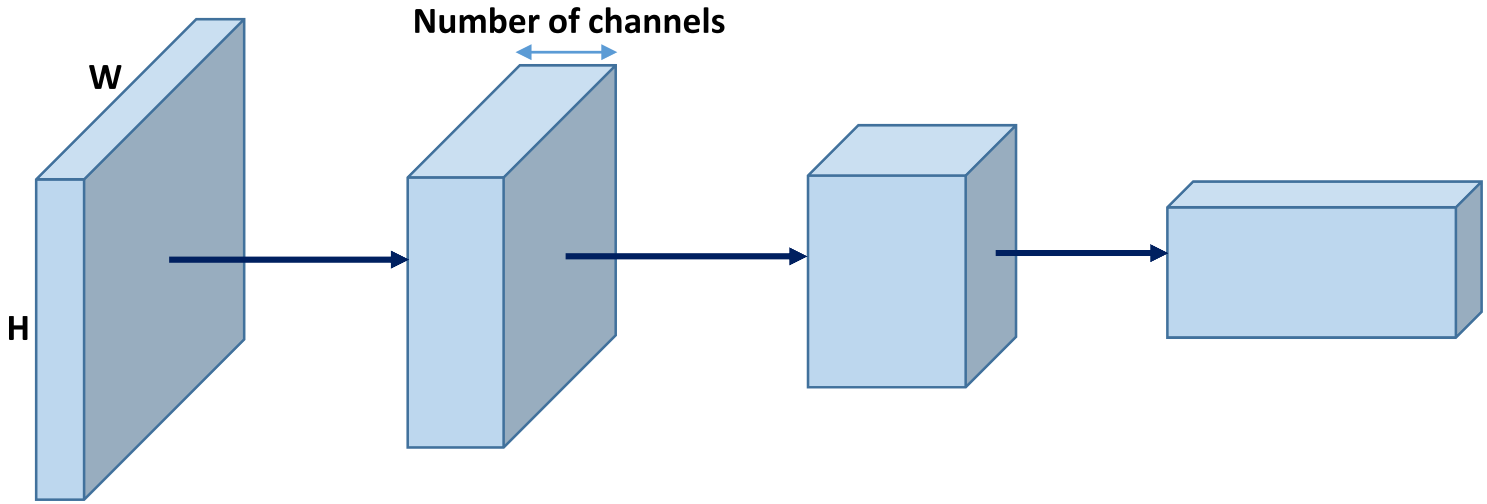


Table of Contents

Multi-resolution Learning



Encoding
the edges

Arranging
the edges

Encoding
the face

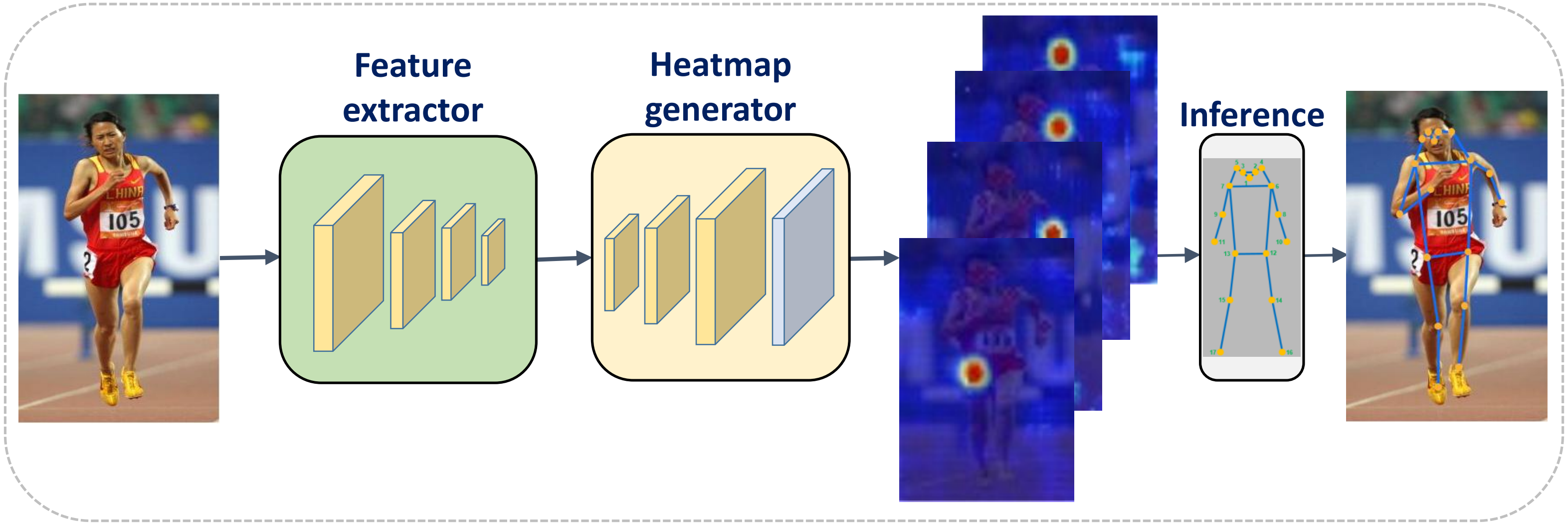
Encoding
the eyes

Observation

Left wrist is occluded

- We can infer the wrist location thanks to other keypoints such as elbow, shoulder, or even human skeleton
- The model needs not only specific features (elbow, shoulder, etc.) but also overall patterns (human skeleton, etc.)

Motivation and Approach

