

Vision AI

Deep High-Resolution Representation Learning for Human Pose Estimation

2022-06-29 | JiHyun Lee

Deep High-Resolution Representation Learning for Human Pose Estimation

Ke Sun^{1,2*†} Bin Xiao^{2*} Dong Liu¹ Jingdong Wang²

¹University of Science and Technology of China ²Microsoft Research Asia

{sunk,dongeliu}@ustc.edu.cn, {Bin.Xiao,jingdw}@microsoft.com

<https://arxiv.org/pdf/1902.09212.pdf>
[git code](#)

https://arxiv.org › cs ▾

Deep High-Resolution Representation Learning for Human ...

K Sun 저술 · 2019 · 1868회 인용 — In this work, we are interested in the human pose estimation problem with a focus on learning reliable high-resolution representations.

Cite as: arXiv:1902.09212

이 페이지를 2번 방문했습니다. 최근 방문 날짜: 22. 6. 26

Comments: accepted by CVPR2019

Subjects: Computer Vision and Pattern Recognition (cs.CV)

Cite as: arXiv:1902.09212 [cs.CV]

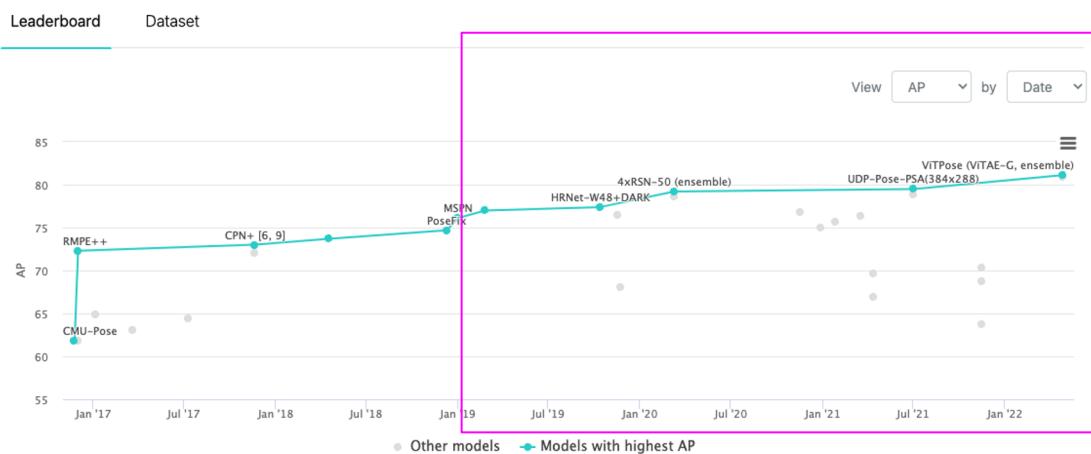
(or arXiv:1902.09212v1 [cs.CV] for this version)

<https://doi.org/10.48550/arXiv.1902.09212> ⓘ

Deep High-Resolution Representation Learning for Human Pose Estimation

Leader Board : Pose Estimation on COCO test-dev

- 2019년부터 약 2년 간, HR-Net 모델이 SoTA 를 유지
- UDP-Pose-PSA
 - HR-Net 을 Backbone 으로, Polarized Self-attention (PSA) 적용
 - Input size 가 다른 경우 모두 , 기존의 vanilla HR-Net 보다 높은 성능
- DARK : Distribution-Aware coordinate Representation of Keypoint (DARK)
 - Model agnostic 접근. 최종 Heatmap 예측 시, Key-point localization 의 정확성 높이는 효과



Rank	Model	AP	AP50	AP75	APL	APM	AR	Paper	Code	Result	Year
1	VITPose (VITAE-G, ensemble)	81.1	95.0	88.2	86.0	77.8	85.6	VITPose: Simple Vision Transformer Baselines for Human Pose Estimation	Link	Result	2022
2	VITPose (VITAE-G)	80.9	94.8	88.1	85.9	77.5	85.4	VITPose: Simple Vision Transformer Baselines for Human Pose Estimation	Link	Result	2022
3	UDP-Pose-PSA (384x288)	79.5	93.6	85.9	84.3	76.3	81.9	Polarized Self-Attention: Towards High-quality Pixel-wise Regression	Link	Result	2021
4	4xRSN-50 (ensemble)	79.2	94.4	87.1	76.1	83.8	84.1	Learning Delicate Local Representations for Multi-Person Pose Estimation	Link	Result	2020
5	UDP-Pose-PSA (256x192)	78.9	93.6	85.8	83.6	76.1	81.4	Polarized Self-Attention: Towards High-quality Pixel-wise Regression	Link	Result	2021
6	4xRSN-50	78.6	94.3	86.6	75.5	83.3	83.8	Learning Delicate Local Representations for Multi-Person Pose Estimation	Link	Result	2020
7	HRNet-W48+DARK	77.4	92.6	84.6	83.7	73.6	82.3	Distribution-Aware Coordinate Representation for Human Pose Estimation	Link	Result	2019
8	HRNet-W48 + extra data	77	92.7	84.5	83.1	73.4	82	Deep High-Resolution Representation Learning for Human Pose Estimation	Link	Result	2019

- **Human Pose Estimation**
 - Pose Estimation
 - Human Pose Estimation Task
 - Brief History
- **HR-Net**
 - Motivation
 - Model Structure
 - Experiments

