# *HRNet*

## Deep High-Resolution Representation Learning for Human Pose Estimation

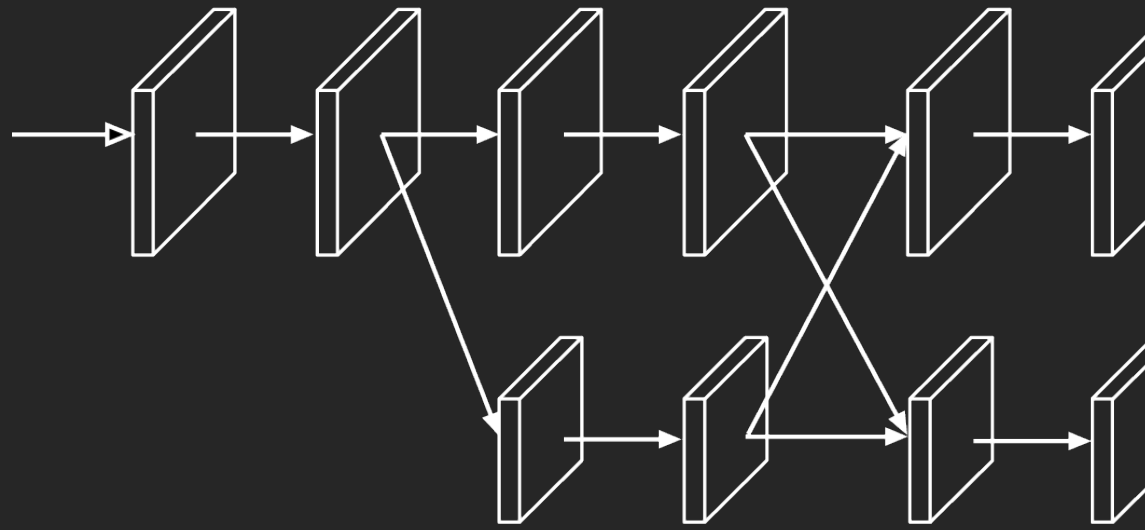



**Image Processing Team | Detection**

Byunghyun Kim, Chanhyeok Lee, Eungi Hong, Jongsik Ahn,
Hyeonjin Kim, Jaewan Park, Chungchun Hyun, Seonok Kim(🙋‍♀️)



CVPR 2019

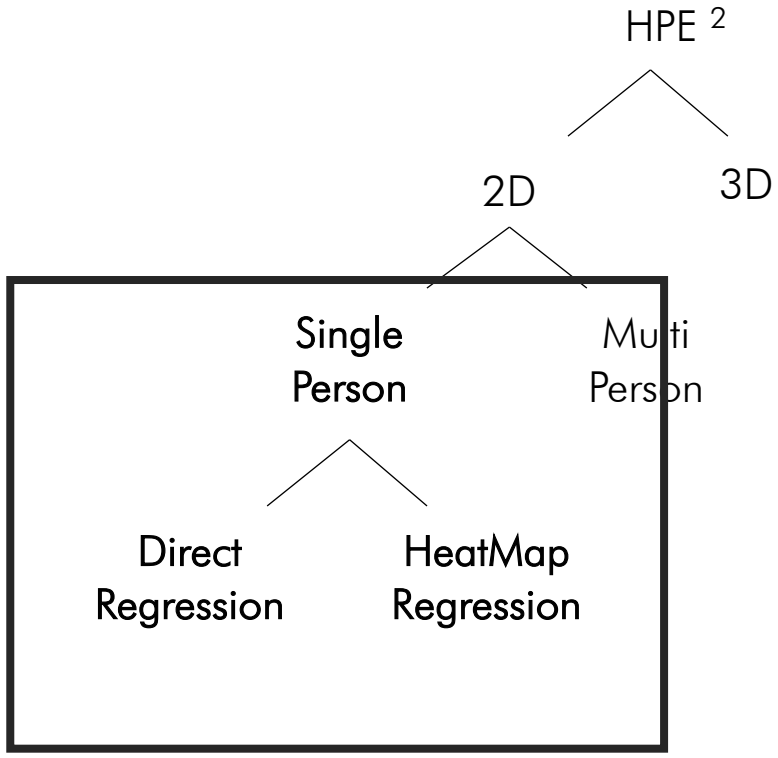


Authors: Ke Sun[1,2], Bin Xiao[2], Dong Liu[1], Jingdong Wang[2],
[1]University of Science and Technology of China, [2]Microsoft Research Asia

# Introduction

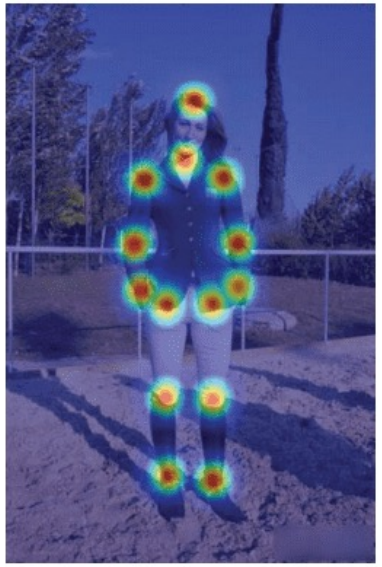
# Pose Estimation

- A computer vision task that represents the orientation of a person in a graphical format.
- Widely applied to predict a person's body parts or joint position.




