

Ch3. HRNet

Deep High-Resolution Representation Learning for Human Pose Estimation

Human Pose Estimation

Single Person Pose Estimation,
Multi Person Pose Estimation

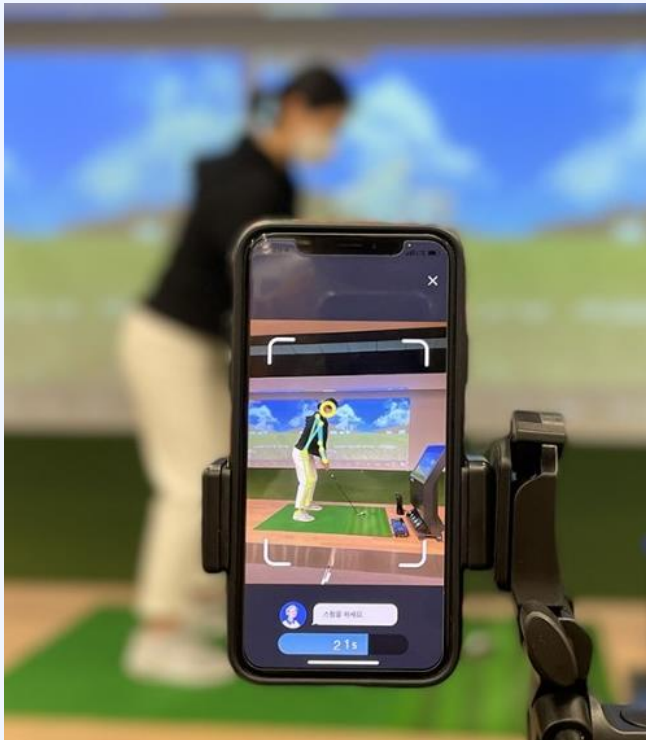
Human Pose Estimation

- Pose Estimation
 - 주어진 영상 속 Human Object의 자세(pose)를 추정하 것
 - Key-points detection, Pose recognition
 - 특정 Pose를 만들어내는 Key-points들을 찾아내는 task



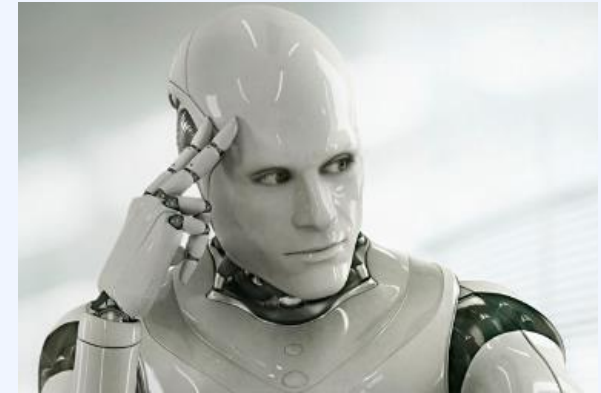
Human Pose Estimation

- 인간의 눈을 대체할 수 있는 시각 기능
- 시각 기능을 통해 획득한 자세에 대한 정보
- 올바른 자세에 대한 지식



이미지 입력

AI 코칭



Human Pose Estimation 활용

- Animation 제작
- 게임, 아바타 동작(메타버스)
- 스포츠 영상 분석을 통한 서비스
- Medical assistance
- CCTV

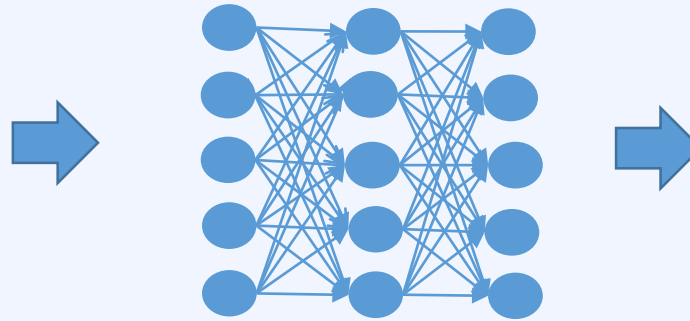


Human pose estimation Task

- Input Image 상에서 Key points의 (x,y) 좌표 값 예측



Input image

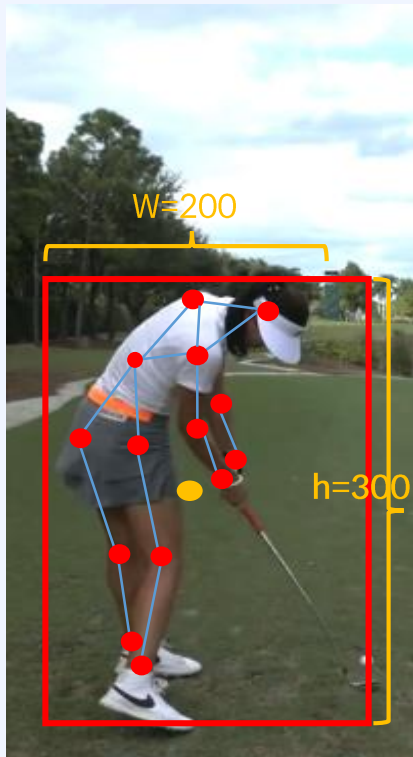


Pose Estimation Network



Human pose estimation 입력 및 출력

- Human pose estimation 입력 및 출력 데이터 예시
 - 입력 데이터: RGB 이미지
 - 출력 데이터: 사람의 Bounding Box와 탐지하고자 하는 point 좌표(x, y)



- 사람에 대한 Bounding Box(x,y,h,w)
 - Box의 중심: $(x,y) = (120,150)$
 - 높이와 너비:h,w
- 탐지하고자 하는 point 개수 : 14개
 - 머리 $(x,y) = (120,370)$
 - 왼쪽 어깨 $(x,y)=(120,330)$
 - 왼쪽 팔꿈치 $(x,y)=(100,330)$
 - .
 - .
 - .
 - .
 - .
 - .
 - .
 - .
 - .
 - .
 - .
 - 오른쪽 발목 $(x,y)=(50,70)$

Human Pose Estimation

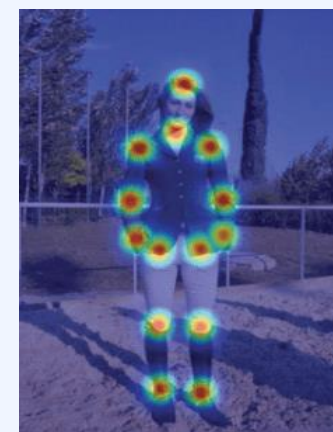
Single Person Pose Estimation

Single person pose estimation

- 입력 이미지 내 사람 한 명만 존재하는 경우
- Direct regression
 - 관절 별 좌표를 예측
- Heatmap based estimation
 - 특정 관절이 존재할 만한 곳을 Heatmap 형태로 출력