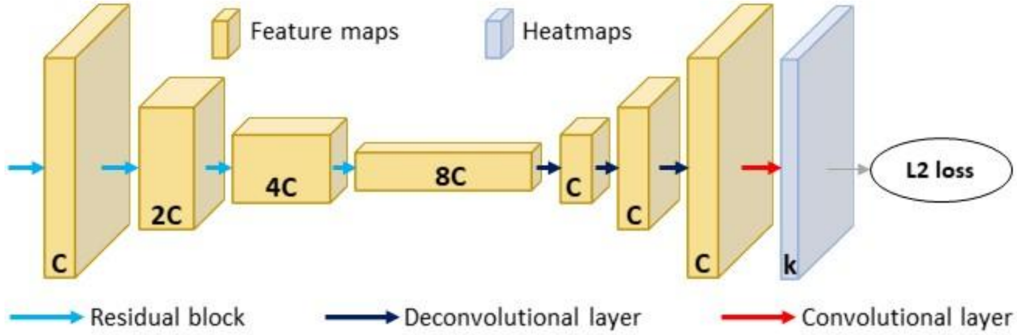


Human pose estimation using heatmaps

- Xiao et al. [1] proposed a simple architecture for human pose estimation:
  - Generating heatmaps only from lowest-resolution feature maps
  - Achieving better accuracy compared to previous methods
- **Argument**: the simple architecture could be ameliorated if it can learn the features from multiple resolutions
  - The high resolution allows capturing overall information
  - The low resolution aims to extract specific characteristics