## Direct Regression

regresses the key body points directly from the feature maps. [3]

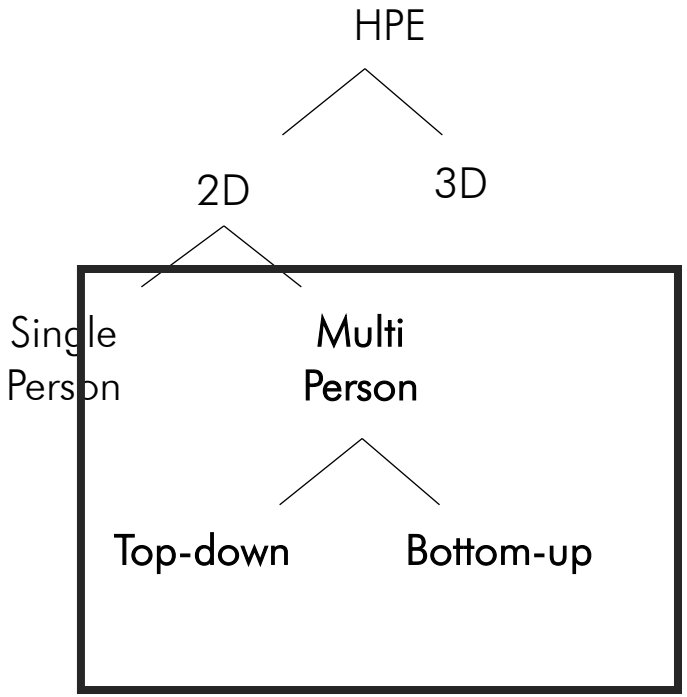


## HeatMap Regression

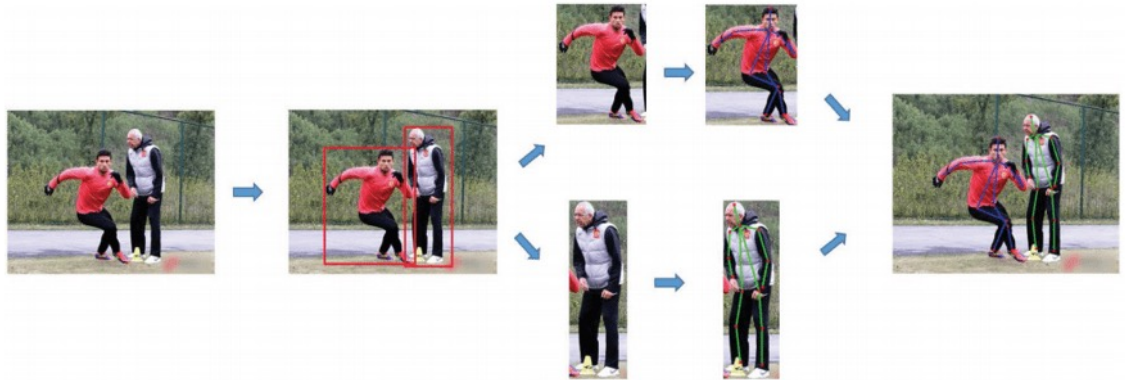
estimates the probability of the existence of a key point in each pixel of the image.

A Comprehensive Guide on Human Pose Estimation (Analytics Vidhya)
Human Pose Estimation (Open DMQA Seminar)
Human Pose Estimation with Deep Learning (Neuralet), Deep Learning Based 2D Human Pose Estimation: A Survey (TUP, 2019)

# Pose Estimation




**Top-down** is detecting all individuals in a given image using a human detector module.

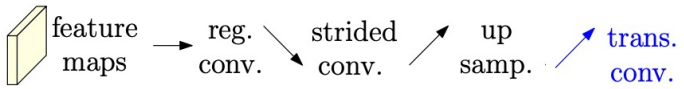


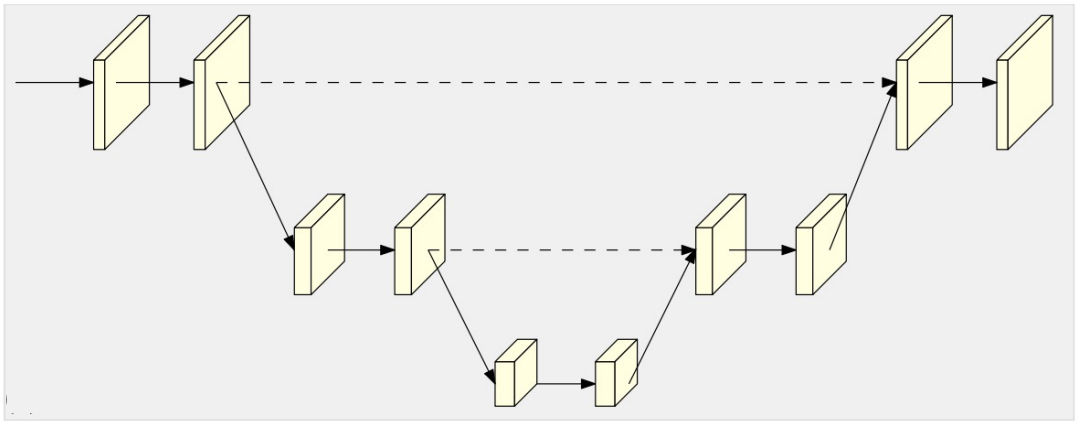
**Bottom-up** is detecting all key-points (body parts) in an instance agnostic manner and then associating key-points to build a human instance


Deep Learning Based 2D Human Pose Estimation: A Survey (2019)

# Previous Methods

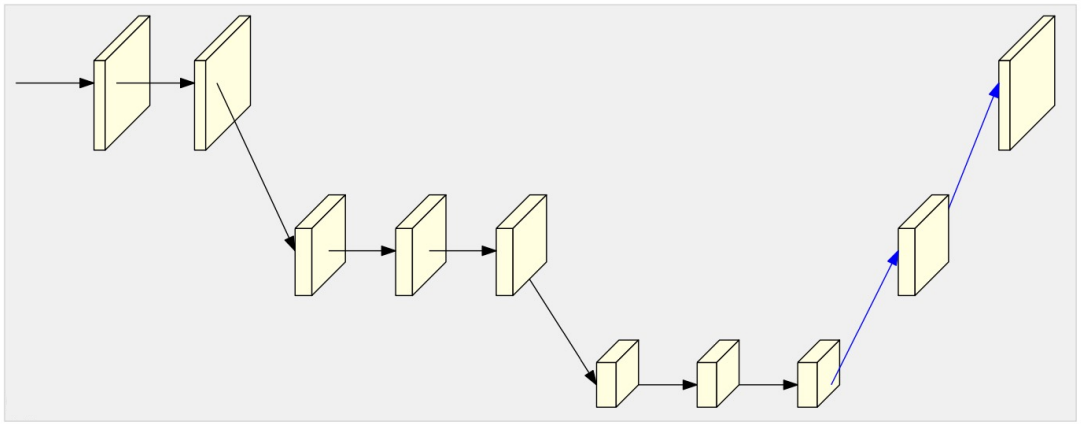
- Most existing methods pass the input through a network, typically consisting of high-to-low resolution sub-networks that are connected in series, and then raise the resolution.