Direct regression

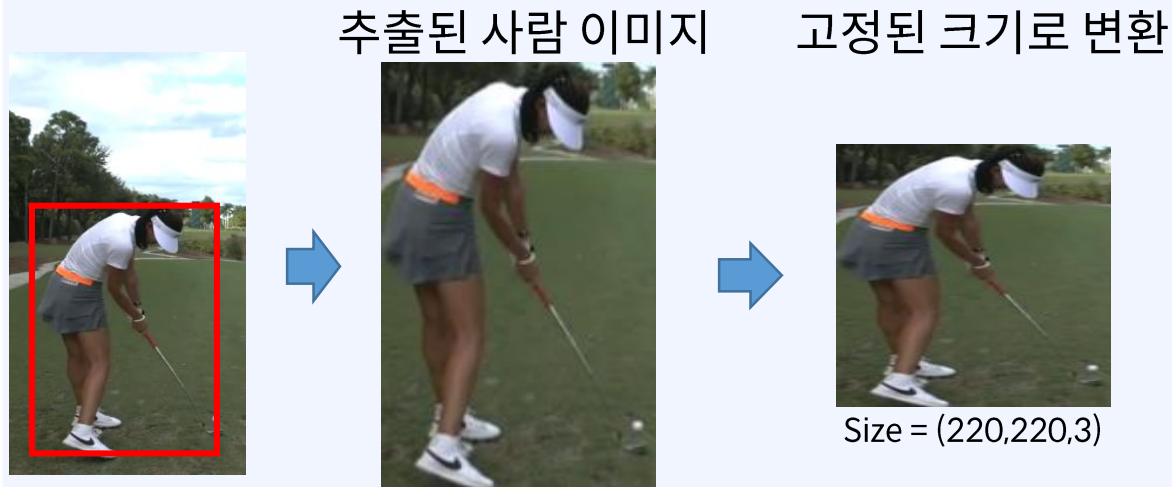


Heatmap based estimation

Direct regression

- DeepPose 모델 – 데이터 입력에서 예측 과정
 - Bounding Box를 사용해 사람이 존재하는 영역만 추출
 - Bounding Box(x,y,w,h) = (120,150,200,300)
 - 관절 별 좌표는 추출 전 이미지 내 좌표 → 변환 필요

- 머리에 대한 좌표 변환
 - 변환된 x좌표 = $\frac{1}{200} (120 - 120)$
 - 변환된 y좌표 = $\frac{1}{300} (370 - 150)$



	X좌표	Y좌표	변환된 x좌표	변환된 y좌표
머리	120	370	0	0.73
왼쪽 어깨	120	330		
왼쪽 팔꿈치	100	330	-0.1	0.60
왼쪽 손목	80	350		
...		
오른쪽 발목	50	70		

Direct regression

- DeepPose 모델 – 데이터 입력에서 예측 과정
 - Bounding Box를 사용해 사람이 존재하는 영역만 추출
 - Bounding Box(x,y,w,h) = (120,150,200,300)
 - 관절 별 좌표는 추출 전 이미지 내 좌표 → 변환 필요

$$N(y_i, b) = \begin{pmatrix} \frac{1}{w} & 0 \\ 0 & \frac{1}{h} \end{pmatrix} (y_i - b_c)$$

- i : 관절 인덱스($i=1,2,\dots,K$)
- w : Bounding box 너비
- h : Bounding box 높이
- y_i : i 번째 관절에 대한 좌표
- b : Bounding box
- b_c : Bounding box 중심 좌표

	X좌표	Y좌표	변환된 x좌표	변환된 x 좌표
머리	120	370	0	0.73
왼쪽 어깨	120	330	0	0.60
왼쪽 팔꿈치	100	330	-0.1	0.60
왼쪽 손목	80	350	-0.2	0.67
...
오른쪽 발목	50	70	-0.35	-0.27

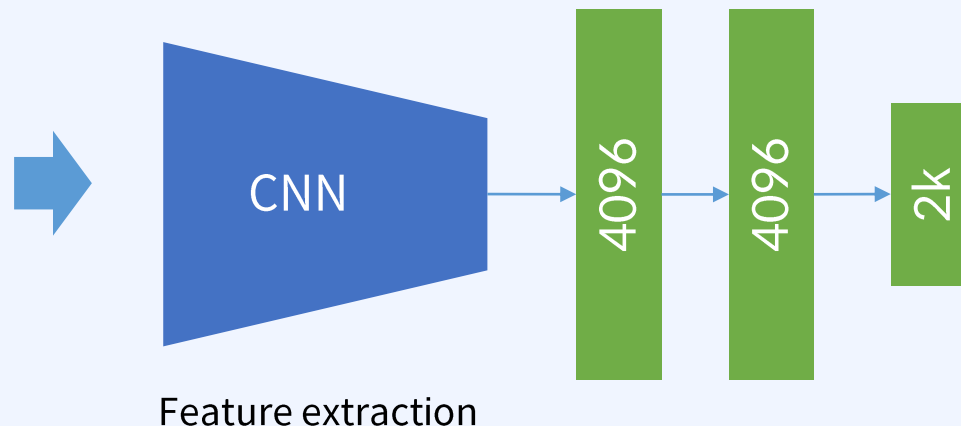
Direct regression

- 처리된 입력 데이터를 이용해 예측값을 산출
- Feature Extractor를 사용해 Representation 벡터 산출
- 해당 벡터를 Fully connected layer에 입력하여 관절별 예측 값 산출
 - 2k:k개의 관절별(x,y)

$$MSE = \frac{1}{2k} \sum_{i=1}^{2k} (y_i - \hat{y}_i)^2$$



Size = (220,220,3)



	예측 (\hat{y}_i)	실제(y_i)
1	0.02	0
2	0.01	0.73
3	-0.03	0
4	0	0.06
...
2k	-0.3	-0.27

Direct regression

- 앞에서 산출된 예측 관절 위치를 실제 이미지 내 위치로 역 변환
 - 오른쪽 엉덩이 위치에 예측 값(-0.3, -0.7) → 실제 이미지 내 위치로 변환
 - 왼쪽 어깨 위치 예측값(0.2, 0.2) → 실제 이미지 내 위치로 변환



