Deep High-Resolution Representation Learning for Human Pose Estimation

Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang

(CVPR 2019)

09.06.2022

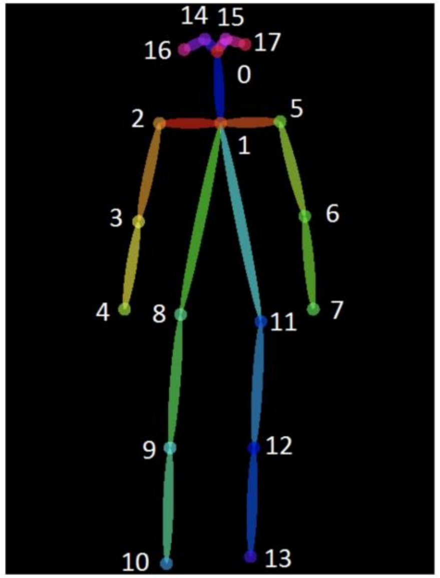
전북대학교 학부생 김세희

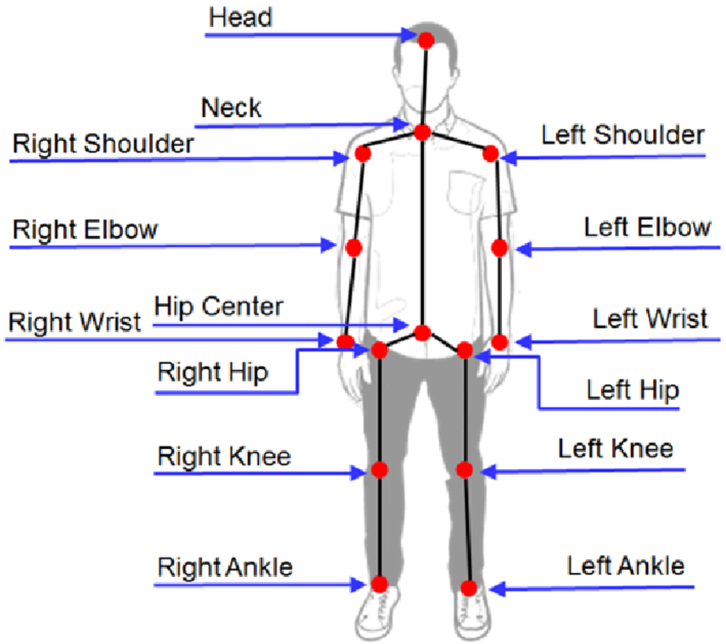
Preliminary

Background

- What is Human Pose Estimation?

  : a way of identifying and classifying the joints in the human body.





Preliminary

Regression vs Heatmap

- Regression: (x, y)   e.g. COCO Dataset

[Image] Image

[Annotation] Key points

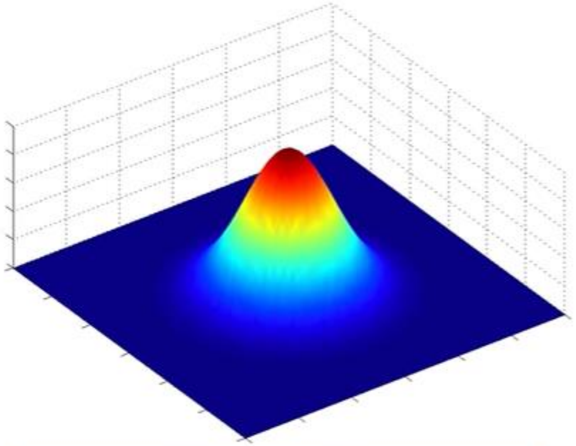
	코	눈(좌)	...	발목(우)
x	367	374	...	396
y	81	73	...	341
z	2	2	...	2

- x, y : (x, y), 2D image 좌표
- z : visibility flag
  - 0 : 이미지 내 존재하지 않는 키 포인트(not labeled)
  - 1 : 이미지 내 존재하지만, 겉으론 보이지 않는 키 포인트
  - 2 : 이미지 내 존재하고, 겉으로도 보이는 키 포인트