## Stacked hourglass

recovers the high resolution through a symmetric low-to-high process.



## Simple Baseline

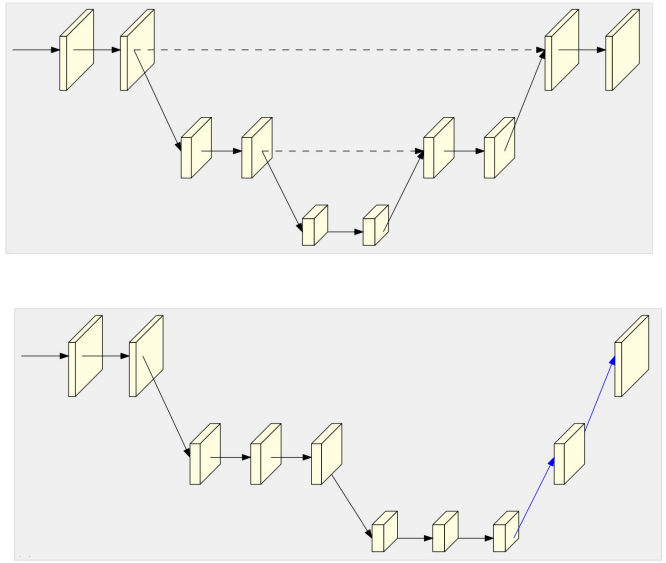
uses transposed convolutions for low-to-high processing.

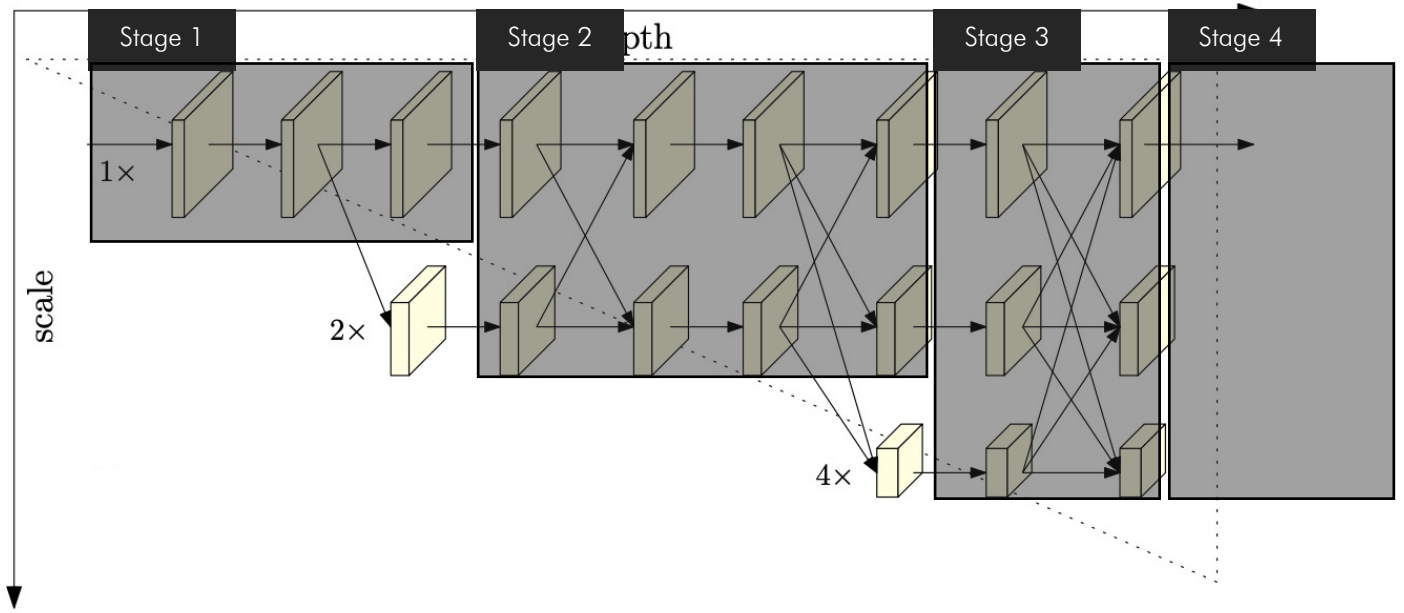
Stacked Hourglass Networks for Human Pose Estimation (2016)
Simple Baselines for Human Pose Estimation and Tracking (ECCV 2018)

# Approach

# HRNet

- HRNet starts from a high-resolution subnetwork as the first stage, gradually add high-to-low resolution subnetworks one by one to form more stages and connect the mutli-resolution subnetworks in parallel.






## Previous Methods

typically consist of high-to-low resolution sub-networks that are connected in series, and then raise the resolution.

## Proposed Method

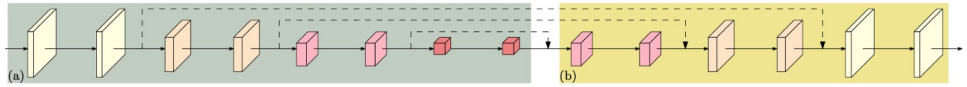
consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion)

# Parallel Multi-resolution Subnetworks

- Parallel Subnetworks start from a high-resolution subnetwork as the first stage, gradually add high-to-low resolution subnetworks one by one, forming new stages, and connect the multi-resolution subnetworks in parallel.
- $N_{sr}$ : the subnetwork, $s$ : $s$th stage, $r$ : the resolution index

## Sequential Subnetworks

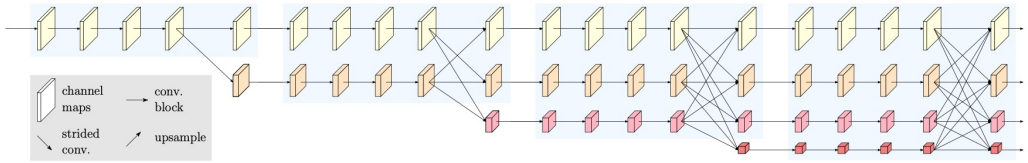
Existing networks for pose estimation are built by connecting high-to-low resolution subnetworks in series.




$$\mathcal{N}_{11} \rightarrow \mathcal{N}_{22} \rightarrow \mathcal{N}_{33} \rightarrow \mathcal{N}_{44}.$$

## Parallel Subnetworks

The resolutions for the parallel sub-networks of a later stage consists of the resolutions from the previous stage, and an extra lower one.