Size = (220,220,3)

- 오른쪽 엉덩이에 대한 좌표 역 변환
 - $-0.3 = \frac{1}{200}(x - 120) \rightarrow x = 60$
 - $-0.7 = \frac{1}{300}(y - 300) \rightarrow y = 90$
- 왼쪽 어깨에 대한 좌표 역 변환
 - $-0.2 = \frac{1}{200}(x - 120) \rightarrow x = 160$
 - $-0.2 = \frac{1}{300}(y - 300) \rightarrow y = 360$



실제 이미지

Direct regression

- 전체 관절 예측 위치 → 관절별로 예측
 - 실제 이미지 내에서 왼쪽 어깨 예측값과 오른쪽 엉덩이 예측값 사이 거리 계산
 - $\sqrt{(160 - 60)^2 + (360 - 90)^2} = 287.92$
 - 실제 이미지에서 예측한 관절의 위치를 중심으로 하는 Bounding box 생성
 - Bounding box의 너비와 높이 = $\delta \times 287.92$



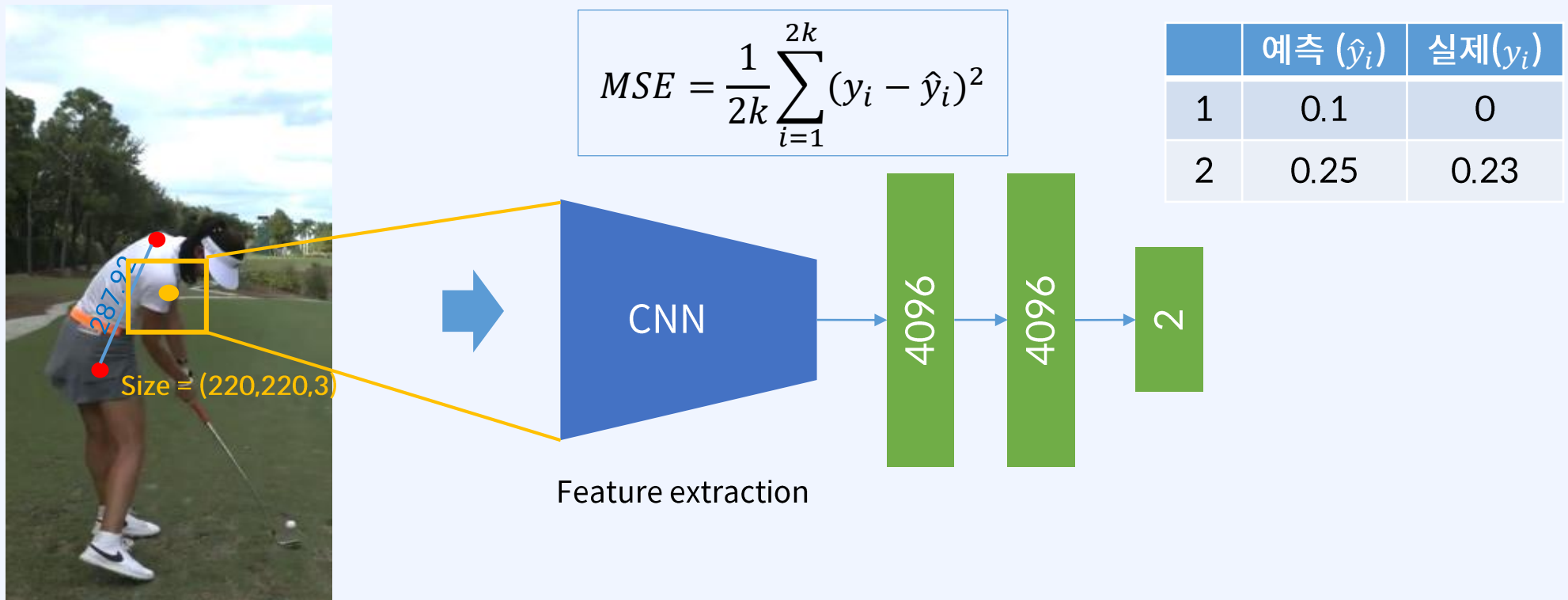
Size = (220,220,3)



실제 이미지

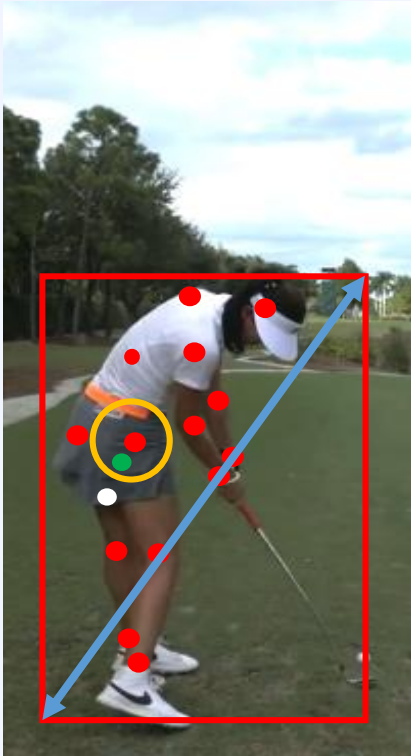
Direct regression

- Bounding box의 너비와 높이 = $\delta \times 287.92$
- 예측된 각 관절의 bounding box 부분을 추출하여 앞의 과정을 진행
- 앞에서 학습한 모델에 새로운 bounding box를 입력해 예측 및 학습



Human Pose Estimation 평가 지표

- Percent of Detected Joints(PDJ) 지표



- 사람의 **길이**(파란선)를 계산
- (특정 임계값x**길이**)의 반지름 원(노란색)을 생성
 - 반지름=임계값 X **길이** = 0.05 X 500 = 25
- 예측 위치가 원 내부에 있는지 확인
 - 원 내부에 있는 경우(녹색 점): 1(correct)
 - 원 외부에 있는 경우 (흰색 점): 0(incorrect-예측x)
- $PDJ = (\text{맞춘 개수}) / (\text{전체 관절 수})$

Human Pose Estimation

Muti Person Pose Estimation

Muti Person Pose Estimation

- 입력 이미지 내 사람이 두 명 이상 존재 하는 경우
- Top-down approach
 - 사람을 우선적으로 탐지 후 탐지 결과 내에서 관절 별 좌표를 예측
- Bottom-up approach
 - 탐지하고자 하는 관절에 대한 위치 예측 후 사람 별로 나누는 과정 진행