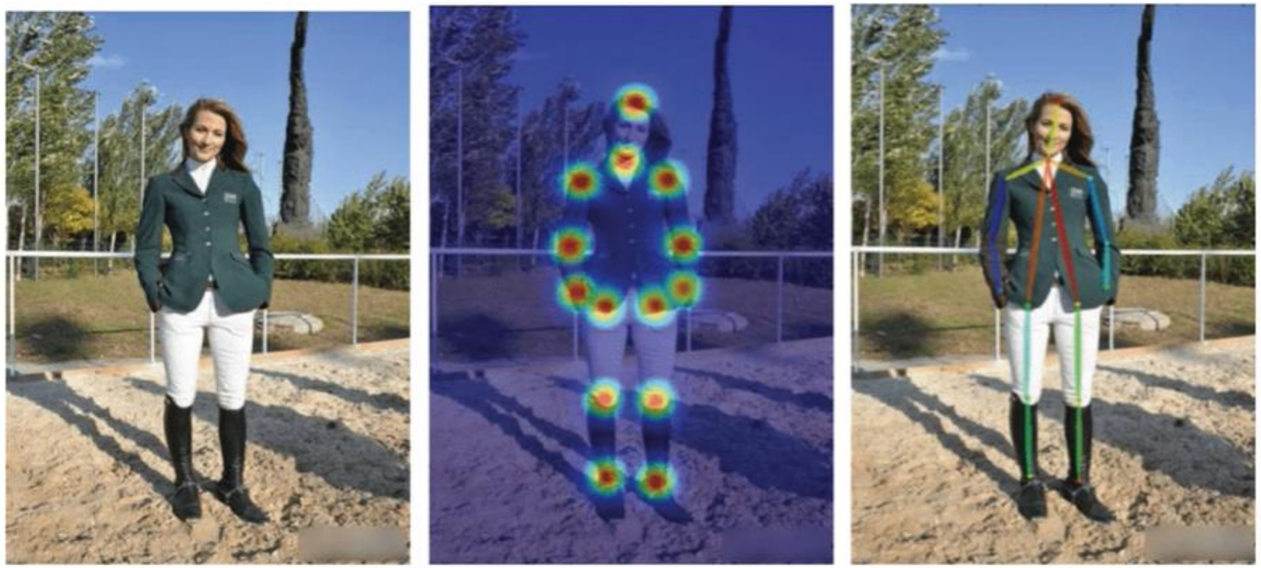
Preliminary

Regression vs Heatmap

- Heatmap [loc = (x, y)]



0	0	0	0	0	0	0
0	0	0	0.4	0	0	0
0	0	0.4	0.6	0.4	0	0
0	0.4	0.6	0.8	0.8	0.4	0
0.4	0.6	0.8	1.0	0.8	0.6	0.4
0	0.4	0.6	0.8	0.6	0.4	0
0	0	0.4	0.6	0.4	0	0
0	0	0	0.4	0	0	0
0	0	0	0	0	0	0



(a) (b) (c)

Preliminary

Human Pose Estimation Task

1. **Train Global + Local Feature**
2. **Recover High-Resolution**

✓ Trade off: Global information vs High resolution

High global information **Low resolution**

When up-sampling, lose information

Increase receptive field size

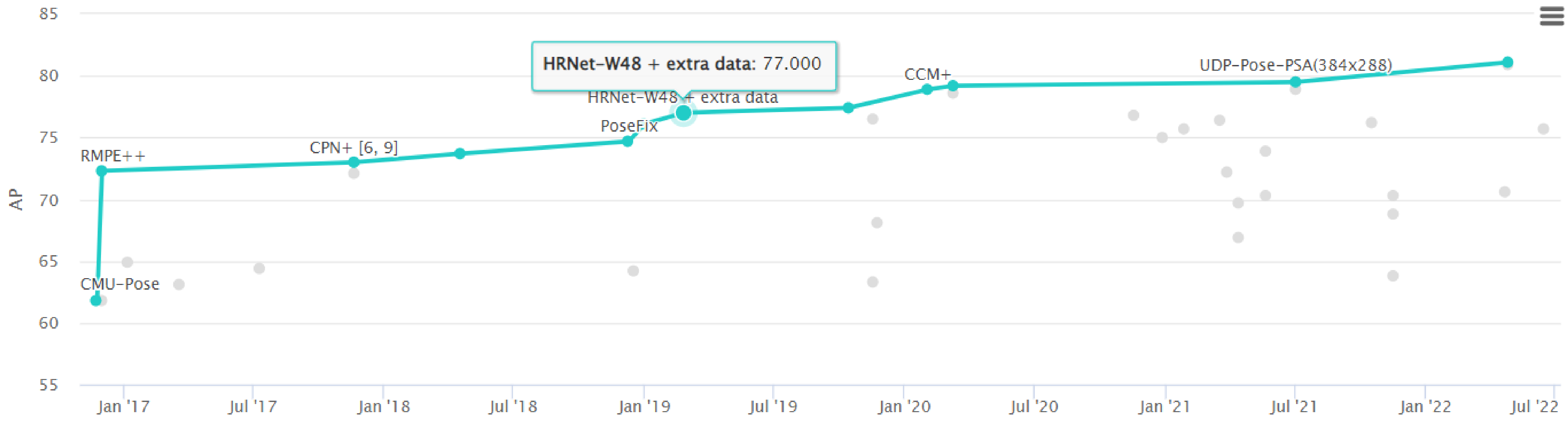
Negative effect for Pixel-wise prediction