[1]. B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018

Motivation and Approach

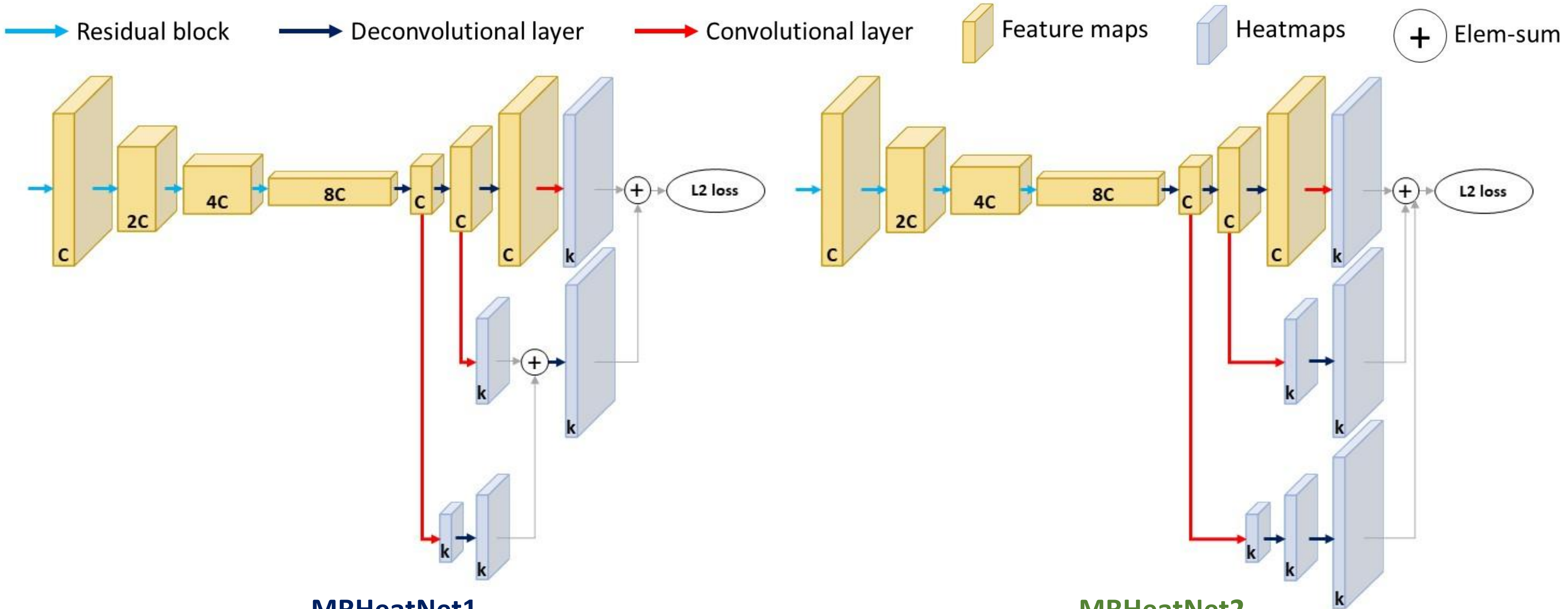


Baseline for human pose estimation using heatmaps

Two approaches:

- **Multi-resolution heatmap learning:**
  - Achieves the multi-resolution heatmaps after the lowest-resolution feature maps are obtained
  - Branches off at each resolution of the heatmap generator and adds extra layers for heatmap generation
- **Multi-resolution feature map learning:**
  - Directly learns the heatmap generation at each resolution of the feature extractor