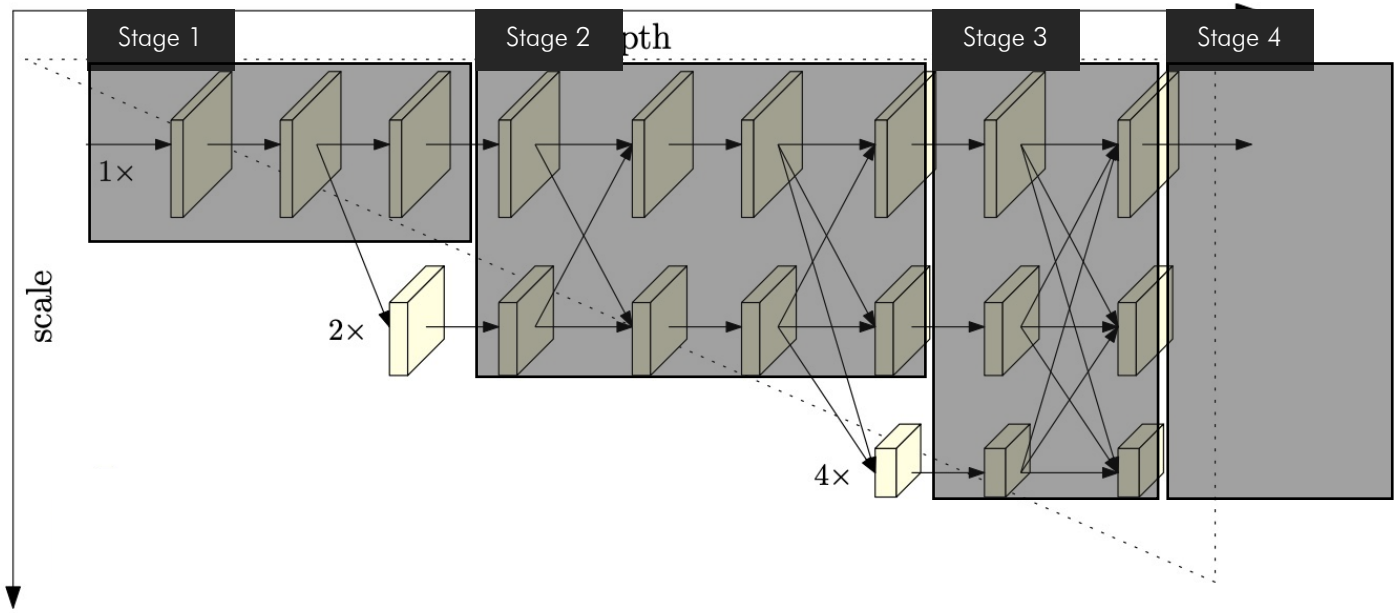



$$\begin{array}{ccccccc} \mathcal{N}_{11} & \rightarrow & \mathcal{N}_{21} & \rightarrow & \mathcal{N}_{31} & \rightarrow & \mathcal{N}_{41} \\ & \searrow & \mathcal{N}_{22} & \rightarrow & \mathcal{N}_{32} & \rightarrow & \mathcal{N}_{42} \\ & & & \searrow & \mathcal{N}_{33} & \rightarrow & \mathcal{N}_{43} \\ & & & & & \searrow & \mathcal{N}_{44}. \end{array}$$

High-Resolution Representations for Labeling Pixels and Regions (2019)

# Repeated Multi-scale Fusion

- HRNet contains four stages with four parallel subnetworks, whose resolution is gradually decreased to a half and accordingly the width (the number of channels) is increased to the double.
- The resolution is $\frac{1}{2^{r-1}}$ of the resolution of the first subnetwork.

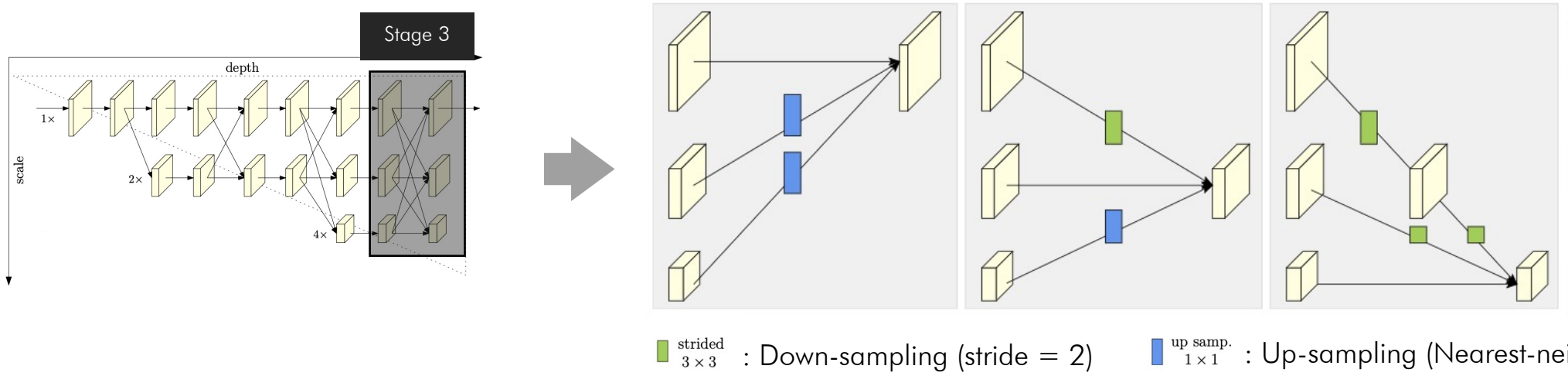



## The Stages of HRNet

- There are four stages. The 1st stage consists of high-resolution convolutions.
- The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks.