Human Pose Estimation

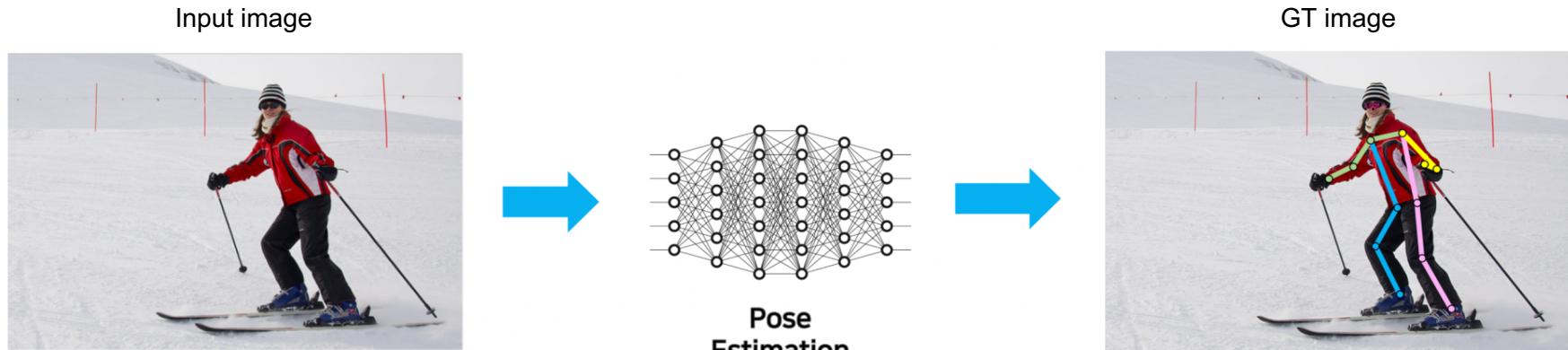
- **Pose Estimation**

- 주어진 영상 속 Human Object 의 자세 (pose) 를 추정하는 것
- Pose 추정 = 주요한 관절을 찾는 것
 - 관절 = joints = key points: 사람의 자세를 구성하는 주요 관절 포인트
 - 해당 과제는 아래와 같이 불리기도 함
 - Key-point detection, Pose recognition
- i.e. 특정 pose 를 만들어내는 Key-points 들을 찾아내는 task (= Key-points Localization)



Human Pose Estimation; Human Pose Estimation Task

- Overall Process



- Input image 상 17 개 Key points 의 (x, y) 좌표 값 예측
- Classification task 와는 다르게 up-sampling 을 통해 high-resolution 으로 이미지를 잘 복원하는지가 핵심

- Overall Process

- Data, Annotation (e.g. COCO Dataset)

[Input image]



[Annotation] Key points

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

- 3개 tuple (x, y, z)
- x, y: (x, y), 2D image 좌표
- z: visibility flag
 - 0 : 이미지 내 존재하지 않는 키 포인트 (not labeled)
 - 1 : 이미지 내 존재하지만, 겉으로 보이지 않는 키 포인트
 - 2 : 이미지 내 존재하고, 겉으로도 보이는 키 포인트

Human Pose Estimation; Human Pose Estimation Task

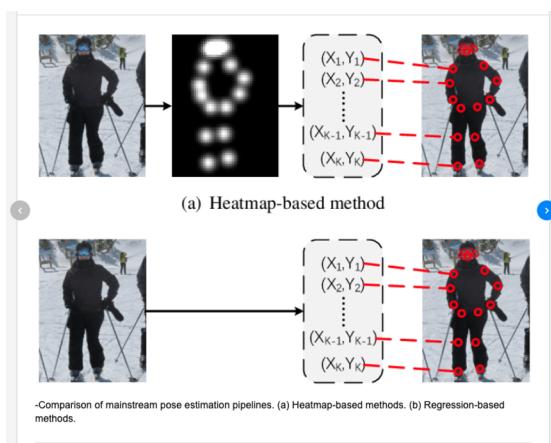
- Regression vs Heatmap

- Regression 접근 (x, y)

- (2013) DeepPose : Human Pose Estimation via Deep Neural Networks (Toshev, CVPR)
 - 최초의 딥러닝 기반 2D Human pose estimation model
 - CNN regressor 를 통해 key-points 의 좌표 값 (x, y) 를 직접 추정
 - 기존 방법론 대비 큰 성능 향상을 이루진 못함

- Heatmap 접근 [$\text{loc} = (x,y)$]

- (2015) Efficient Object Localization Using Convolutional Networks (Jonathan Tompson, CVPR)
 - Output 을 (x, y) 좌표 대신, Heatmap 을 통해 상대적인 빈도, 확률값으로 가져가도록 함
 - Input image 의 여러 pose 로 인해 생기는 다양성을 잘 포용하고, 보다 Robust 한 모델을 가능하게 함
 - 좌표 기반 방법론보다 2배 이상 높은 성능



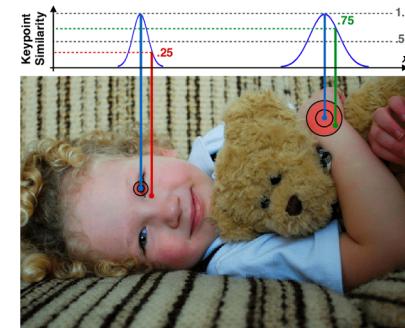
(a)



(b)



(c)



HUMAN TOUCH IN MEDICINE

Human Pose Estimation; Human Pose Estimation Task

- **Single vs Multi**

- Single-Person Pose Estimation (SPPE)
- Multi-Person Pose Estimation
 - Top-Down Approaches
 - Bottom-Up Approaches



Figure 5: Single-Person Vs. Multi-Person Pose Estimation

Human Pose Estimation; Human Pose Estimation Task

- **Single vs Multi**

- Single-Person Pose Estimation (SPPE)

- **Multi-Person Pose Estimation**

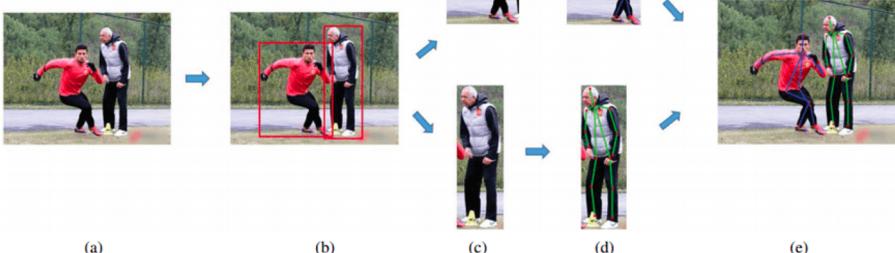
- **Top-Down Approaches** ⇒ HR-Net
- Bottom-Up Approaches