Multi-resolution Heatmap Learning



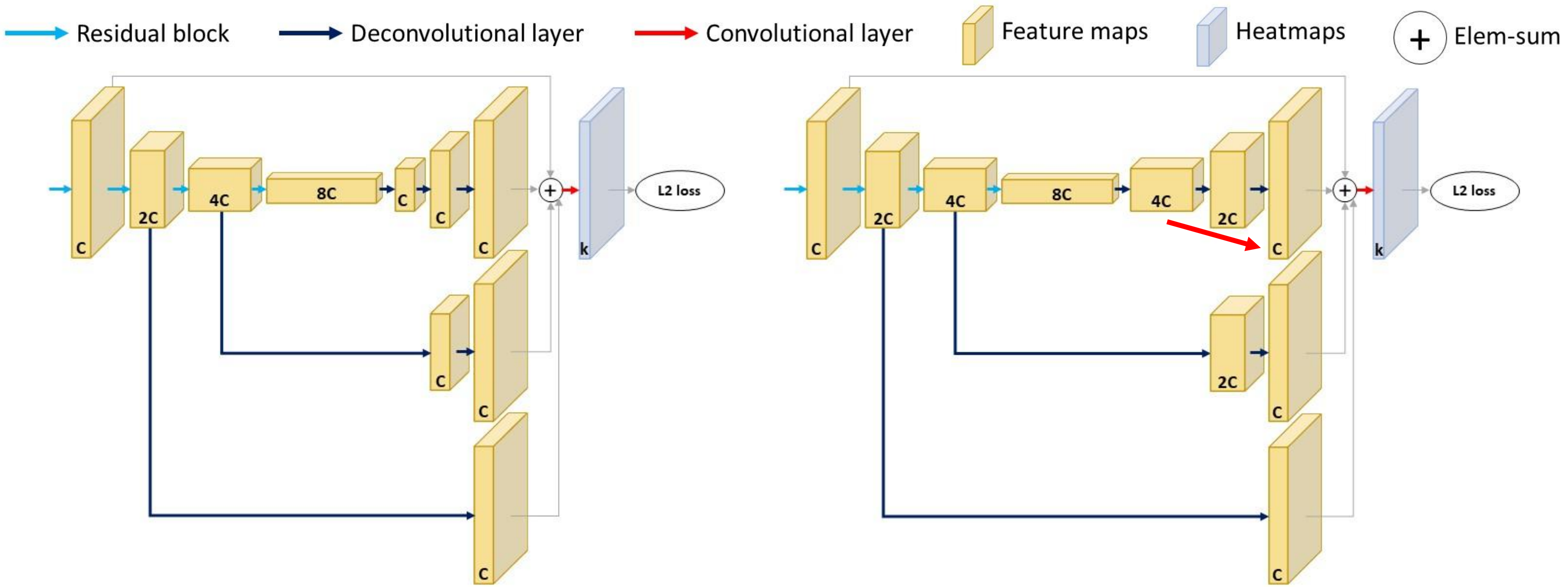
MRHeatNet1

- The lowest-resolution heatmaps are upsampled to the higher resolution (called medium resolution) and then combined with the heatmaps generated at this medium resolution
- The result of the combination is fed into a deconvolutional layer to obtain the highest-resolution heatmaps

MRHeatNet2

- The heatmaps at each resolution are upsampled to the highest-resolution heatmaps independently and then combined at the end

Multi-resolution Feature Map Learning



MRFeaNet1

- The number of output channels of deconvolutional layers is kept unchanged

MRFeaNet2

- The number of output channels is different among the deconvolutional layers
- To avoid the loss of previously learned information

Table of Contents

Evaluation Metric

- **COCO dataset:**
  - Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i [exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)]}{\sum_i [\delta(v_i > 0)]}$$

  - OKS plays the same role as the IoU in object detection → the average precision (AP) and average recall (AR) scores could be computed

- **MPII dataset:**
  - Percentage of Correct Keypoints (PCK):

$$\frac{\|y_i - \hat{y}_i\|_2}{\|y_{rhip} - y_{lsho}\|_2} \leq r$$

    - The percentage of correct detection that falls within a tolerance range which is a fraction of torso diameter
  - Percentage of Correct Keypoints with respect to head (PCKh):
    - Is almost the same as PCK except that the tolerance range is a fraction of head size

Results on COCO *val2017* dataset

Method	Backbone	Pretrain	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
8-stage Hourglass [4]	8-stage Hourglass	N	66.9	-	-	-	-	-	-	-	-	-
CPN [5]	ResNet-50	Y	68.6	-	-	-	-	-	-	-	-	-
CPN + OHKM [5]	ResNet-50	Y	69.4	-	-	-	-	-	-	-	-	-
SimpleBaseline [6]	ResNet-50	Y	70.4	88.6	78.3	67.1	77.2	76.3	92.9	83.4	72.1	82.4
MRHeatNet1	ResNet-50	Y	70.2	88.5	77.6	66.8	77.2	76.2	92.8	83.0	71.8	82.4
MRHeatNet2	ResNet-50	Y	70.3	88.5	78.0	67.2	77.0	76.4	92.9	83.1	72.1	82.4
MRFeaNet1	ResNet-50	Y	70.6	88.7	78.1	67.3	77.5	76.5	92.9	83.3	72.1	82.7
MRFeaNet2	ResNet-50	Y	70.9	88.8	78.3	67.2	78.1	76.8	93.0	83.6	72.2	83.4
SimpleBaseline [6]	ResNet-101	Y	71.4	89.3	79.3	68.1	78.1	77.1	93.4	84.0	73.0	83.2
MRFeaNet2	ResNet-101	Y	71.8	89.1	79.6	68.5	78.8	77.8	**93.5**	84.5	73.5	84.0
SimpleBaseline [6]	ResNet-152	Y	72.0	89.3	79.8	68.7	78.9	77.8	93.4	84.6	73.6	83.9
MRFeaNet2	ResNet-152	Y	**72.6**	**89.4**	**80.4**	**69.4**	**79.3**	**78.2**	93.4	**85.2**	**74.1**	**84.2**