Top-down approach



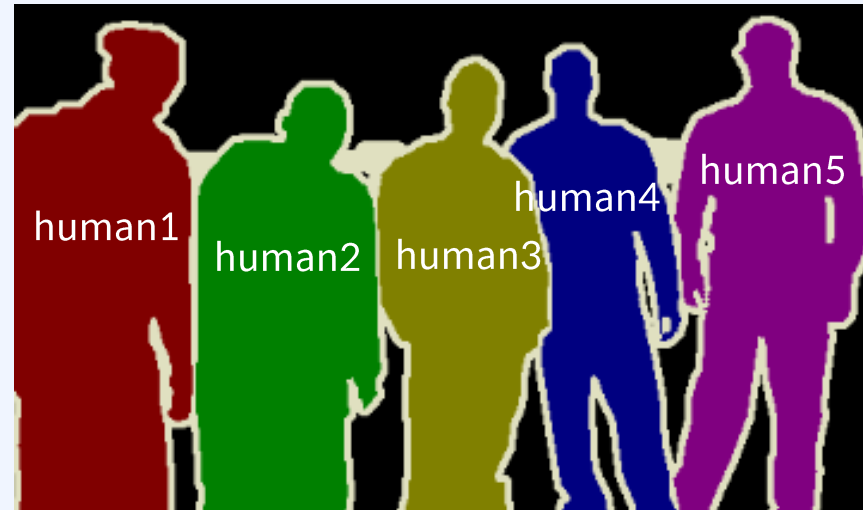
Bottom-up approach

Top-down approach

- Object Detection & Instance segmentation
 - 일반적으로 RGB image를 입력으로 사용
 - Object detection: 탐지하고자 하는 범주에 대해 bounding box regression & classification
 - Instance segmentation: 관심 있는 객체를 찾고 찾은 객체에 대해 Pixel-wise classification



Object detection

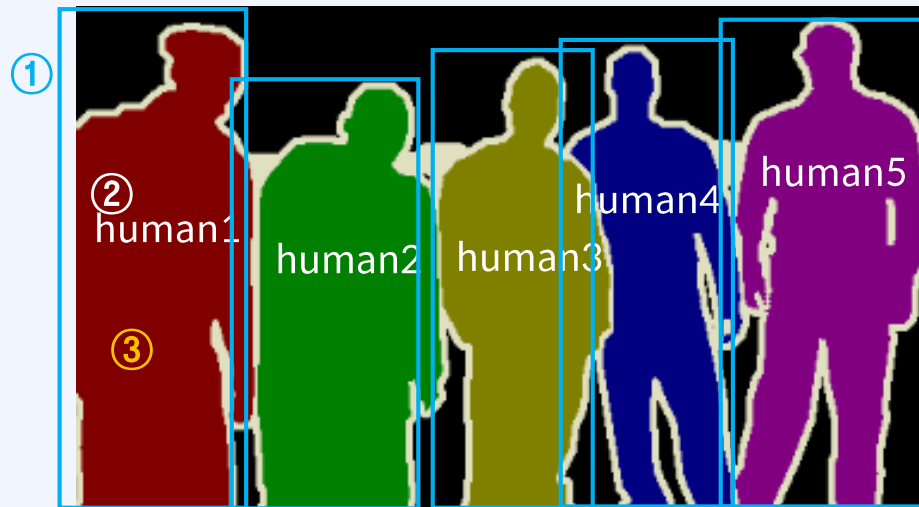


Instance segmentation

Top-down approach

- Mask R-CNN 동작

- 1) 객체가 있을만한 영역탐지(Bounding box regression)
- 2) 탐지한 영역 내 어떠한 범주가 있는지 예측(classification)
- 3) 상자 내 픽셀이 탐지한 범주인지 아닌지 분류(Segmentation, Pixel-wise classification)

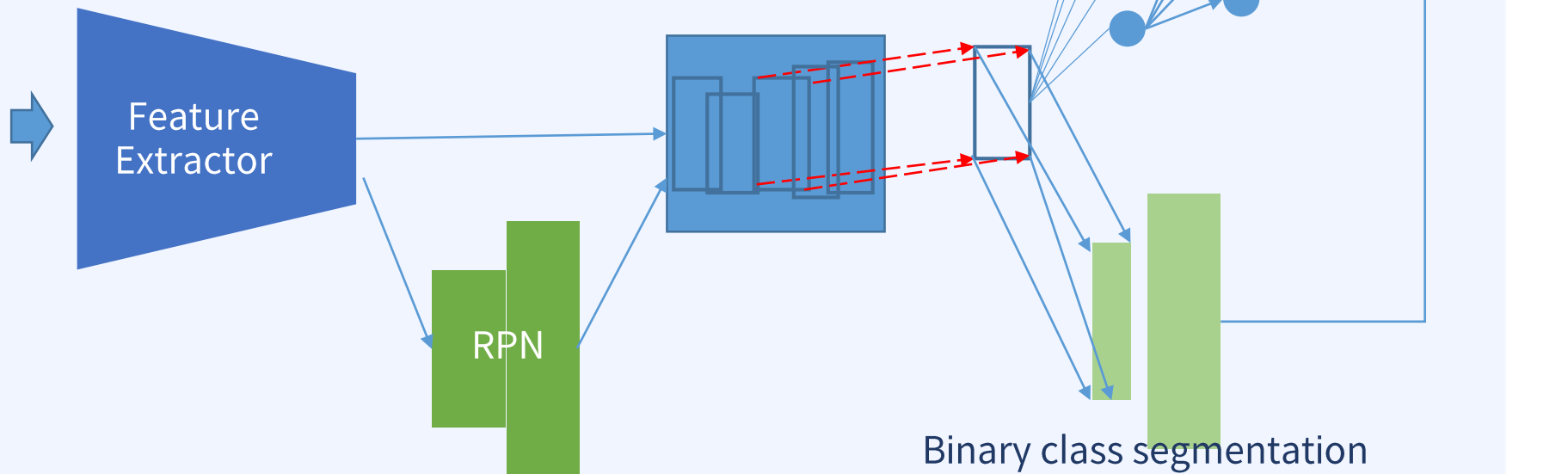


Top-down approach

- Mask R-CNN 구조 : 네 가지 모듈로 구성
 - Feature extractor(backbone)
 - Region Proposal Network(RPN)
 - Bounding box regression and Classification
 - Binary class segmentation