Original Image



Result



- 사실상 Single Pose Estimation Task
 1. Object (사람) Detection
 2. Key Points Detection (매 사람마다)
- 속도 느림, 정확도 높음

Human Pose Estimation; Human Pose Estimation Task

- **Single vs Multi**

- Single-Person Pose Estimation (SPPE)
- **Multi-Person Pose Estimation**
 - Top-Down Approaches
 - Bottom-Up Approaches



1. Key Points Detection (모든 사람)
 2. 각 Key Points 의 연결
- 속도 빠름, 정확도 낮음

Human Pose Estimation; Brief History

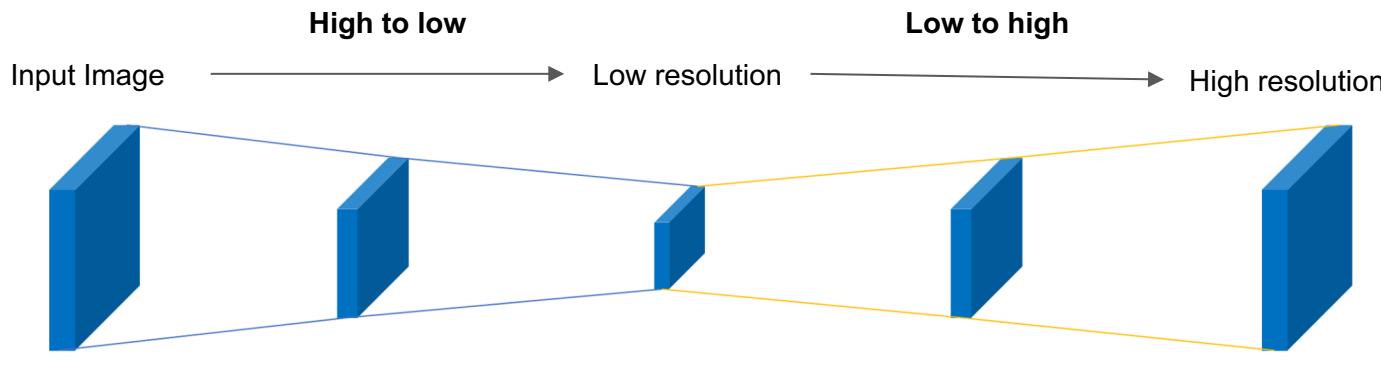
- ML (1970~)
 - Pictorial Structure Model
 - **AlexNet (2012)**
 - **DeepPose (2013)**
 - Deep Learning Network 최초 접목
 - ResNet (2015)
 - Deconvolution Network (2015)
 - Efficient Object Localization Using Convolutional Networks (2015)
 - Heatmap based approach
 - Convolutional Pose Machines (2016)
 - Receptive field 확대, Global Context 학습
 - Stacked hourglass networks for human pose estimation (2016)
 - Global + Local Context 학습
 - Deeper Cut (2016)
 - Cascaded Pyramid Network for Multi-Person Pose Estimation (2018)
 - Global Net + Refine Net
 - Simple Baselines for Human Pose Estimation and Tracking (2018)
 - + Deconvolution
 - **HR-Net (2019)**
 - Parallel Multi-Resolution Feature structure
 - Polarized Self-Attention (2021)
-
- The diagram illustrates the evolution of Human Pose Estimation models over time. It features a vertical timeline of models on the left and three large curly braces on the right, each grouping models into distinct categories based on their architectural or methodological advancements.
- Early ML models (1970~):** Pictorial Structure Model.
 - Deep learning models (2012-2015):** AlexNet, DeepPose, ResNet, Deconvolution Network, Efficient Object Localization Using Convolutional Networks. This group is bracketed by a single brace on the right.
 - Advanced models (2016-2021):** Convolutional Pose Machines, Stacked hourglass networks for human pose estimation, Deeper Cut, Cascaded Pyramid Network for Multi-Person Pose Estimation, Simple Baselines for Human Pose Estimation and Tracking, HR-Net (2019), and Polarized Self-Attention. This group is bracketed by two braces on the right.
- The bracketing indicates a progression from early machine learning approaches to deep learning, and finally to more complex, multi-resolution feature-based systems.

HR-Net

- 기준의 접근

- High-to-low
- Low-to-high

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



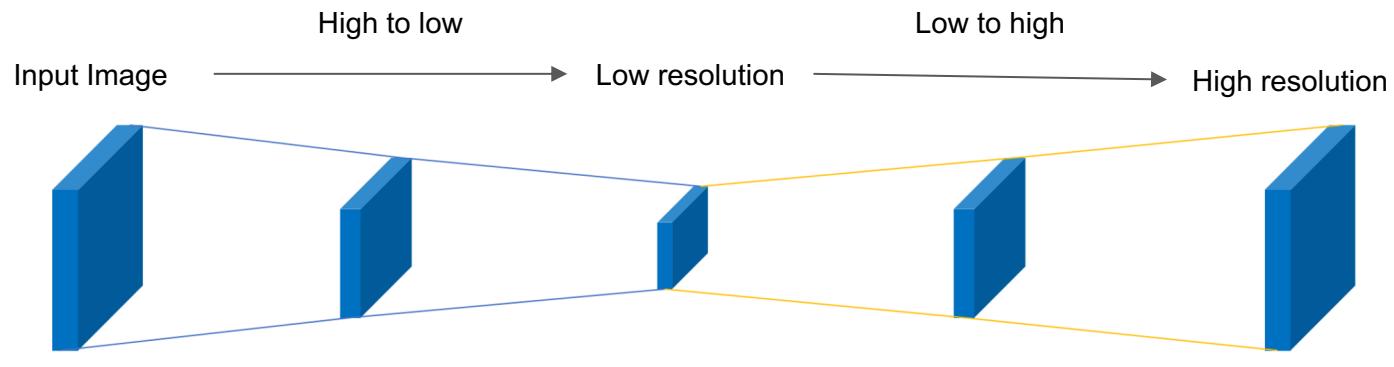
- Classification 수행 시, image feature extraction
 - Strided Convolution
 - Pooling
- Small Object 혹은 Detail 한 spatial information 의 손실