- Our architectures outperform Hourglass and CPN
- With ResNet-50 backbone, Online Hard Keypoints Mining (OHKM) helps CPN gain the AP by 0.8 points, but still being 1.5 points lower than the AP of MRFeaNet2
- In comparison with SimpleBaseline, MRHeatNet has slightly worse performance, but MRFeaNet is superior

[4]. A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, 2016
[5]. Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018
[6]. B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision*, 2018

Results on COCO *test-dev* dataset

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Bottom-up approach: keypoint detection and grouping												
OpenPose [10]	-	-	61.8	84.9	67.5	57.1	68.2	-	-	-	-	-
Associative Embedding [11]	-	-	65.5	86.8	72.3	60.6	72.6	70.2	89.5	76.0	64.6	78.1
PersonLab [12]	ResNet-152	-	68.7	89.0	75.4	64.1	75.5	75.4	92.7	81.2	69.7	**83.0**
MultiPoseNet [13]	-	-	69.6	86.3	76.6	65.0	76.3	73.5	88.1	79.5	68.6	80.3
Top-down approach: person detection and single-person keypoint detection												
Mask-RCNN [14]	ResNet-50-FPN	-	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
G-RMI [15]	ResNet-101	353 × 257	64.9	85.5	71.3	62.3	70.0	69.7	88.7	75.5	64.4	77.1
Integral Pose Regression [16]	ResNet-101	256 × 256	67.8	88.2	74.8	63.9	74.0	-	-	-	-	-
G-RMI + extra data [15]	ResNet-101	353 × 257	68.5	87.1	75.5	65.8	73.3	73.3	90.1	79.5	68.1	80.4
SimpleBaseline [6]	ResNet-50	256 × 192	70.0	90.9	77.9	66.8	75.8	75.6	94.5	83.0	71.5	81.3
SimpleBaseline [6]	ResNet-101	256 × 192	70.9	91.1	79.3	67.9	76.7	76.7	**94.9**	84.2	72.7	82.2
SimpleBaseline [6]	ResNet-152	256 × 192	71.6	**91.2**	**80.1**	68.7	77.2	77.2	**94.9**	**85.0**	73.4	82.6
Our multi-resolution representation learning models												
MRHeatNet1	ResNet-50	256 × 192	69.7	90.8	77.8	66.6	75.4	75.4	94.4	82.9	71.3	81.1
MRHeatNet2	ResNet-50	256 × 192	69.9	90.8	78.3	66.9	75.6	75.6	94.5	83.3	71.6	81.2
MRFeaNet1	ResNet-50	256 × 192	70.1	90.7	78.4	67.0	75.9	75.8	94.3	83.3	71.7	81.3
MRFeaNet2	ResNet-50	256 × 192	70.4	90.9	78.7	67.3	76.3	76.2	94.6	83.7	72.0	81.9
MRFeaNet2	ResNet-101	256 × 192	71.2	91.0	79.6	68.2	76.9	77.0	94.7	84.5	72.9	82.5
MRFeaNet2	ResNet-152	256 × 192	**71.8**	**91.2**	**80.1**	**68.9**	**77.5**	**77.4**	94.8	84.9	**73.5**	82.8