# Repeated Multi-scale Fusion

- Exchange units: each subnetwork repeatedly receives the information from other parallel subnetworks.

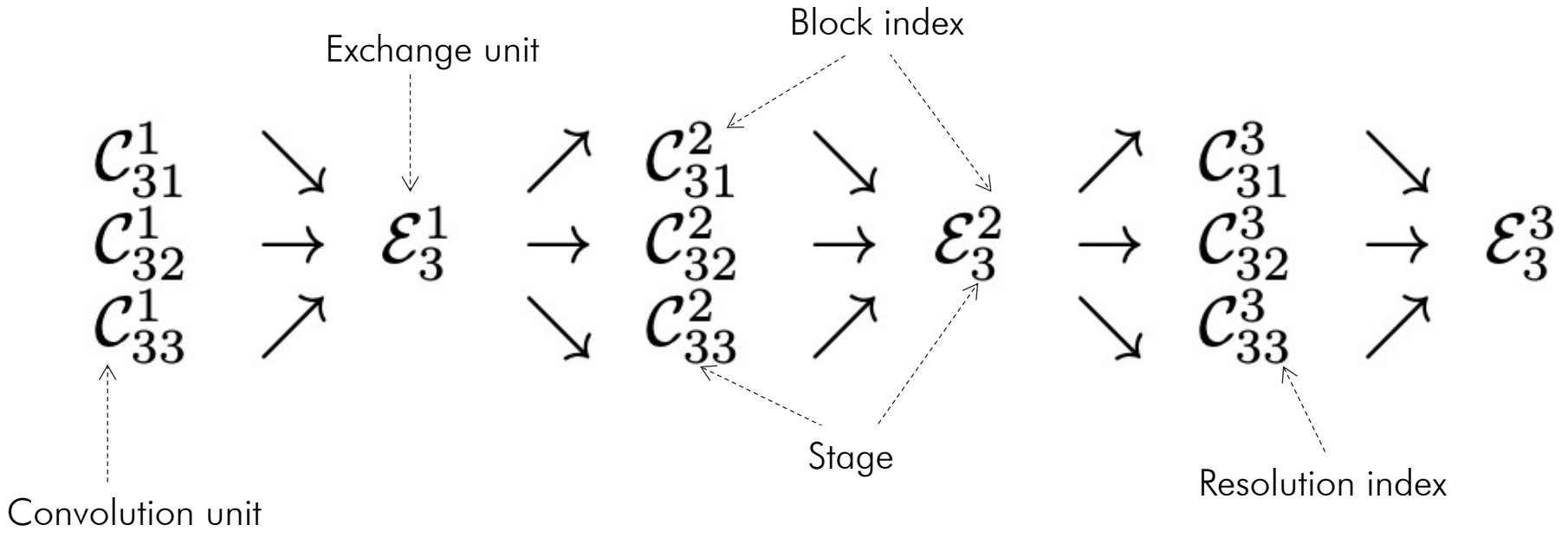



- Up-sampling or down-sampling $\mathrm{X}_i$ from resolution $i$ to resolution $k$: $\alpha(\mathrm{X}_i, k)$
- If : $i = k$ , $\alpha(\mathrm{X}_i, k) = \mathrm{X}_i$.
- Each output is an aggregation of the input maps: $\Upsilon_k = \sum_{i=1}^{s} \alpha(\mathrm{X}_i, k)$
- The exchange unit across stages has an extra output map: $\Upsilon_{s+1} = \alpha(\Upsilon_s, s+1)$

- Inputs : $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_s\}$
- Outputs : $\{\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_s\}$
- $s$th stage : $s$

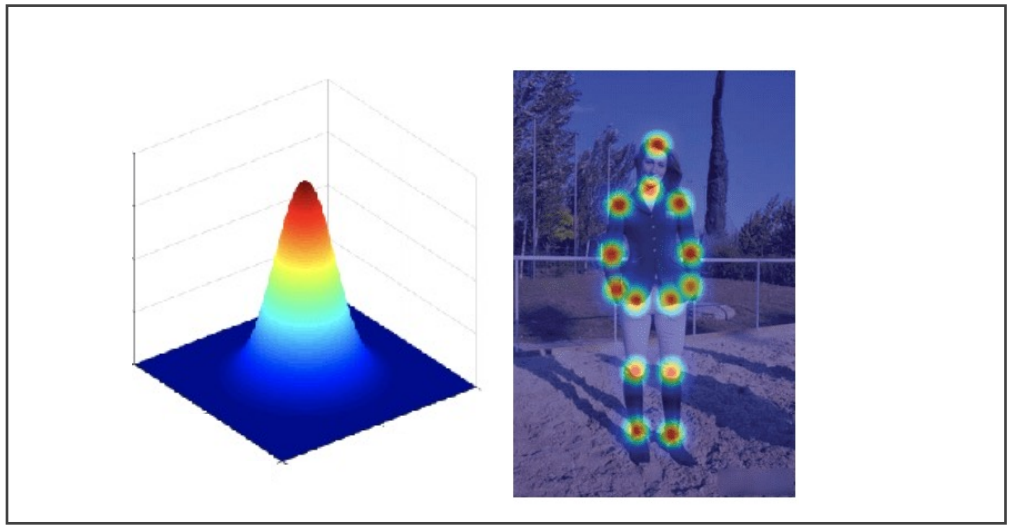
# Repeated Multi-scale Fusion

- Exchange units: each subnetwork repeatedly receives the information from other parallel subnetworks.
- Each block is composed of 3 parallel convolution units with an exchange unit across the parallel units, which is given as follows:



# Heatmap Estimation

## Gaussian heatmap



- The groundtruth heatmaps are generated by applying 2D Gaussian.
- The mean squared error is applied for regressing the heatmaps.