Preliminary

HR-Net Record

- Leader Board: Pose Estimation on COCO test-dev

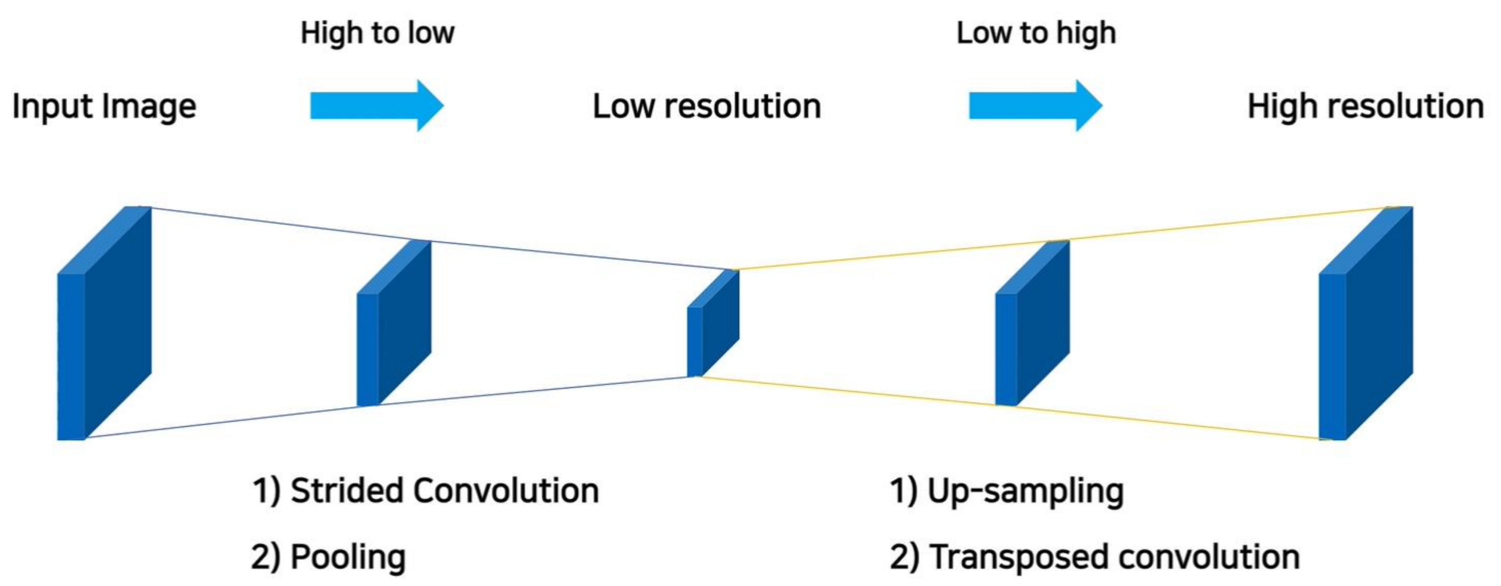
  : It maintained SOTA for about two years.



Introduction

Previous Approach

- Existing networks for pose estimation are built by connecting high-to-low resolution subnetworks in series. e.g. Simple baseline(2018)



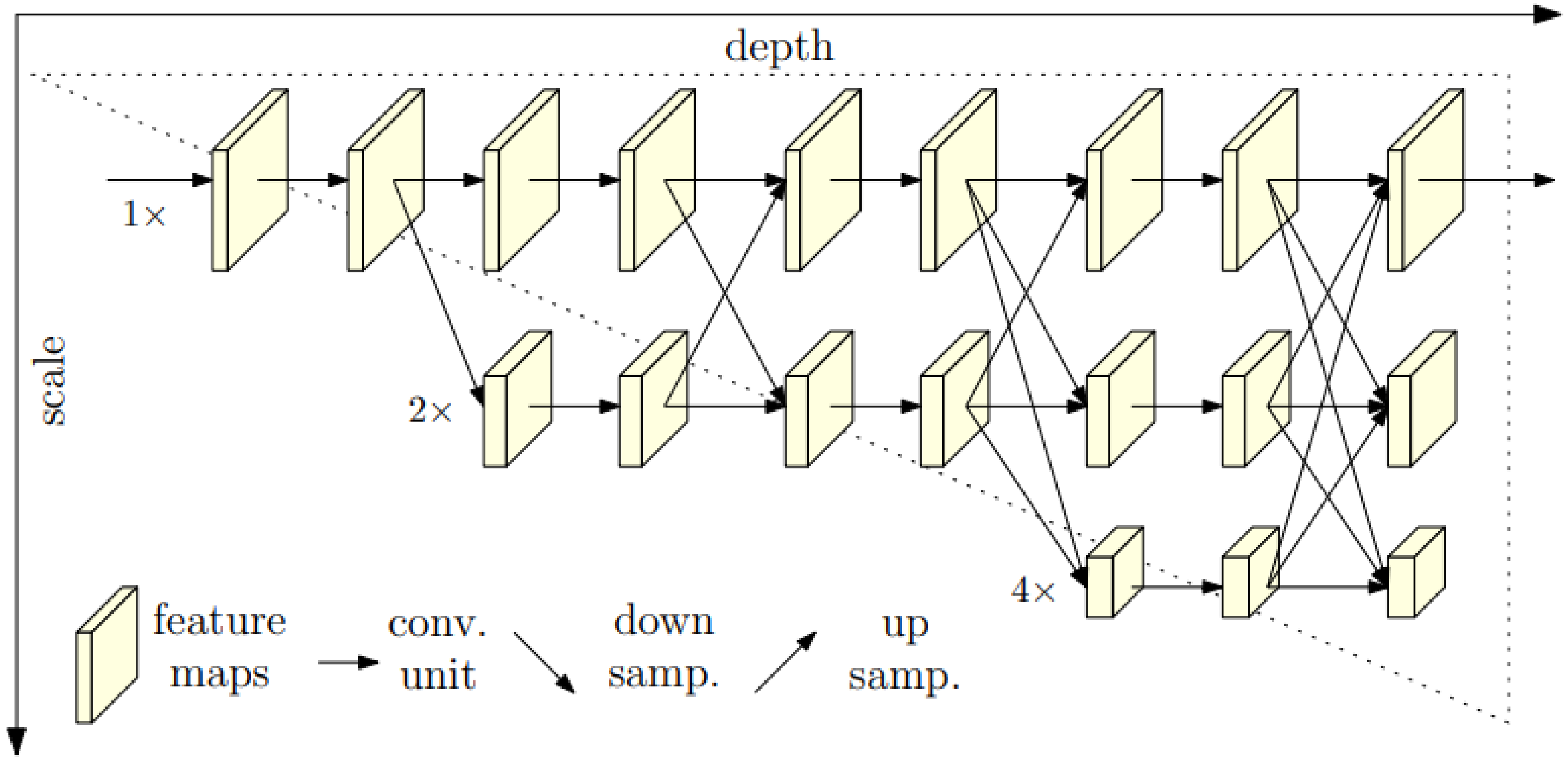
Lose small object or detailed spatial information