1.9

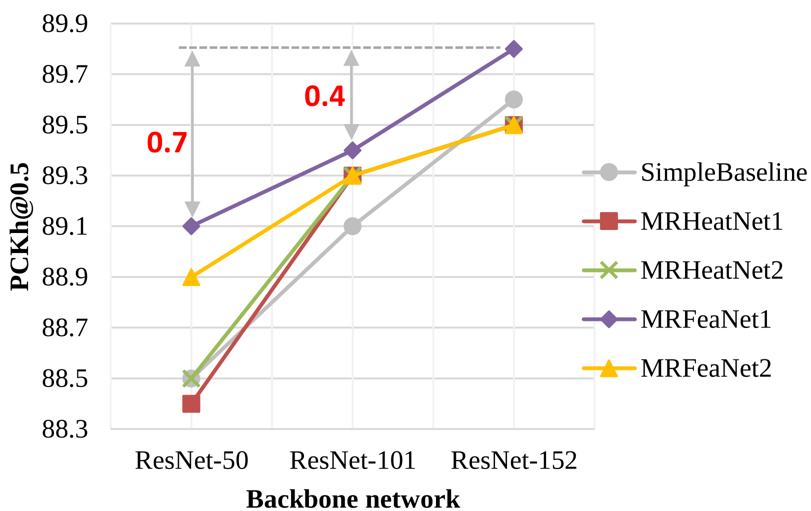
- Our architectures outperform bottom-up and top-down approaches
- In comparison with SimpleBaseline, MRFeaNet improves the AP by **0.4**, **0.3**, and **0.2** points in the case of using the ResNet-50, ResNet-101, and ResNet-152 backbone, respectively

[10]. Cao et al., 2017
[11]. Newell et al., 2017
[12]. Papandreou et al., 2018
[13]. Kocabas et al., 2018
[14]. He et al., 2017
[15]. Papandreou et al., 2017
[16]. Sun et al., 2018
[6]. Xiao et al., 2018