## Network Instantiation

| The type of nets | The widths(C) of the high-resolution subnetworks in last three stages |
|---|---|
| HRNet-W32 | 64, 128, 256 |
| HRNet-W48 | 96, 192, 384 |

- The network is instantiated by following the design rule of ResNet.
- The resolution is gradually decreased to a half, and accordingly, the width (the number of channels) is increased to double.

UniPose+: A unified framework for 2D and 3D human pose estimation in images and videos (TPAMI 2021)

# Question 🤔

# Experiments

# Dataset

Figure 4. Qualitative results of some example images in the MPII (top) and COCO (bottom) datasets: containing viewpoint and appearance change, occlusion, multiple persons, and common imaging artifacts.

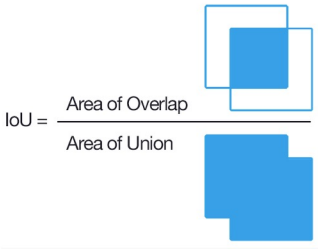
## COCO Keypoint Dataset

- 17 keypoints
- 200K images and 250K person instances
- Train, validation, test set
- Metric: OKS, AP

## MPII Human Pose Dataset

- 16 keypoints
- 25K images with 40K subjects
- Train, test set
- Metric: PCKh

# Evaluation Metric : Object Keypoint Similarity (OKS)




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

$d_i$ Euclidean distances between each ground truth and detected keypoint

$v_i$ Visibility flags of the ground truth

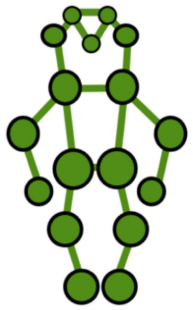
$v = 0$ : not labeled
$v = 1$ : labeled but not visible
$v = 2$ : labeled and visible

$s * k_i$ Scale*keypoint constant, the standard deviation of this gaussian

Perfect predictions will have OKS = 1
Predictions with wrong keypoints will have OKS~0

Constants for joints

| Keypoint | $k_i$ |
|---|---|
| hips | 0.107 |
| ankles | 0.089 |
| knees | 0.087 |
| shoulders | 0.079 |
| elbows | 0.072 |
| wrists | 0.062 |
| ears | 0.035 |
| nose | 0.026 |
| eyes | 0.025 |



$AP^{50}$ AP at OKS = 0.50

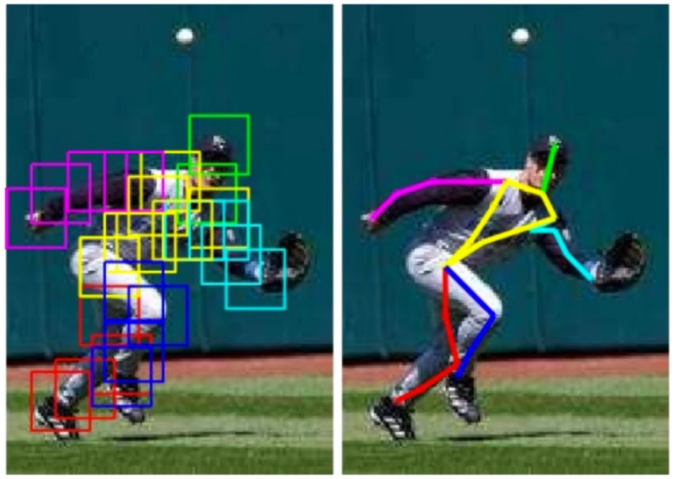
$AP^{M,L}$ The mean of AP scores at 10 positions*, for medium objects and large objects

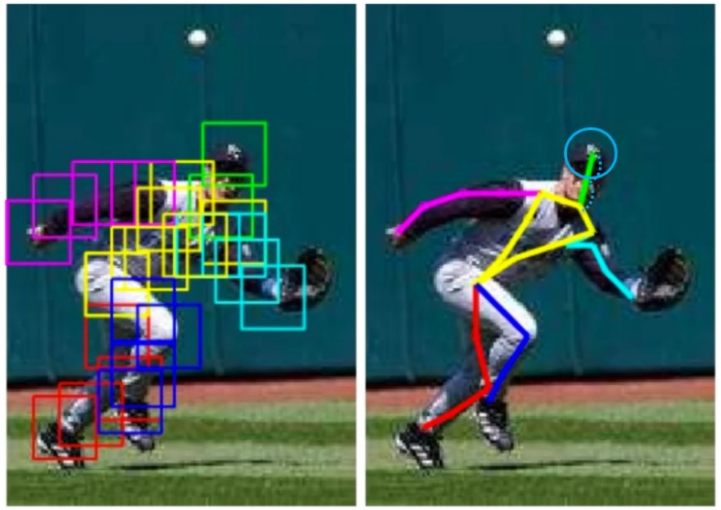
$AR$ Average recall at 10 positions*

*10 positions : OKS = 0.50, 0.55, . . . , 0.90, 0.95

Towards Accurate Multi-person Pose Estimation in the Wild (CVPR 2017)
Pose Estimation. Metrics. (stasiuk.medium.com)

# Evaluation Metric: Head-normalized Probability of Correct Keypoint (PCKh)





## PCK

- A detected joint is considered ‘correct’ if the distance between the predicted and the true joint is within a certain threshold.

## PCKh

- PCKh uses head size instead of bounding box size.
- PCKh@0.5 is when the threshold = 50% of the head bone link

Articulated Human Pose Estimation with Flexible Mixtures of Parts (CVPR 2011)

# COCO Keypoint Detection

- HRNet is significantly better than bottom-up approaches.
- It outperforms all the other top-down approaches and is more efficient in terms of model size and computation complexity.