Negative effect for Pixel-wise prediction

Method

Architecture of the proposed HRNet

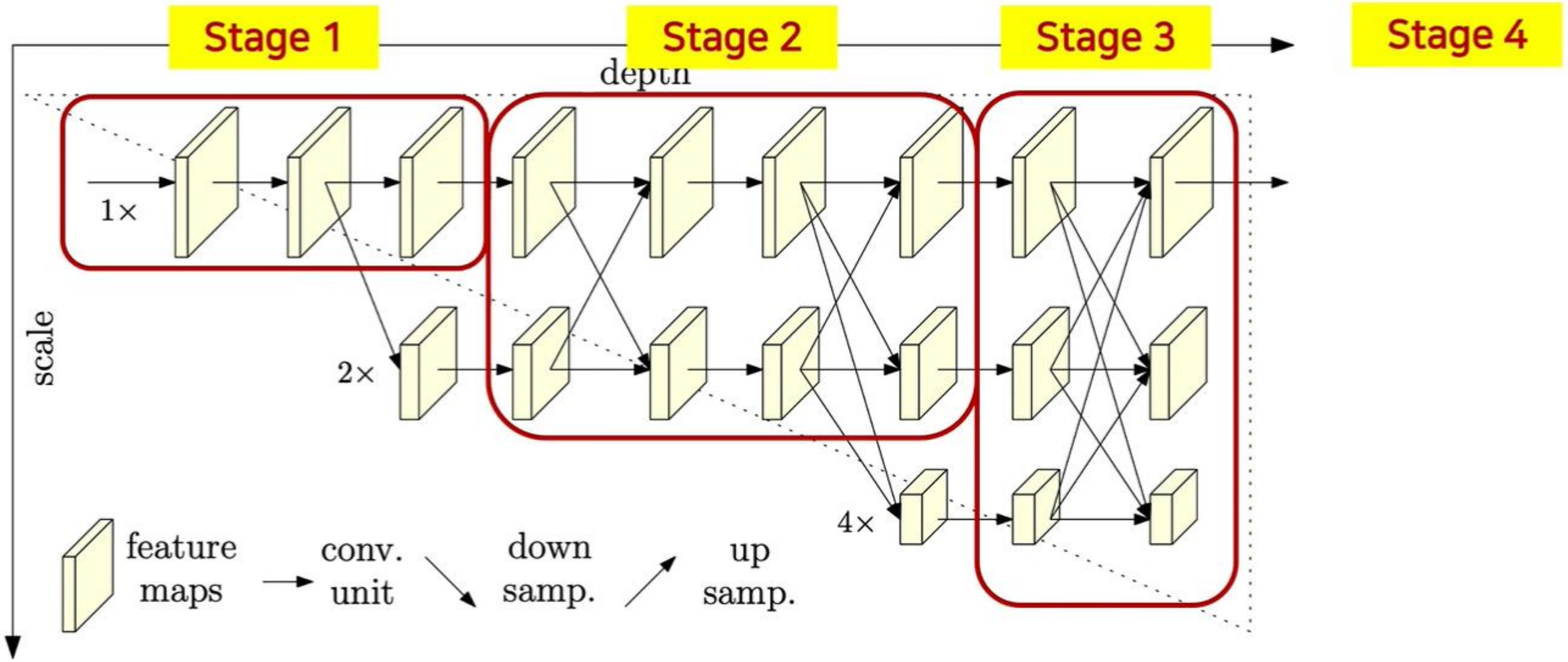
- It consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion)