Results on MPII dataset

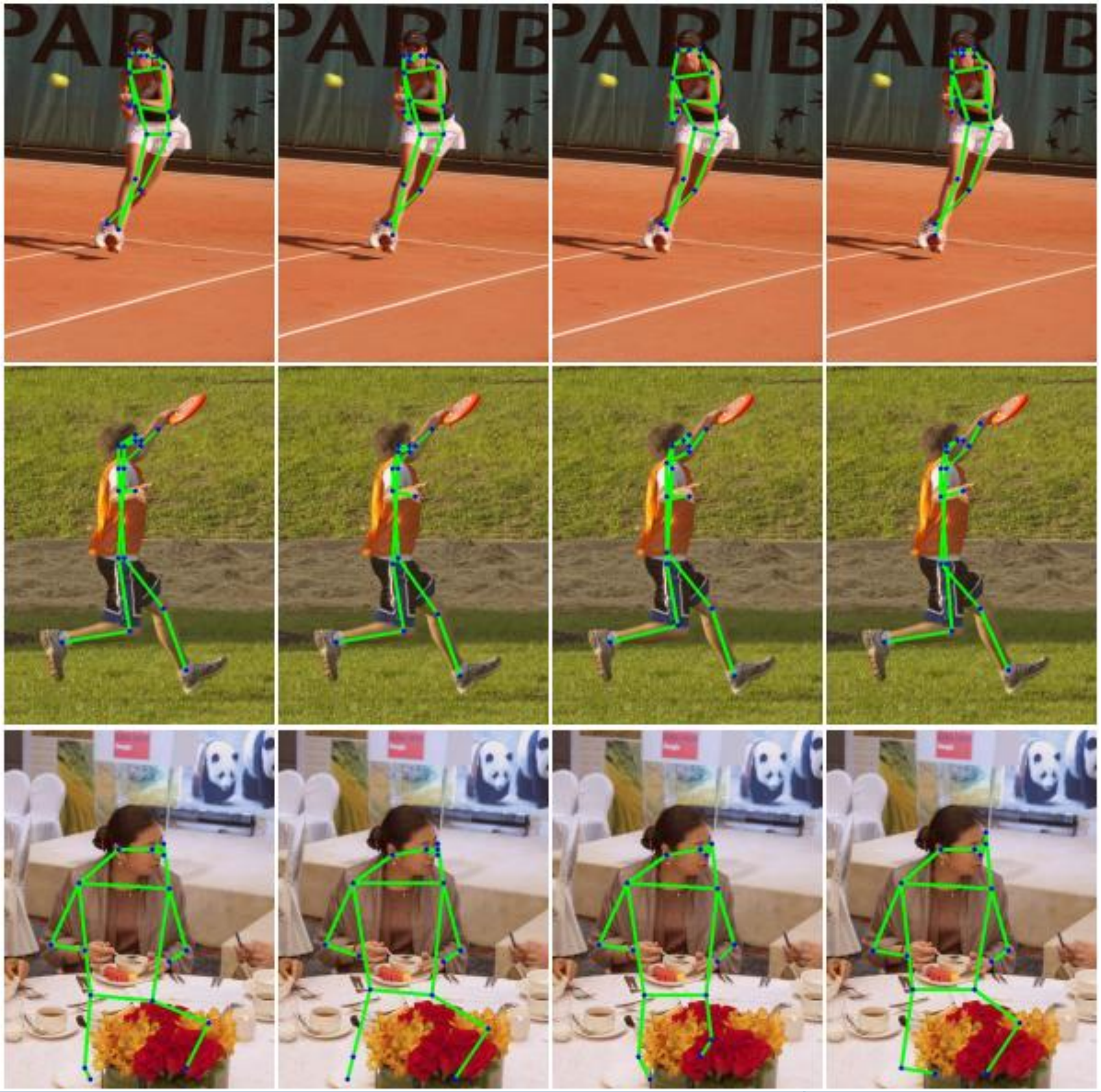
Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Pishchulin et al. [18]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson et al. [19]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al. [20]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al. [2]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu et al. [21]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin et al. [22]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al. [23]	**97.8**	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary et al. [24]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al. [25]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis et al. [26]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov et al. [27]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [3]	**97.8**	95.0	88.7	84.0	88.4	82.8	79.4	88.5
SimpleBaseline^{50} [6]	96.4	95.3	89.0	83.2	88.4	84.0	79.6	88.5
MRHeatNet1^{50}	96.7	95.2	88.9	83.8	88.1	83.6	78.6	88.4
MRHeatNet2^{50}	96.8	95.5	88.6	83.8	88.5	83.6	78.7	88.5
MRFeaNet1^{50}	96.5	95.5	89.6	84.3	88.6	84.6	80.6	89.1
MRFeaNet2^{50}	96.6	95.4	88.9	83.9	88.5	84.6	80.9	88.9
SimpleBaseline^{101} [6]	96.9	95.9	89.5	84.4	88.4	84.5	80.7	89.1
MRHeatNet1^{101}	96.7	95.7	89.7	84.4	89.1	84.7	81.4	89.3
MRHeatNet2^{101}	97.4	95.6	89.3	84.2	89.0	84.9	81.2	89.3
MRFeaNet1^{101}	96.8	95.6	89.4	84.6	89.2	85.2	81.2	89.4
MRFeaNet2^{101}	96.6	95.2	89.3	84.2	89.2	85.9	81.6	89.3
SimpleBaseline^{152} [6]	97.0	95.9	90.0	85.0	89.2	85.3	81.3	89.6
MRHeatNet1^{152}	96.8	**96.0**	90.1	84.4	88.9	85.3	81.4	89.5
MRHeatNet2^{152}	96.9	95.6	89.9	84.6	88.9	**86.0**	81.2	89.5
MRFeaNet1^{152}	97.2	95.9	**90.2**	**85.3**	**89.3**	85.4	**82.0**	**89.8**
MRFeaNet2^{152}	96.7	95.4	89.9	85.1	88.8	85.7	81.8	89.5

- Our architectures outperform numerous previous methods
- MRFeaNet1 gains PCKh@0.5 score by **0.6**, **0.3** and **0.2** points compared to SimpleBaseline in the case of using the ResNet-50, ResNet-101, and ResNet-152 backbone, respectively
- The performance could be improved if using the larger backbone network