- Image 복원
 - Upsampling
 - Transposed convolution

Pixel-wise prediction에 부정적 영향

- **기준의 접근**

- High-to-low
- Low-to-high

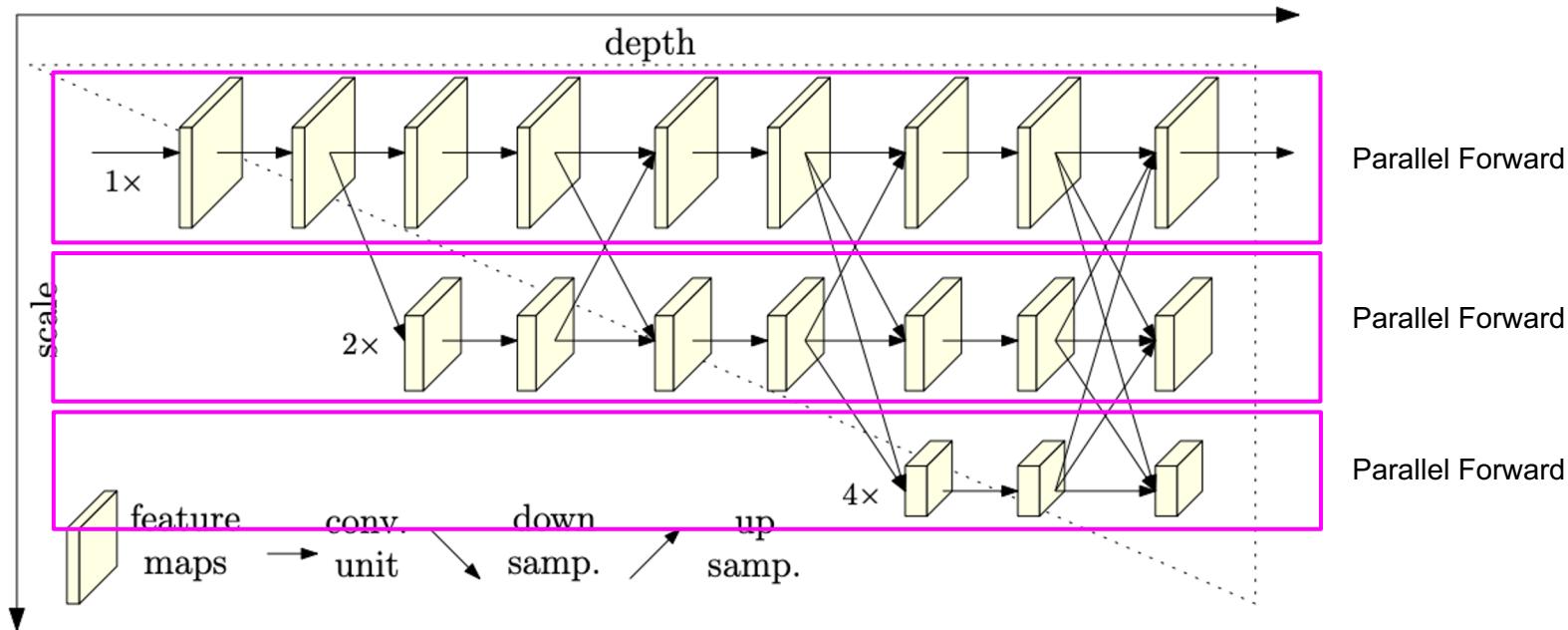


- **Network 직렬화**

- Local, Global 특징 추출과 학습의 과정이 직렬화되어 있어, 최종적으로 Up-sampling에 모든 process 가 과도한 의존
 - 직렬 구조를 벗어나면 receptive field 를 확장하면서도 local information 손실이 크지 않을 것.

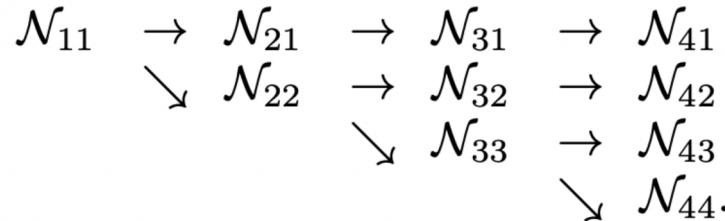
- **HR-Net**

- 병렬화 (parallel)
- 병렬적인 하위 네트워크들로 Multi-scale Resolution 을 그대로 유지하며 다양한 scale 의 spatial 정보 학습
- 병렬적인 하위 네트워크 간 학습 정보를 공유하게 해, Multi-scale 의 spatial 정보를 더욱 풍부하게 학습할 수 있음