Method

Architecture of the proposed HRNet

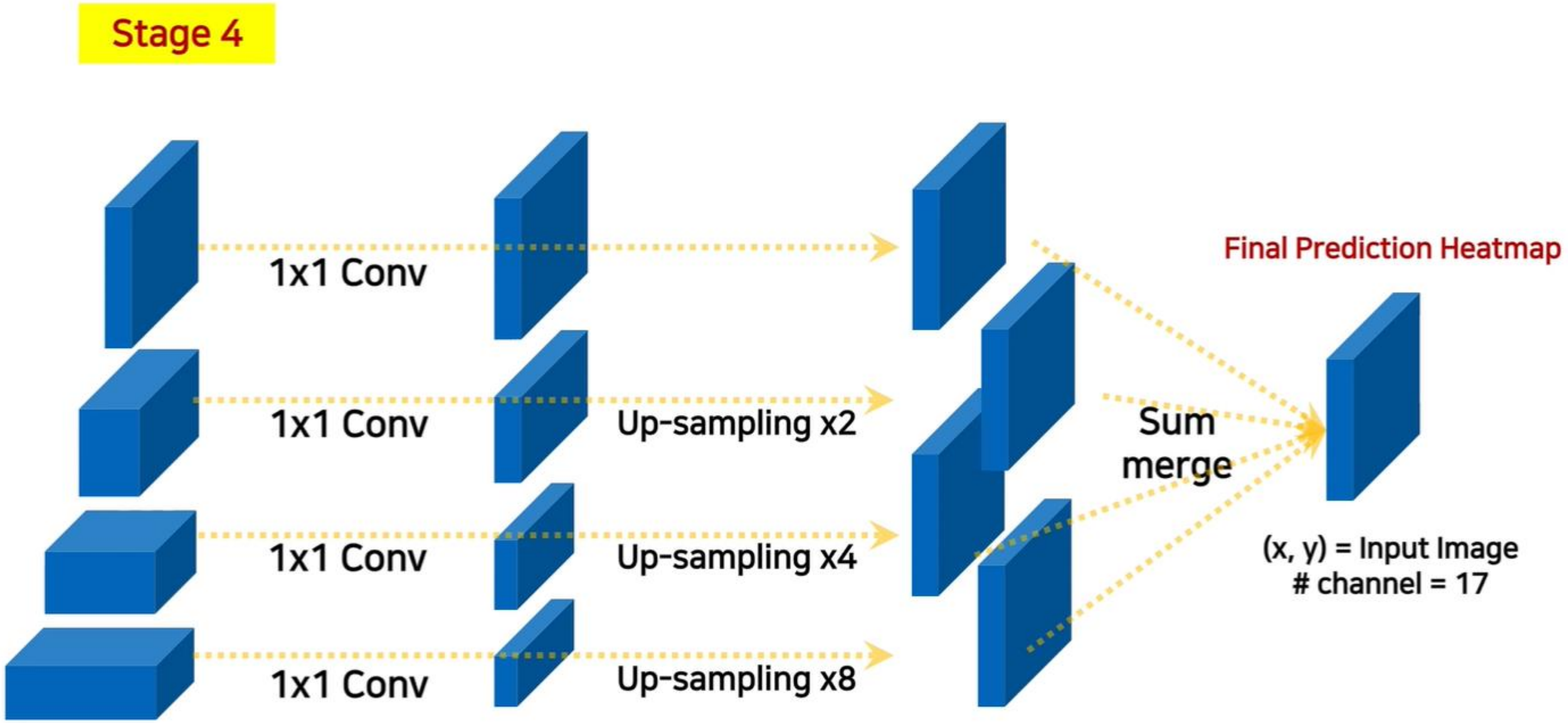
- The resolutions for the parallel subnetworks of a later stage consists of the resolutions from the previous stage, and an extra lower one.



Method

Architecture of the proposed HRNet

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



Experiments

Dataset (COCO Dataset)

- Train dataset: 57,000 images (150K person instances)
- Evaluation: 5,000 images(val), 20,000 images(test-dev)
- Key-points: 17

Evaluation metric

- OKS (= Object Keypoint Similarity)
- AP (=Average Precision)

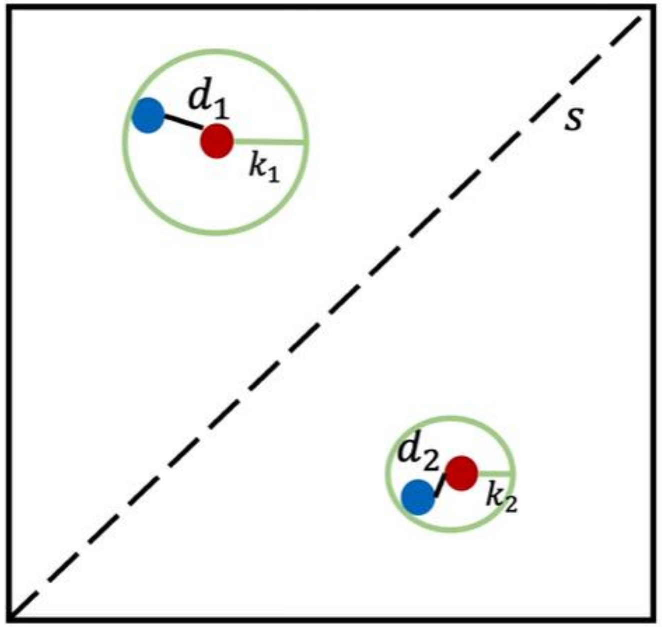
Experiments

Evaluation metric

* Object Detection task에서의 'IOU(Intersection over Union)' 개념

1. 키 포인트 유사성 측정 지표 : OKS (=Object Keypoint Similarity)

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



d_i	Euclidean 거리(Ground-Truth 관절(key-point), 예측 관절)
v_i	Visibility flag ($v_i > 0$: 이미지 내 존재하는 모든 키 포인트)
s	객체 Bounding box 대각선 길이
k_i	관절 종류마다 사전 설정되어 있는 상수

✓ OKS는 1(Best)과 0(Worst)사이의 값을 가짐

- 완벽한 예측의 경우, d_i는 0이 되어 OKS값이 0이 됨
- 반대의 경우, d_i값이 매우 커져 $\exp(-x)$값이 0에 접근하며 OKS가 0이 됨

- $s^2 k_i^2$ 역할 : 각 Image, 관절 마다 일종의 '정규화'

Experiments

Evaluation metric

2. 평가 Metric : AP (= Average Precision)

[Average Precision]

1) OKS Threshold에 따른 'Precision', 'Recall' 값 계산
2) Recall에 따른 Precision 값을 도식화 해 'Precision-Recall' curve
3) 'Precision-Recall' curve 의 면적을 'Average Precision' 으로 정의

AP	The mean of AP at 10 positions (OKS = [0.5, 0.05, 0.95])
AP^{50}	AP at OKS = 0.5
AP^{75}	AP at OKS = 0.75
AP^{M}	AP for medium object ($32^2 < segmentation\ area < 96^2$)
AP^{L}	AP for Large object ($segmentation\ area > 96^2$)
AR	Average Recall