Qualitative results on COCO dataset

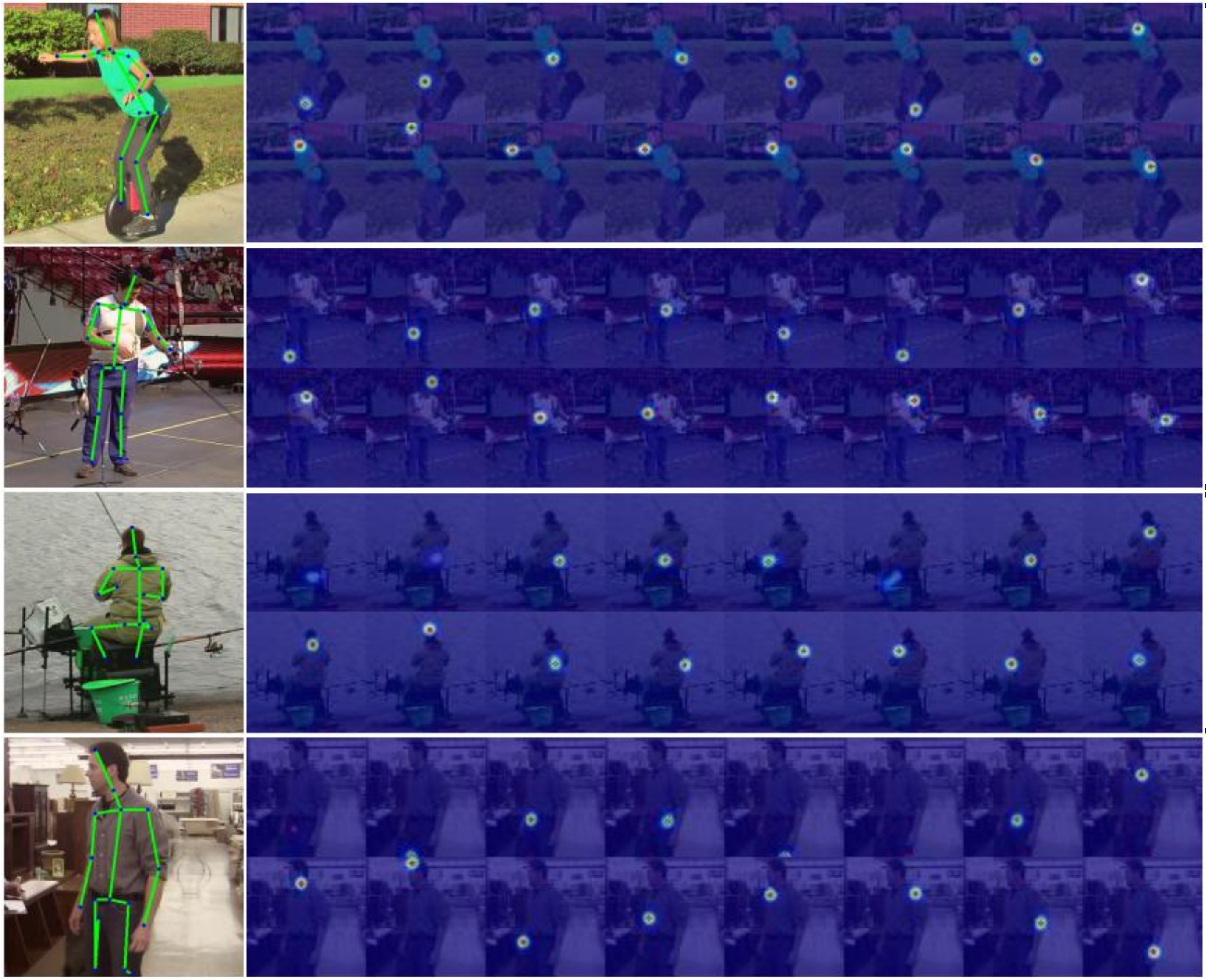
MRHeatNet1 | MRHeatNet2 | MRFeaNet1 | MRFeaNet2



- The case of occluded keypoints
- MRFeaNet still relatively precisely predicts the human keypoints

- Both legs of the woman are hidden under the table
- Our models can make their opinion

Qualitative results on MPII dataset



All keypoints are predicted with high confidence

Right leg and left ankle are occluded → the prediction has low confidence

Two ankles are not displayed → the prediction has very low confidence

Table of Contents

Conclusion and Future Work

- We introduce two novel approaches for multi-resolution representation learning:
  - The first approach reconciles a multi-resolution representation learning strategy with the heatmap generator where the **heatmaps are generated at each resolution of the deconvolutional layers**
  - The second approach achieves the **heatmap generation from each resolution of the feature extractor**

- Our architectures are simple yet effective, and experiments show the **superiority** of our methods over numerous methods

- Our approaches **could be applied to other tasks** which have the architecture of encoder (feature extractor) and decoder (specific tasks) such as image captioning or image segmentation

Thank you!