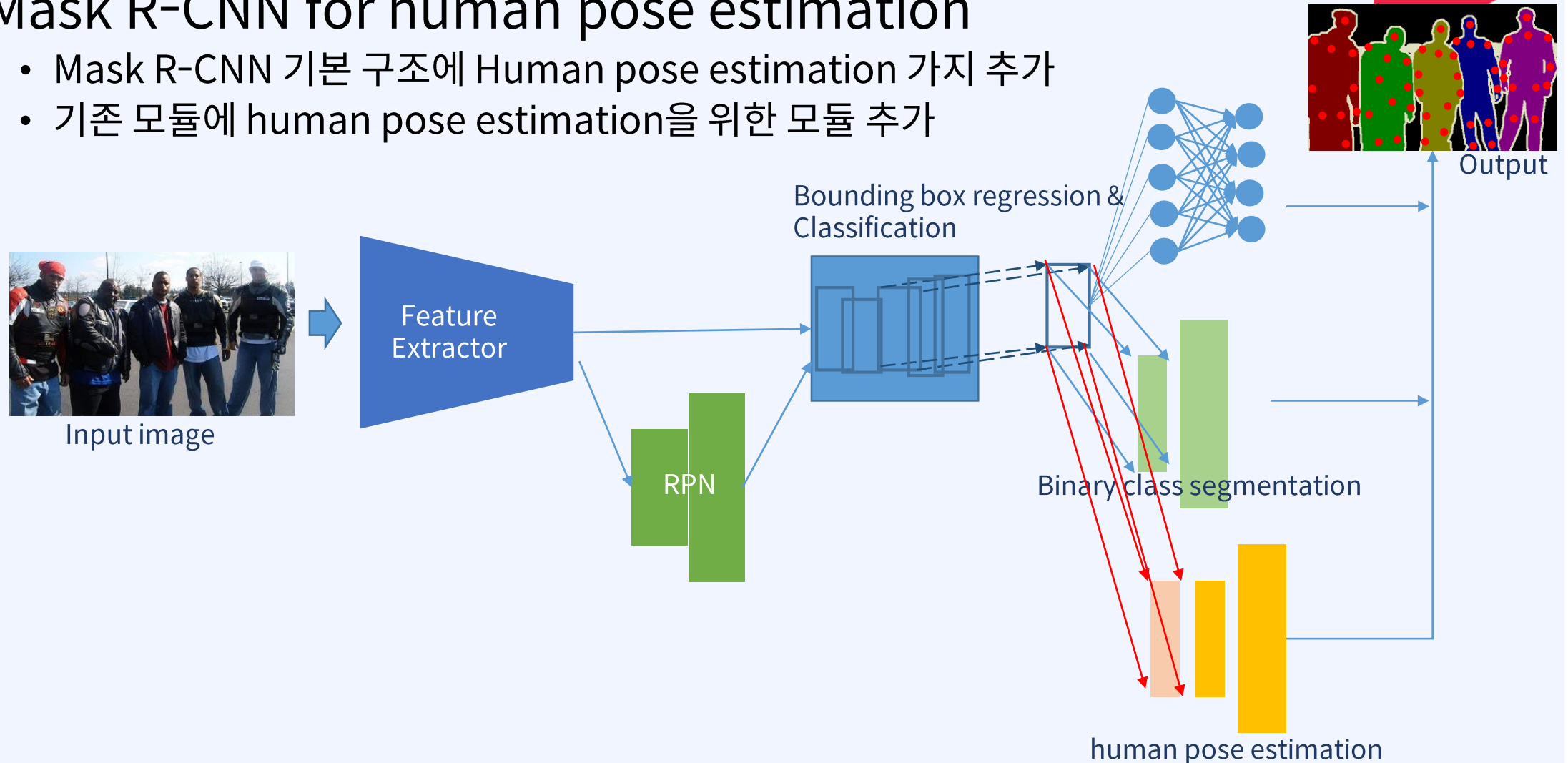


Input image



Top-down approach

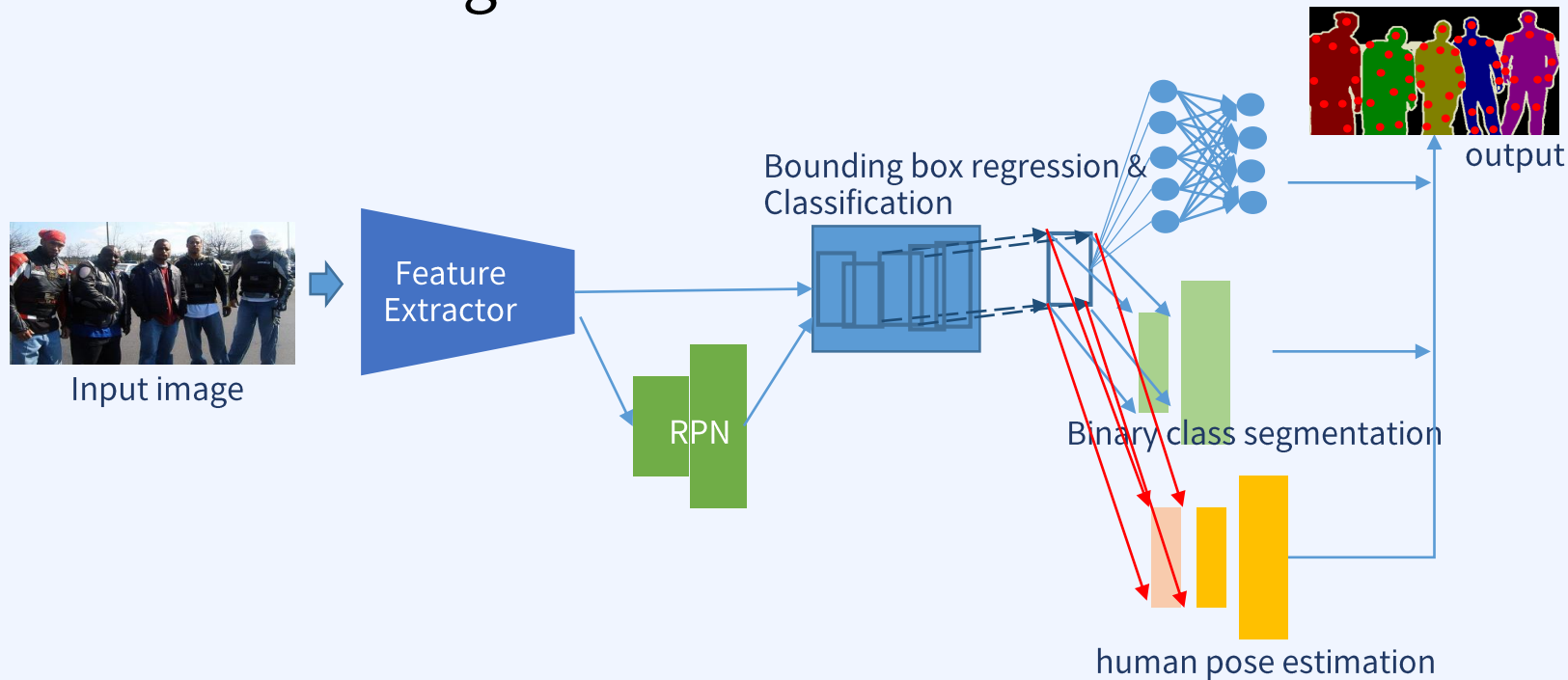
- Mask R-CNN for human pose estimation
 - Mask R-CNN 기본 구조에 Human pose estimation 가지 추가
 - 기존 모듈에 human pose estimation을 위한 모듈 추가