Table 2. Comparisons on the COCO test-dev set. #Params and FLOPs are calculated for the pose estimation network, and those for human detection and keypoint grouping are not included.

| Method | Backbone | Input size | #Params | GFLOPs | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ | AR |
|---|---|---|---|---|---|---|---|---|---|---|
| Bottom-up: keypoint detection and grouping | | | | | | | | | | |
| OpenPose [6] | – | – | – | – | 61.8 | 84.9 | 67.5 | 57.1 | 68.2 | 66.5 |
| Associative Embedding [39] | – | – | – | – | 65.5 | 86.8 | 72.3 | 60.6 | 72.6 | 70.2 |
| PersonLab [46] | – | – | – | – | 68.7 | 89.0 | 75.4 | 64.1 | 75.5 | 75.4 |
| MultiPoseNet [33] | – | – | – | – | 69.6 | 86.3 | 76.6 | 65.0 | 76.3 | 73.5 |
| Top-down: human detection and single-person keypoint detection | | | | | | | | | | |
| Mask-RCNN [21] | ResNet-50-FPN | – | – | – | 63.1 | 87.3 | 68.7 | 57.8 | 71.4 | – |
| G-RMI [47] | ResNet-101 | 353 × 257 | 42.6M | 57.0 | 64.9 | 85.5 | 71.3 | 62.3 | 70.0 | 69.7 |
| Integral Pose Regression [60] | ResNet-101 | 256 × 256 | 45.0M | 11.0 | 67.8 | 88.2 | 74.8 | 63.9 | 74.0 | – |
| G-RMI + extra data [47] | ResNet-101 | 353 × 257 | 42.6M | 57.0 | 68.5 | 87.1 | 75.5 | 65.8 | 73.3 | 73.3 |
| CPN [11] | ResNet-Inception | 384 × 288 | – | – | 72.1 | 91.4 | 80.0 | 68.7 | 77.2 | 78.5 |
| RMPE [17] | PyraNet [77] | 320 × 256 | 28.1M | 26.7 | 72.3 | 89.2 | 79.1 | 68.0 | 78.6 | – |
| CFN [25] | – | – | – | – | 72.6 | 86.1 | 69.7 | 78.3 | 64.1 | – |
| CPN (ensemble) [11] | ResNet-Inception | 384 × 288 | – | – | 73.0 | 91.7 | 80.9 | 69.5 | 78.1 | 79.0 |
| SimpleBaseline [72] | ResNet-152 | 384 × 288 | 68.6M | 35.6 | 73.7 | 91.9 | 81.1 | 70.3 | 80.0 | 79.0 |
| HRNet-W32 | HRNet-W32 | 384 × 288 | 28.5M | 16.0 | 74.9 | 92.5 | 82.8 | 71.3 | 80.9 | 80.1 |
| HRNet-W48 | HRNet-W48 | 384 × 288 | 63.6M | 32.9 | **75.5** | **92.5** | **83.3** | **71.9** | **81.5** | **80.5** |
| HRNet-W48 + extra data | HRNet-W48 | 384 × 288 | 63.6M | 32.9 | **77.0** | **92.7** | **84.5** | **73.4** | **83.1** | **82.0** |

# MPII Human Pose Estimation

- The result is the best one among the previously-published results on the leaderboard of Nov. 16th, 2018.
- HRNet-W32 achieves a 92.3 PKCh@0.5 score and outperforms the stacked hourglass approach and its extensions.

Table 3. Performance comparisons on the MPII test set (PCKh@0.5).

| Method | Hea | Sho | Elb | Wri | Hip | Kne | Ank | Total |
|---|---|---|---|---|---|---|---|---|
| Insafutdinov et al. [27] | 96.8 | 95.2 | 89.3 | 84.4 | 88.4 | 83.4 | 78.0 | 88.5 |
| Wei et al. [69] | 97.8 | 95.0 | 88.7 | 84.0 | 88.4 | 82.8 | 79.4 | 88.5 |
| Bulat et al. [4] | 97.9 | 95.1 | 89.9 | 85.3 | 89.4 | 85.7 | 81.7 | 89.7 |
| Newell et al. [40] | 98.2 | 96.3 | 91.2 | 87.1 | 90.1 | 87.4 | 83.6 | 90.9 |
| Sun et al. [58] | 98.1 | 96.2 | 91.2 | 87.2 | 89.8 | 87.4 | 84.1 | 91.0 |
| Tang et al. [63] | 97.4 | 96.4 | 92.1 | 87.7 | 90.2 | 87.7 | 84.3 | 91.2 |
| Ning et al. [44] | 98.1 | 96.3 | 92.2 | 87.8 | 90.6 | 87.6 | 82.7 | 91.2 |
| Luvizon et al. [37] | 98.1 | 96.6 | 92.0 | 87.5 | 90.6 | 88.0 | 82.7 | 91.2 |
| Chu et al. [14] | 98.5 | 96.3 | 91.9 | 88.1 | 90.6 | 88.0 | 85.0 | 91.5 |
| Chou et al. [12] | 98.2 | 96.8 | 92.2 | 88.0 | 91.3 | 89.1 | 84.9 | 91.8 |
| Chen et al. [10] | 98.1 | 96.5 | 92.5 | 88.5 | 90.2 | **89.6** | 86.0 | 91.9 |
| Yang et al. [77] | 98.5 | 96.7 | 92.5 | 88.7 | 91.1 | 88.6 | 86.0 | 92.0 |
| Ke et al. [31] | 98.5 | 96.8 | 92.7 | 88.4 | 90.6 | 89.3 | **86.3** | 92.1 |
| Tang et al. [62] | 98.4 | **96.9** | 92.6 | 88.7 | **91.8** | 89.4 | 86.2 | **92.3** |
| SimpleBaseline [72] | 98.5 | 96.6 | 91.9 | 87.6 | 91.1 | 88.1 | 84.1 | 91.5 |
| HRNet-W32 | **98.6** | **96.9** | **92.8** | **89.0** | 91.5 | 89.0 | 85.7 | **92.3** |

# Ablation Study

- Figure 5 implies that the resolution does impact the keypoint prediction quality.
- Figure 6 implies that the improvement for the smaller input size is more significant than the larger input size.