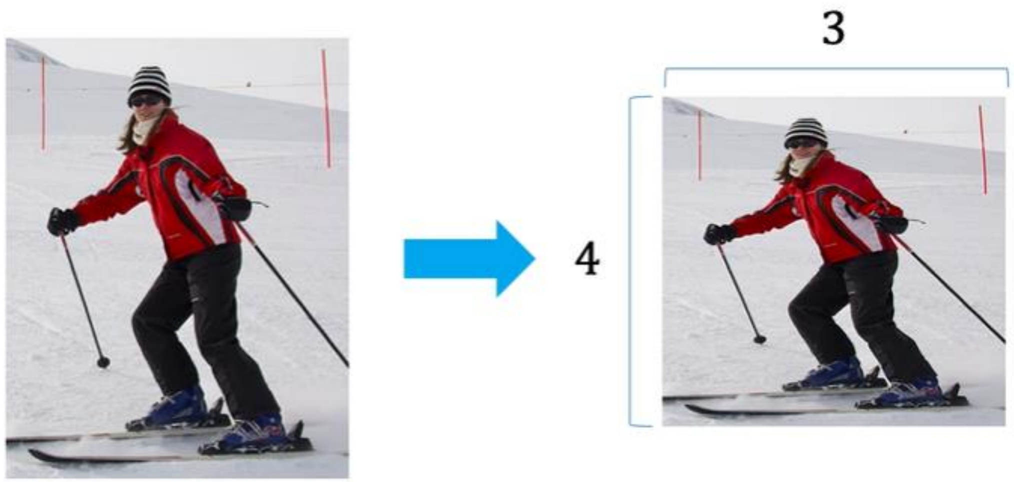
Experiments

Training

1. Extension of Human detection box image

- Fixed Human detection box image ($height : width = 4 : 3$)

  예시) 256×192 or 384×288



2. Data Augmentation

   1) Random Rotation($-45^{\circ}, 45^{\circ}$)

   2) Random scale($0.56, 1.35$)

   3) flipping

   4) half body data augmentation

Results

COCO validation set

Table 1. Comparisons on the COCO validation set. Pretrain = pretrain the backbone on the ImageNet classification task. OHKM = online hard keypoints mining [11].

Method	Backbone	Pretrain	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stage Hourglass [40]	8-stage Hourglass	N	256×192	25.1M	14.3	66.9	–	–	–	–	–
CPN [11]	ResNet-50	Y	256×192	27.0M	6.20	68.6	–	–	–	–	–
CPN + OHKM [11]	ResNet-50	Y	256×192	27.0M	6.20	69.4	–	–	–	–	–
SimpleBaseline [72]	ResNet-50	Y	256×192	34.0M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [72]	ResNet-101	Y	256×192	53.0M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [72]	ResNet-152	Y	256×192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32	HRNet-W32	N	256×192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32	HRNet-W32	Y	256×192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48	HRNet-W48	Y	256×192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
SimpleBaseline [72]	ResNet-152	Y	384×288	68.6M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W32	HRNet-W32	Y	384×288	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNet-W48	HRNet-W48	Y	384×288	63.6M	32.9	**76.3**	**90.8**	**82.9**	**72.3**	**83.4**	**81.2**

Results

COCO validation set

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**

Ablation Study

Table 6. Ablation study of exchange units that are used in repeated multi-scale fusion. Int. exchange across = intermediate exchange across stages, Int. exchange within = intermediate exchange within stages.

Method	Final exchange	Int. exchange across	Int. exchange within	AP
(a)	✓			70.8
(b)	✓	✓		71.9
(c)	✓	✓	✓	73.4

Results

Ablation Study

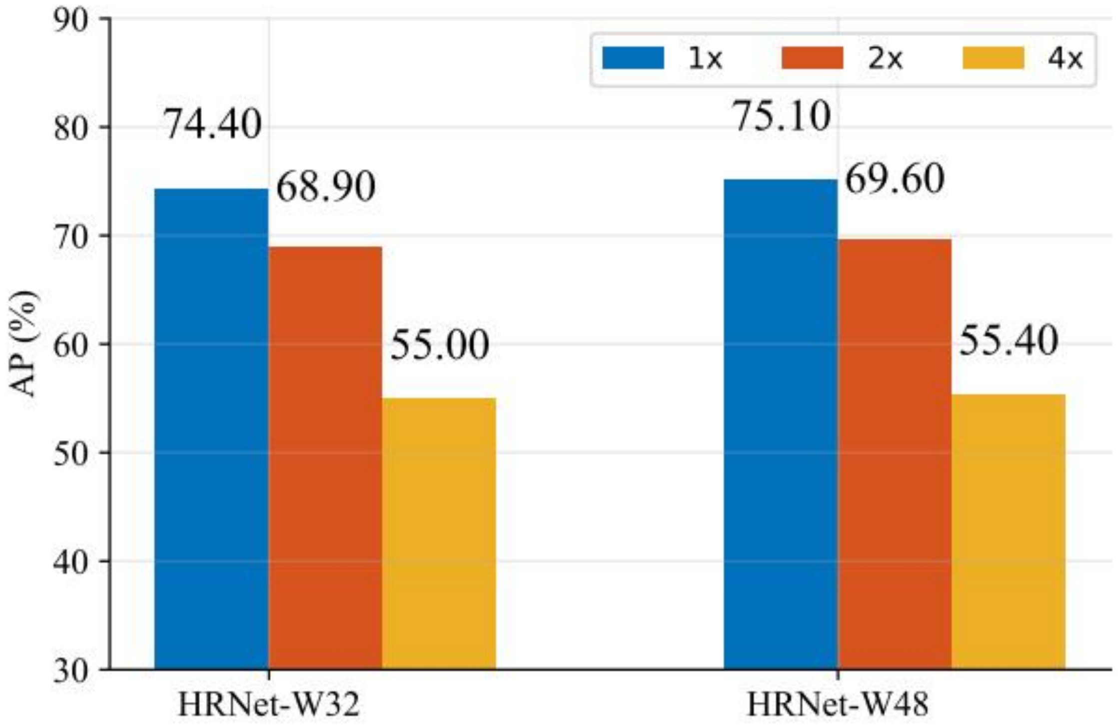


Figure 5. Ablation study of high and low representations. 1×, 2×, 4× correspond to the representations of the high, medium, low resolutions, respectively.