Top-down approach

- 손실함수

- $LOSS_{Mask\ R-CNN} = LOSS_{RPN_reg} + LOSS_{RPN_clf} + LOSS_{BB_reg} + LOSS_{BB_clf} + LOSS_{segment} + LOSS_{hpe}$
- Multi-task learning

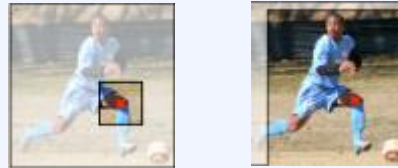


HRNet 구조와 동작

Humans pose estimation

- Humans pose estimation을 위한 목적

- Global + Local Feature 학습
- High-Resolution 복원



- Global information vs High resolution(Trade off)

<High Global Information>

Receptive Field 확대



<Low resolution>

Resolution이 낮아짐

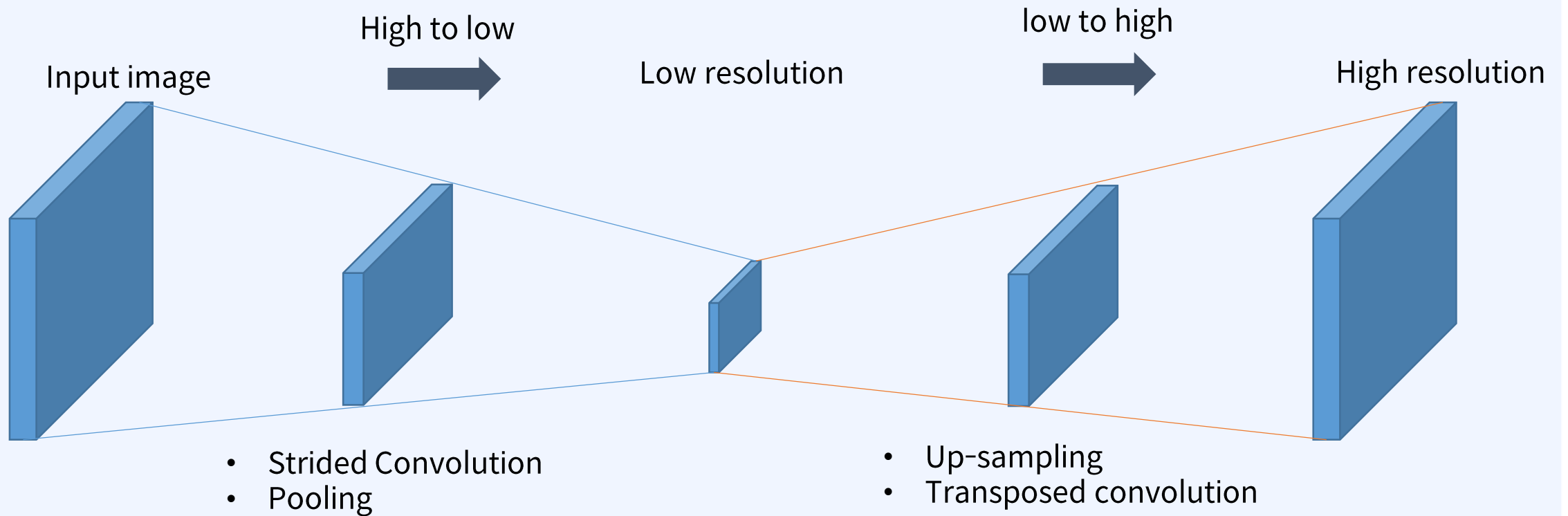
Pixel 단위 이미지 정보 상실
→ Pixel-wise prediction에
부정적인 영향



Up-sampling 시
정보 손실

기존의 접근 방식

• Simple baseline(2018)



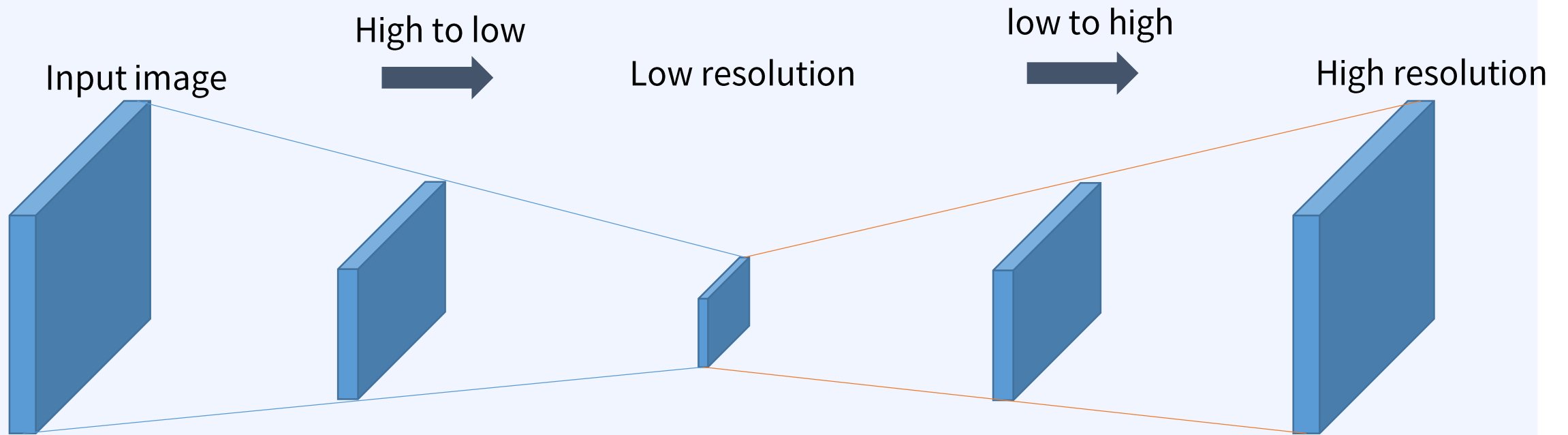
Small Object 혹은 Detail한 spatial information의 손실