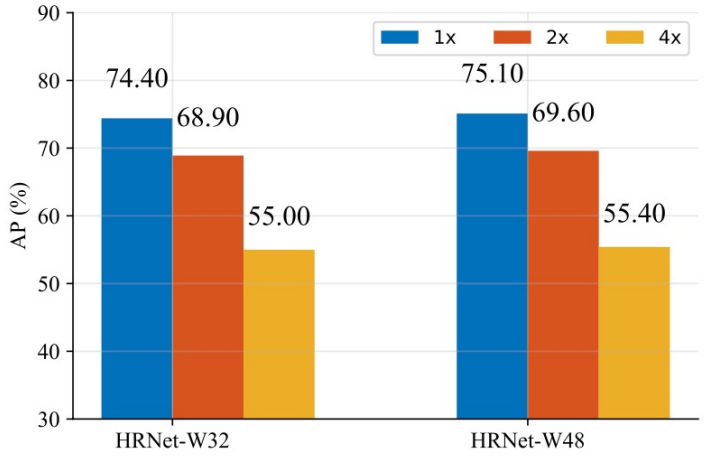



Figure 5. Ablation study of high and low representations. $1\times$, $2\times$, $4\times$ correspond to the representations of the high, medium, low resolutions, respectively.

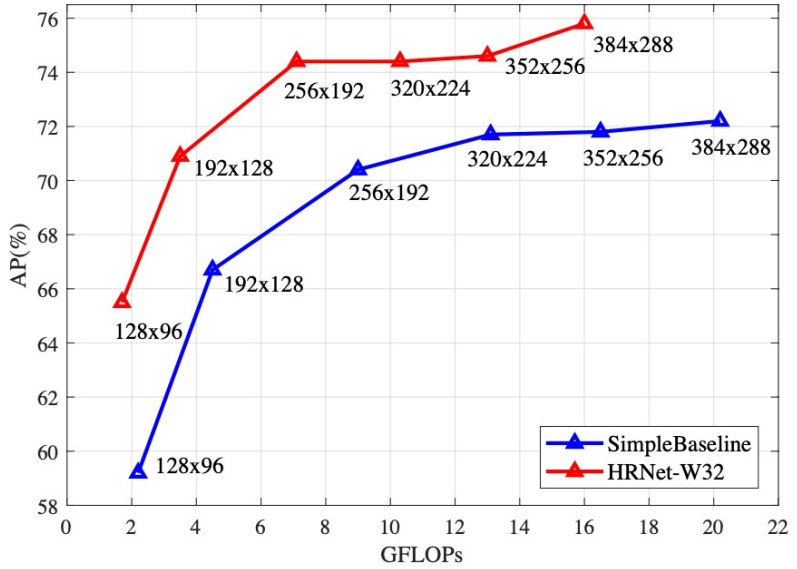
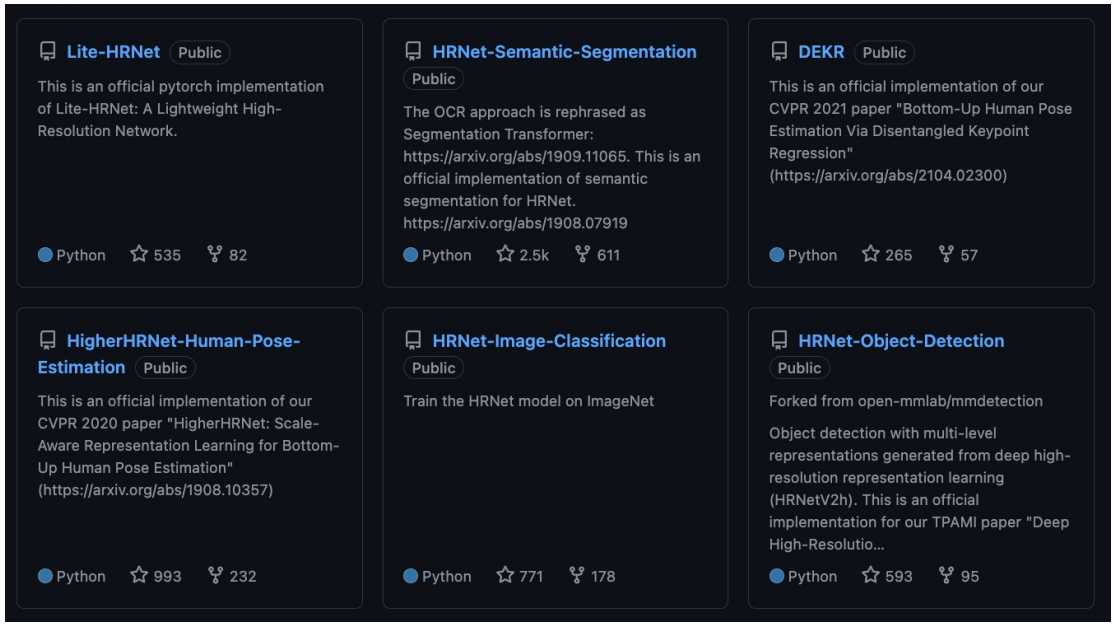



Figure 6. Illustrating how the performances of our HRNet and SimpleBaseline [72] are affected by the input size.

# Conclusion

# Conclusion




## Benchmarks

| Trend | Task | Dataset Variant | Best Model |
|---|---|---|---|
| | **Semantic Segmentation** | Cityscapes test | HRNetV2 + OCR + |
| | **Semantic Segmentation** | Cityscapes val | HRNetV2-OCR+PSA |

- The proposed method maintains the high resolution through the whole process without the need of recovering the high resolution.
- It fuses multi-resolution representations repeatedly, rendering reliable high-resolution representations.
- HRNet gets good performance in other computer vision tasks, including human pose estimation, semantic segmentation, and object detection.

# Question 🤔

# Sources


Paper
- Deep High-Resolution Representation Learning for Human Pose Estimation (CVPR 2019)
- High-Resolution Representations for Labeling Pixels and Regions (arXiv:1904.04514)
- Towards Accurate Multi-person Pose Estimation in the Wild (CVPR 2017)
- Articulated Human Pose Estimation with Flexible Mixtures of Parts (CVPR 2011)
- UniPose+: A unified framework for 2D and 3D human pose estimation in images and videos (TPAMI 2021)

YouTube
- [Paper Review] Deep High-Resolution Representation Learning for Human Pose Estimation (https://www.youtube.com/watch?v=w39bjQxm1eg)

Blog
- Pose Estimation. Metrics. (https://stasiuk.medium.com/pose-estimation-metrics-844c07ba0a78)


# Presenter


Seonok Kim (sokim0991@korea.ac.kr)