- **HR-Net**

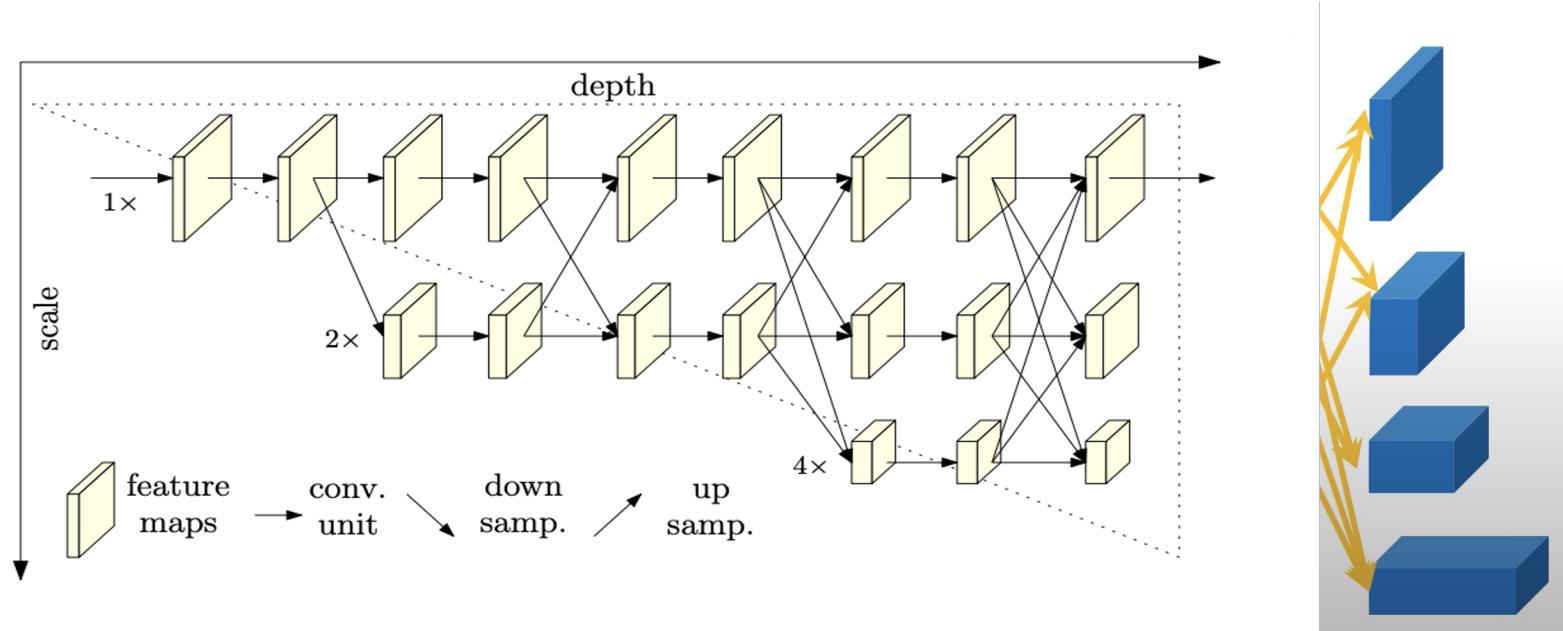
- 하나의 row 안에서 resolution 이 모두 동일하게 구성
- Row 가 밑으로 내려갈수록 high-to-low resolution 네트워크를 통해 down-sampling
- Resolution scale 은 동일하면서, scale 별로 resolution 유지
- 다양한 scale 정보 학습 가능
- 각 row 의 high network 는 수평, 병렬적으로 forward 진행



- N_{sr} 에서 앞자리는 stage, r은 downsample된 단계를 의미한다. \
- N_{sr} 은 첫번째 subnetwork(N_{11})의 해상도의 $\frac{1}{2^{r-1}}$
- high-resolution subnetwork을 처음 stage로 시작한다.
- high-to-low resolution subnetworks을 하나씩 추가한다.

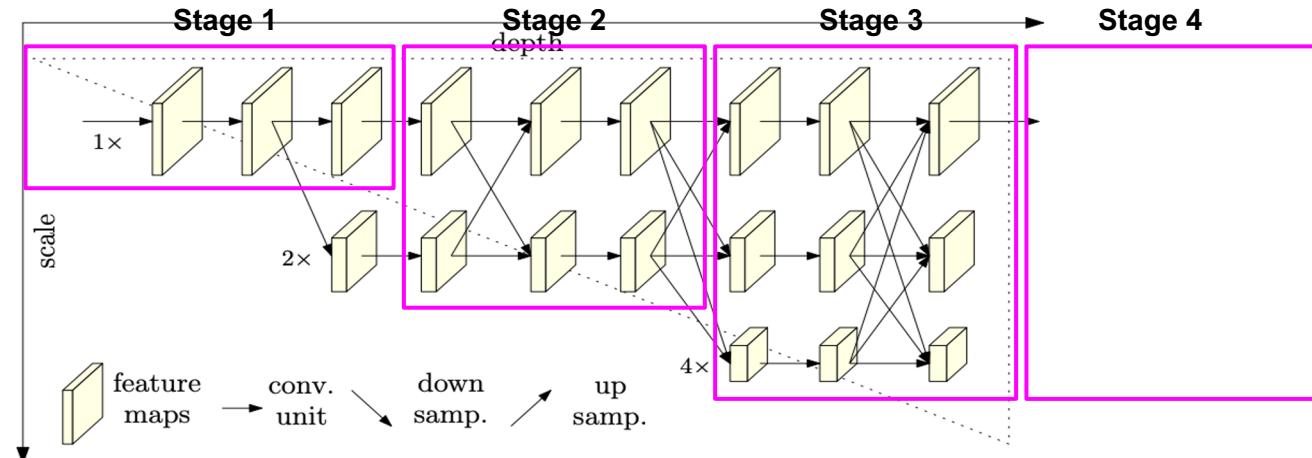
- **HR-Net**

- Forward 만 진행하면 다양한 scale 정보가 따로 놀게 됨.
 - 따라서 forward 과정과 더불어 scale 정보를 fusion 될 수 있도록 함.
- 이러한 과정을 통해서 마지막 stage 에서는 각 resolution 별로 중요한 global, local 정보들이 풍부하게 담길 수 있음.