Pixel-wise prediction에 부정적 영향

기존의 접근 방식

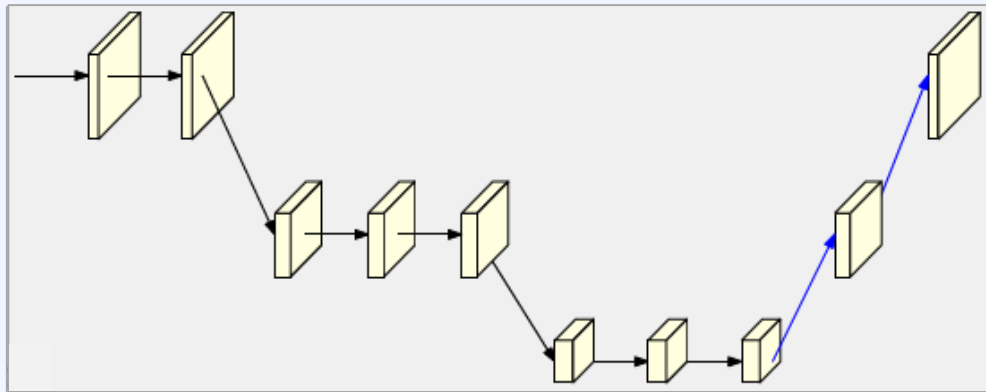
- Simple baseline(2018)



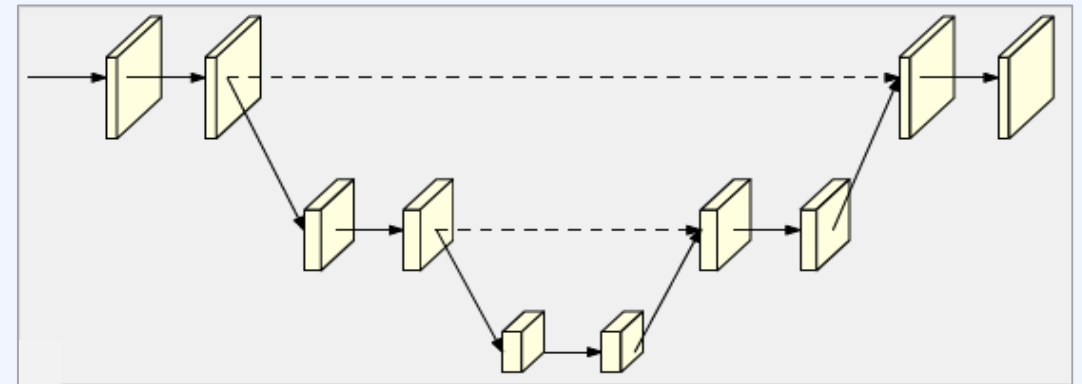
Local, Global 특징 추출과 학습의 과정의 직렬화 → Up-sampling에 과도한 의존

기존의 접근 방식

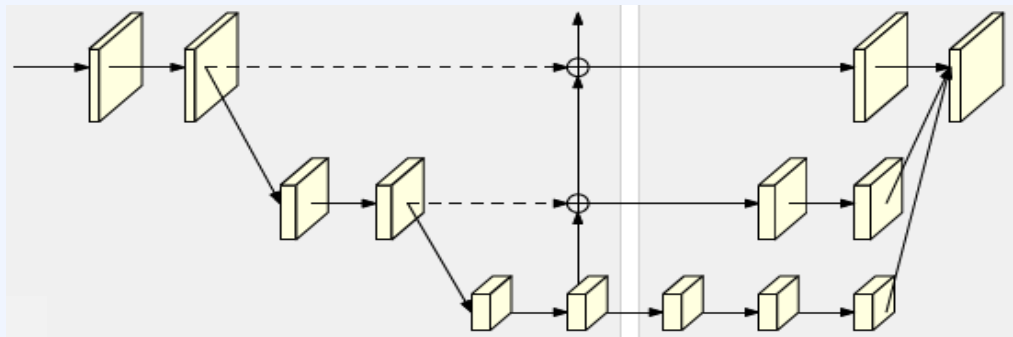
- Representative pose estimation networks