Results

Ablation Study

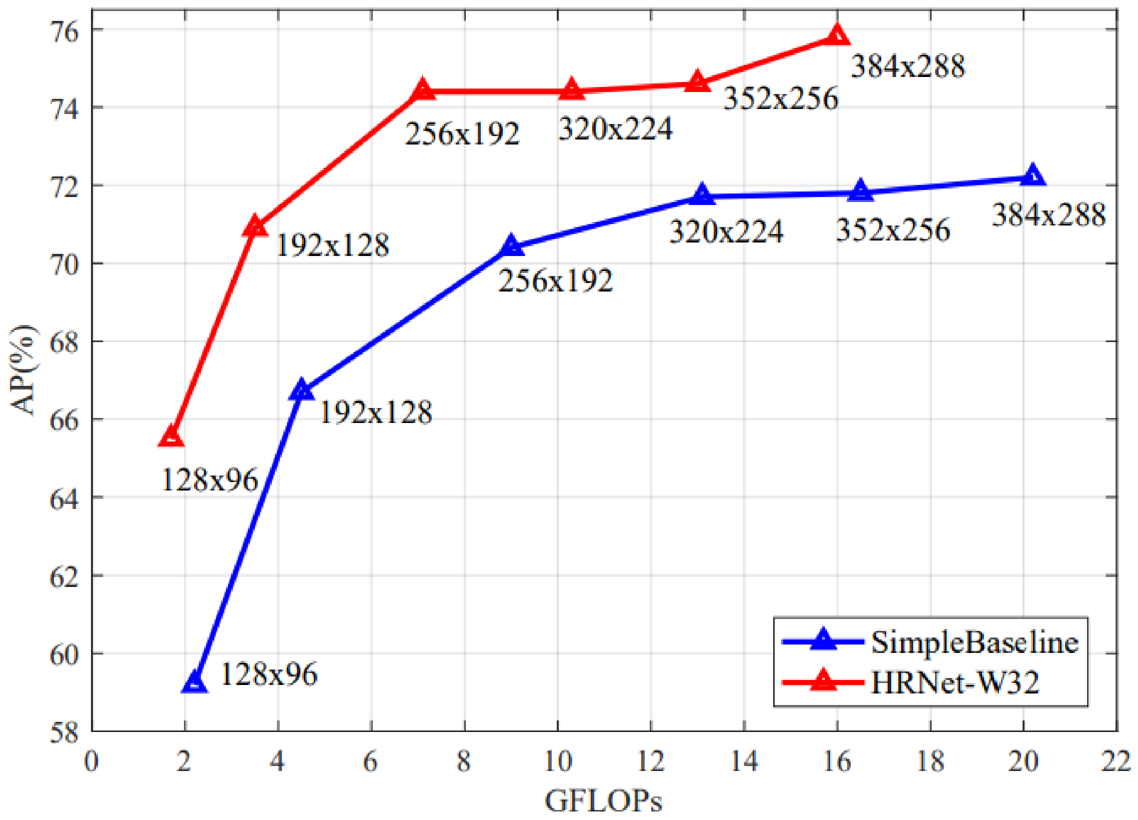
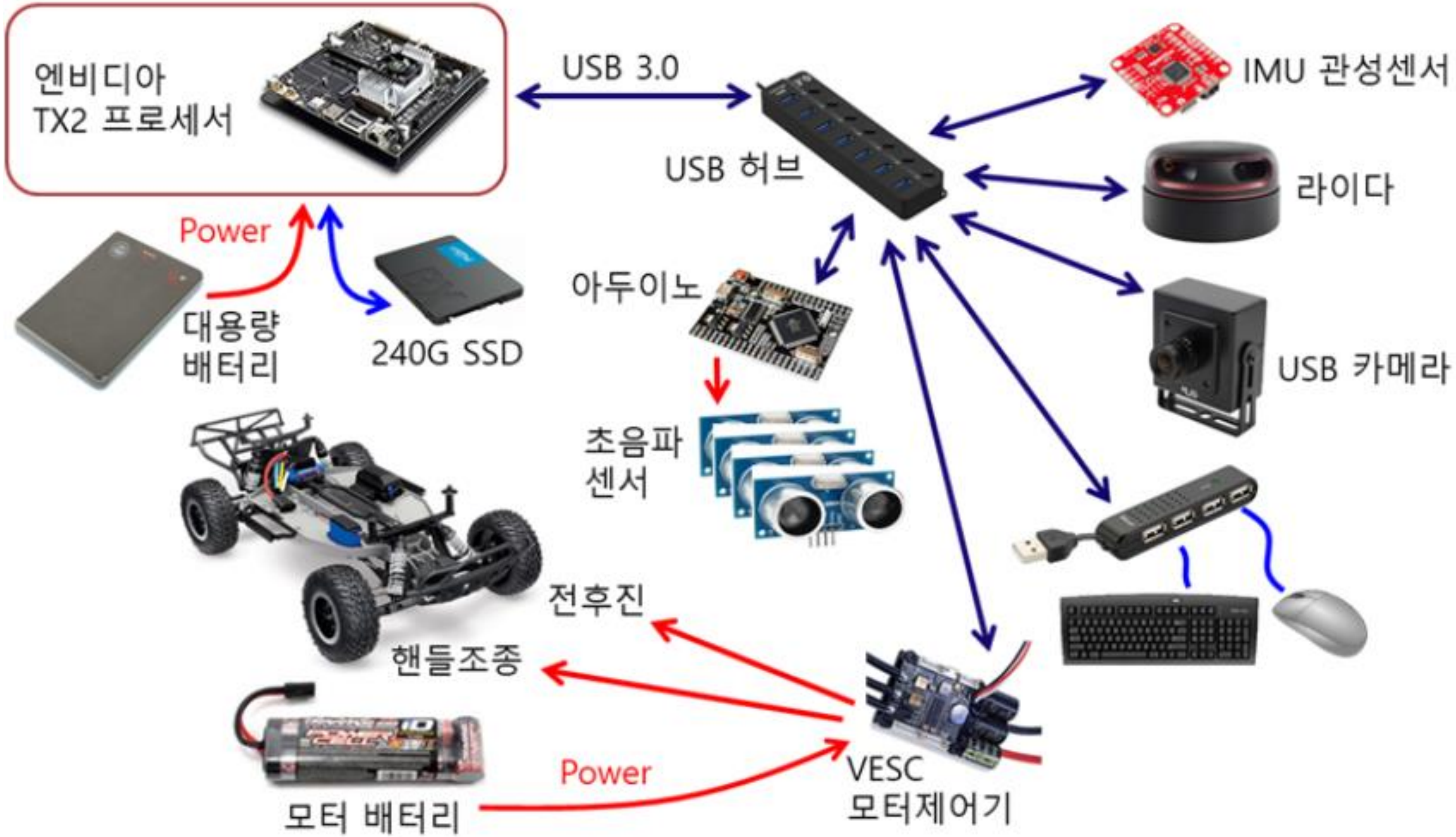
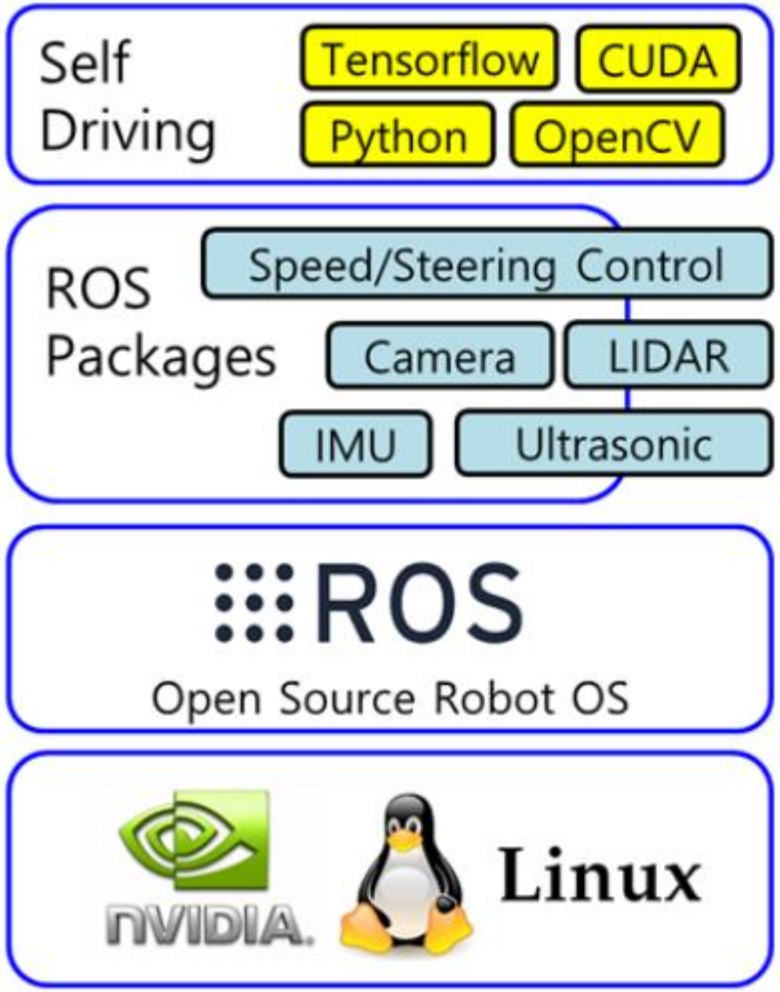


Figure 6. Illustrating how the performances of our HRNet and SimpleBaseline [72] are affected by the input size.

Thank you


엔비디아
TX2 프로세서
USB 3.0
USB 허브
IMU 관성센서
라이다
USB 카메라
Power
대용량
배터리
240G SSD
아두이노
초음파
센서
전후진
핸들조종
Power
모터 배터리
VESC
모터제어기



Self Driving
Tensorflow
CUDA
Python
OpenCV
ROS Packages
Speed/Steering Control
Camera
LIDAR
IMU
Ultrasonic
ROS
Open Source Robot OS
NVIDIA
Linux