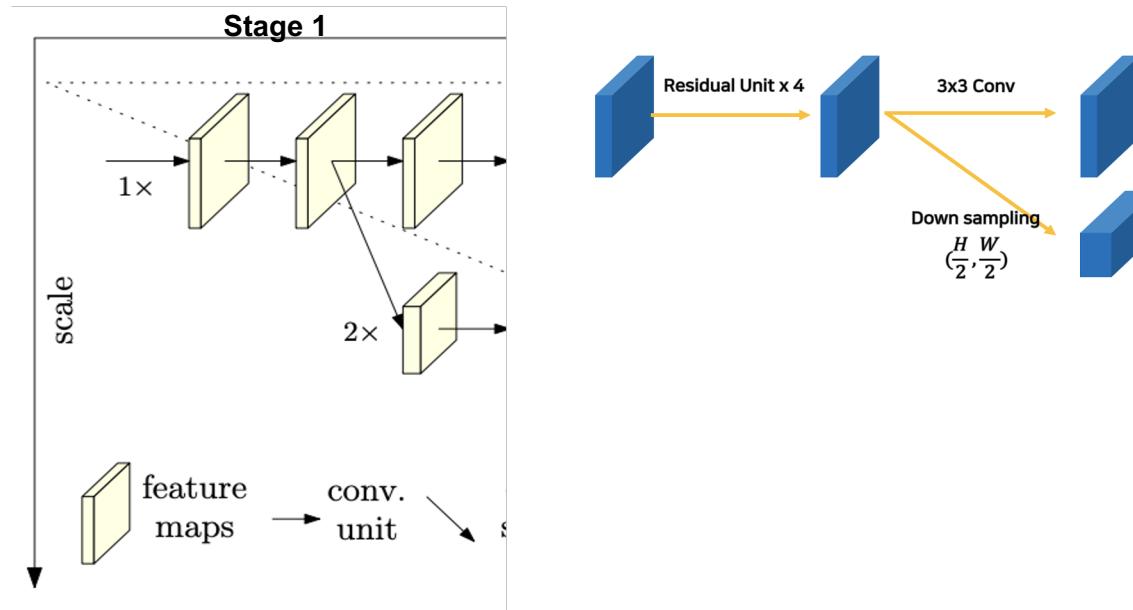
- **HR-Net**

- 병렬적으로 구성된 Sub-네트워크 간 Fusion을 통해, 상단의 High Resolution을 유지하며 Global, Local 정보 학습
 - Low resolution : 넓은 receptive field로 상대적으로 풍부한 semantic information을 가짐
 - High resolution : positional information이 많이 살아 detail한 정보를 가짐
- HR-Net은 4개의 stage로 구성되어 있으며, 이를 서로 병렬적으로 fusion



- **HR-Net**

- **Stage 1**
 - 하나의 resolution 유지
- 병렬적으로 구성된 Sub-네트워크 간 Fusion을 통해, 상단의 High Resolution 을 유지하며 Global, Local 정보 학습

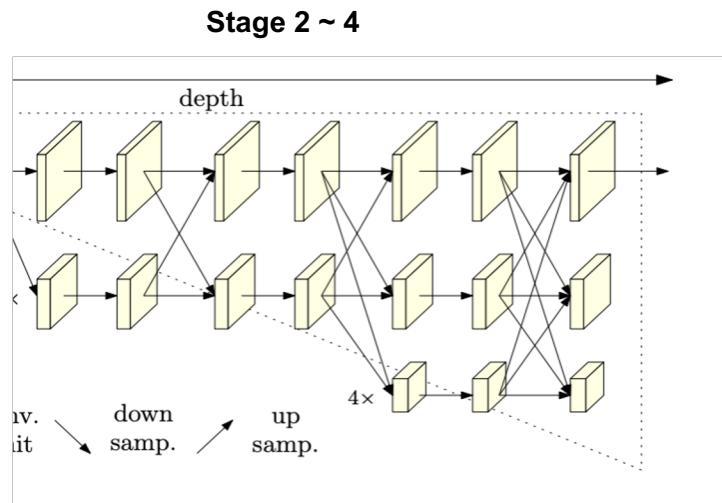


- **HR-Net**

- **Stage 2~4**

- 본격적으로 fusion 과 transition 진행

- 병렬적으로 구성된 Sub-네트워크 간 Fusion 을 통해, 상단의 High Resolution 을 유지하며 Global, Local 정보 학습



1. Multi-resolution 간 Fusion

- a. Exchange Unit

- i. Down-sampling (halve) 3×3 Conv (Stride=2, Padding=1)
- ii. Up-sampling (double) Nearest-neighbor Up-sampling (* 2)

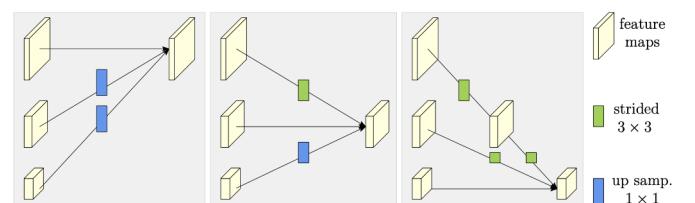


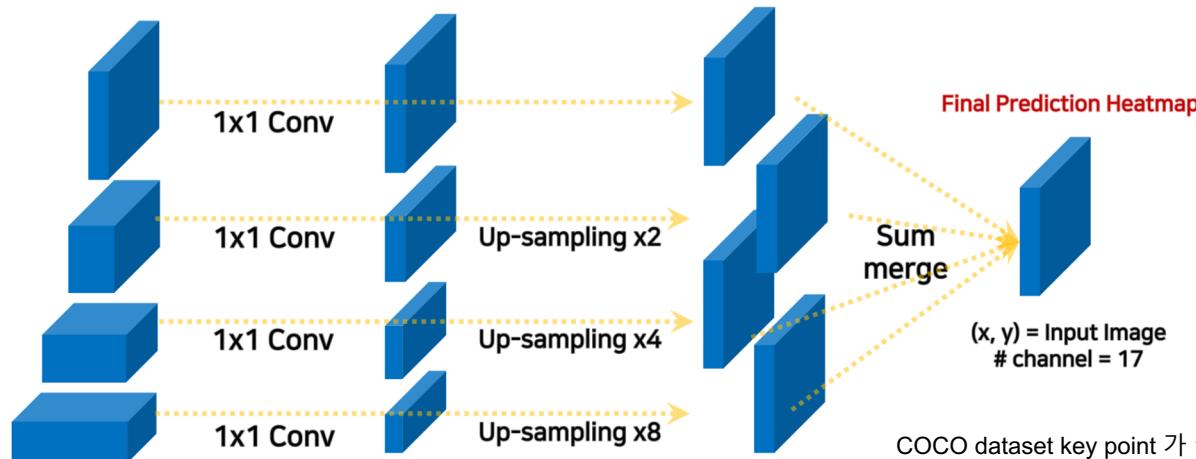
Figure 3. Illustrating how the exchange unit aggregates the information for high, medium and low resolutions from the left to the right, respectively. Right legend: strided 3×3 = strided 3×3 convolution, up samp. 1×1 = nearest neighbor up-sampling following a 1×1 convolution.

1. Resolution scale 확장하는 Transition

- a. Down-sampling : 가로, 세로 절반의 resolution

- **HR-Net**

- Stage 4
 - 원본 이미지 shape에 맞게 upsampling
 - 최종 resolution에 맞춰준 후에, merge하여 final prediction heatmap 생성
 - 최종 heatmap은 (x, y, 17) dimension
 - 마지막 exchange unit으로부터 나온 high-resolution representations output으로 regress
- Loss function은 MSE(평균 제곱 오차)
- GT heatmap은 각 keypoint에 2D 가우시안 분포를 적용해서 구함



COCO dataset key point 17개
i.e. 하나의 heat map이 하나의 key point 관절에 대한 예측값

- 데이터셋

- COCO Keypoint Detection
 - 학습 데이터 : 57,000 images (person instances: 150k)
 - Evaluation : 5,000 Images (val), 20,000 Images (test-dev)
 - # Key-points : 17

- 평가 지표

- 유사성 측정 지표 : OKS (=Object Keypoint Similarity)
 - 키 포인트 유사성 측정을 위함
- 평가 Metric : AP (=Average Precision)

- 평가 지표

- 유사성 측정 지표 : OKS (=Object Keypoint Similarity)
 - 키 포인트 유사성 측정을 위함
 - 0 (Worst) ~ 1 (Best) 사이의 값

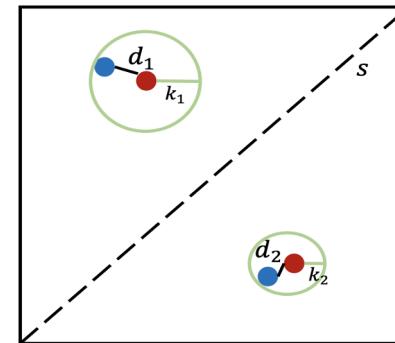
$$\frac{\sum_i \exp(-d_i^2/2s^2 k_i^2) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}$$

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



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

- Average Precision

- OKS Threshold에 따른 Precision, Recall 값
- Precision-Recall curve

- Results

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

- Simple Baseline 대비 더 가볍고, 연산량이 적지만, 높은 모델의 성능 (Small HR-Net, HRNet-W32)
- CPN 계열 모델 대비, 각각 AP 4.6, 4.0 포인트 성능 향상, 다만 HR-Net 의 복잡도 크기는 다소 증가
- 이전 SOTA 인 Simple Baseline 의 small, large 모델 대비 모두 적은 parameter 와 연산량으로 더 높은 성능 달성
- Pre-train 모델 사용할 경우, 모델의 AP 1 point 상승
- HR-Net 의 사이즈 증가 시, AP 1 point 를 비롯해 전반적으로 다소 성능 지표 상승

- Results

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

- Bottom-up Approach 보다 높은 정확도 (AP)
- SimpleBaseline 대비 더 작은 모델 사이즈와 적은 연산 복잡도로 더 높은 성능 달성
- 추가 데이터 (AI Challenger) 훈련 시, 가장 높은 AP (77.0) 달성

- Qualitative results



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.

- **Ablation Study**

- Repeated multi-scale fusion
 - Fusion 의 정도가 증가할수록, 정확도 (AP) 높아짐 (비례)
 - Fusion 의 정도
 - Model C > Model B > Model A

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

- **Ablation Study**

- 다양한 resolution 의 heatmap 각각의 성능 관찰
 - 여러 resolution 별로 heatmap 성능 확인
 - Resolution 이 커질수록 성능이 비례해서 증가
 - 가장 작은 resolution 성능은 AP 가 10 이하이기 때문에 표기하지 않음.
 - **Resolution impact the key-point prediction quality**
- Input size 에 따른 모델의 성능 향상 관찰
 - 이미지 size 별로 baseline, HR-Net 성능 비교
 - Image size 가 작을수록 성능 향상 폭이 큼. 지속적인 성능 향상을 보임.
 - **High resolution 을 유지하는 것이 중요함.**

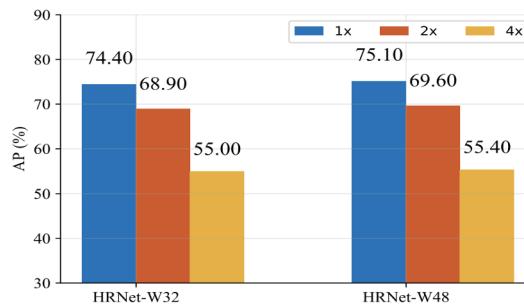


Figure 5. Ablation study of high and low representations. $1\times$, $2\times$, $4\times$ correspond to the representations of the high, medium, low resolutions, respectively.

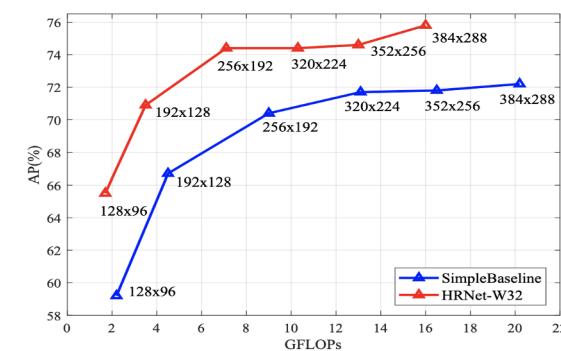
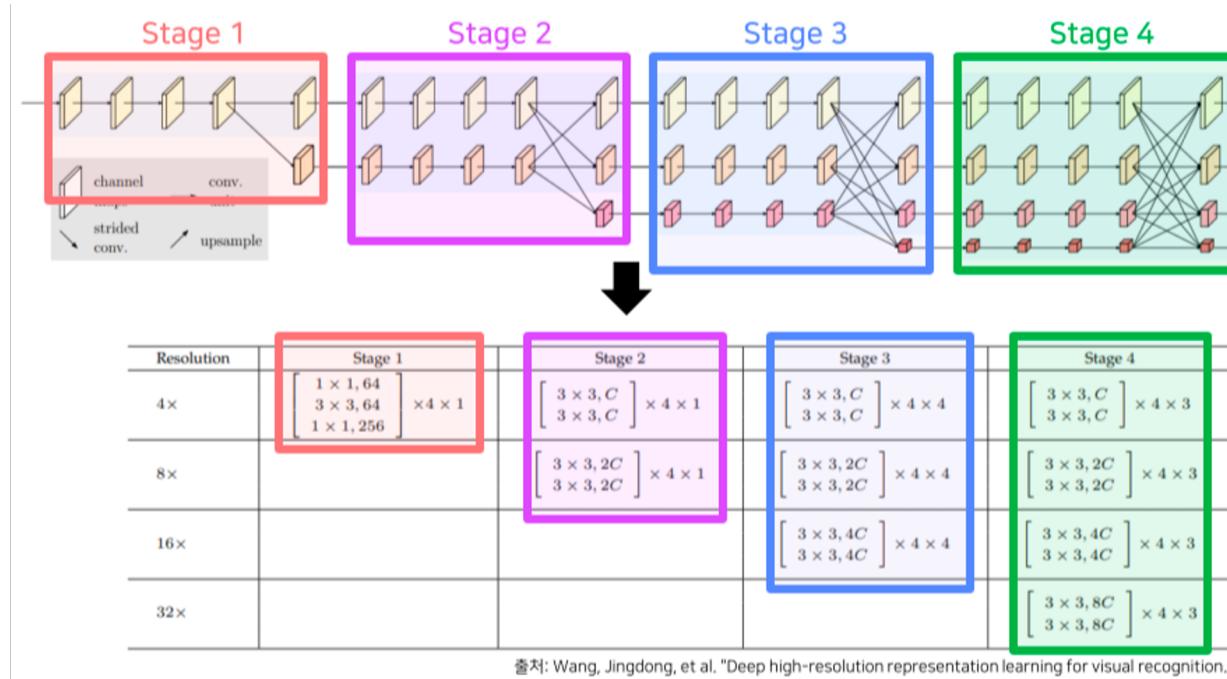


Figure 6. Illustrating how the performances of our HRNet and SimpleBaseline [72] are affected by the input size.



- 복원 (Low resolution \rightarrow High resolution) 없이, 모든 과정에서 High resolution 유지
- Multi-resolution 표상을 반복적으로 fusion 하여, 보다 신뢰할만한 High Resolution Representation 생성

Research Trends

- **Confluence**
 - <https://hutomdev.atlassian.net/wiki/spaces/COM/pages/2564620328/2022+Research+Trends>
- Arxiv Sanity: 가장 최신의 논문을 볼수 있는 최대 저장소
- SCI-HUB: DOI입력시 유료논문 공짜열람
- Arxiv : 논문보자!! 여기서 Artificial Intelligence, Computer Vision and Pattern Recognition 두가지 관심분야 보자!
- Paper With code: 논문뿐만 아니라 코드도 한번에 볼수 있는 곳
- [labml.ai](#): 아카이브 대체 trends 사이트? pytorch Implementation ⇒ 수식으로 변환 하는사이트도 있는 듯?
- [semantic sanity](#): paper recomendation
- [benty-fields](#): research pool ? paper, 세미나, 저널클럽 등 ?
- [shortscience.org](#): paper 요약 사이트
- [deepmonitoring.org](#): 트위터 기반 paper trends. keyword 기반 검색 가능

2022 Research Trends



Created by 지현규

Last updated: Jun 20, 2022 • 1 min read • Add Workflow • 5 people viewed

- Purpose

- 연구 survey 동향 파악 및 분석

- Description

- 정리방법

- arXiv sanity 에 기재된 연구 survey의 top recent 조사
 - labml.ai - weekly papers 에 댓글 및 하트가 많거나, OUR PICK 논문 위주로 조사
 - [Paper With code](#), [labml.ai](#), [deepmonitoring.org](#) 를 이용하여 현재 팀내에서 진행하는 연구분야 및 관심있는 monitoring keyword를 기준으로 논문조사

- Weekly(2주간 10개), monthly 간격으로 업로드

- Monitoring Keyword (monthly 와 함께 업로드)

- Columns

- Paper title: 연구 survey title
 - Research group: 연구 survey research group
 - Correspondence: 연구 survey correspondence
 - Link: link

- 전체 survey list

- 해당 기간 동안 arXiv sanity에 업로드 된 전체 survey ([link](#))

End of the Document