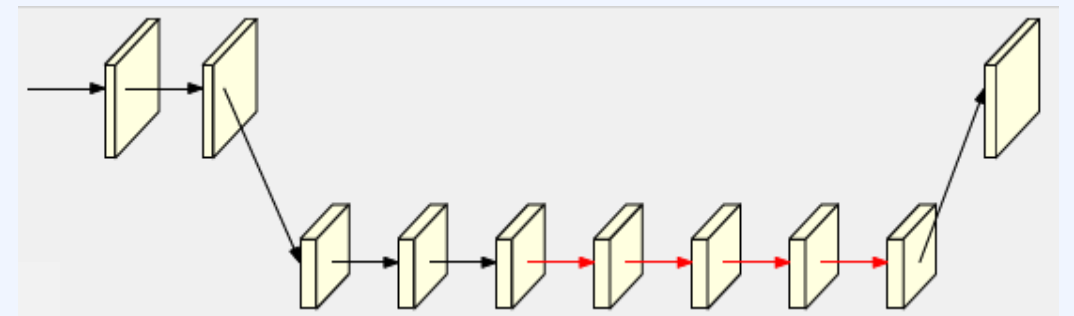
Simple baseline



Stacked hourglass



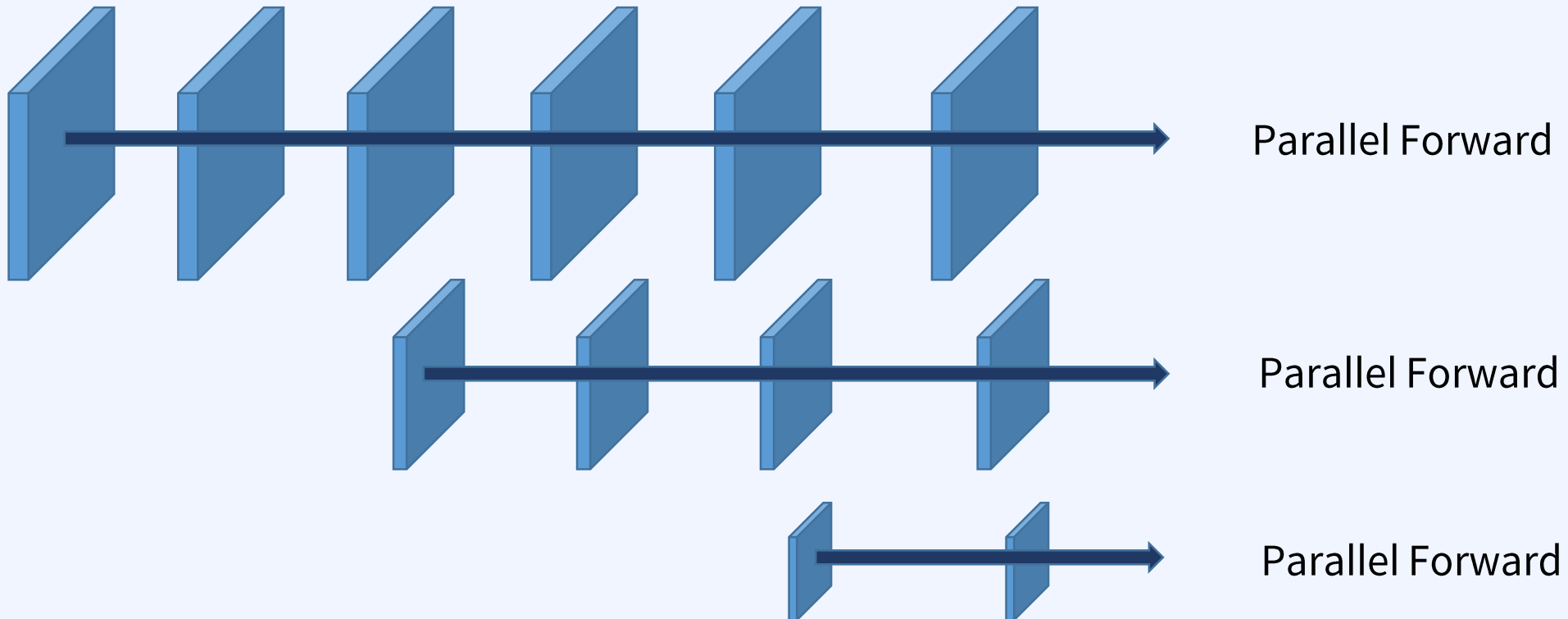
Cascaded pyramid networks



Combination with dilated convolutions

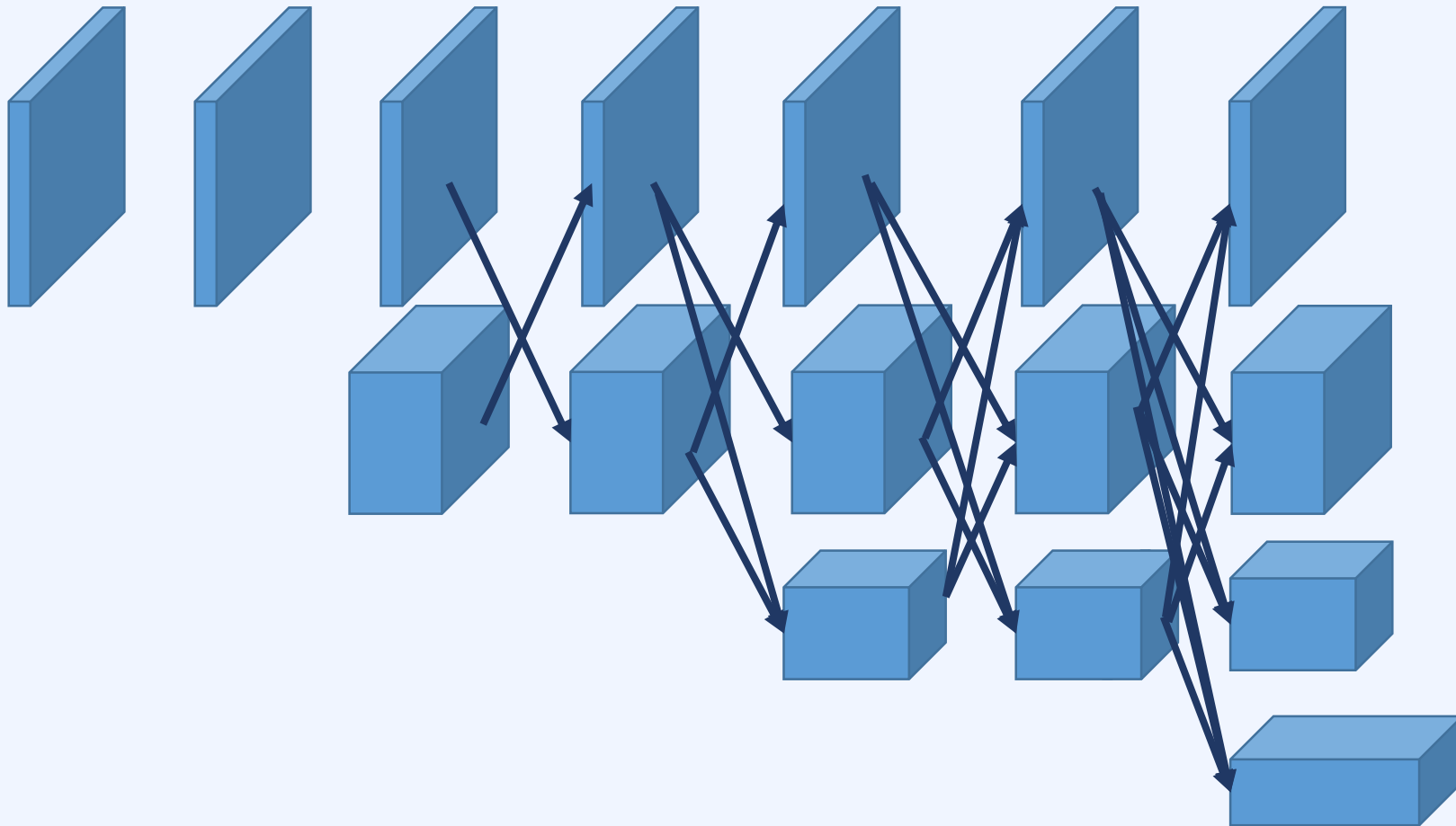
HR-Net

- Parallel
 - 병렬적인 하위네트워크들로 Multi-scale Resolution을 그대로 유지
 - 다양한 scale의 spatial 정보 학습



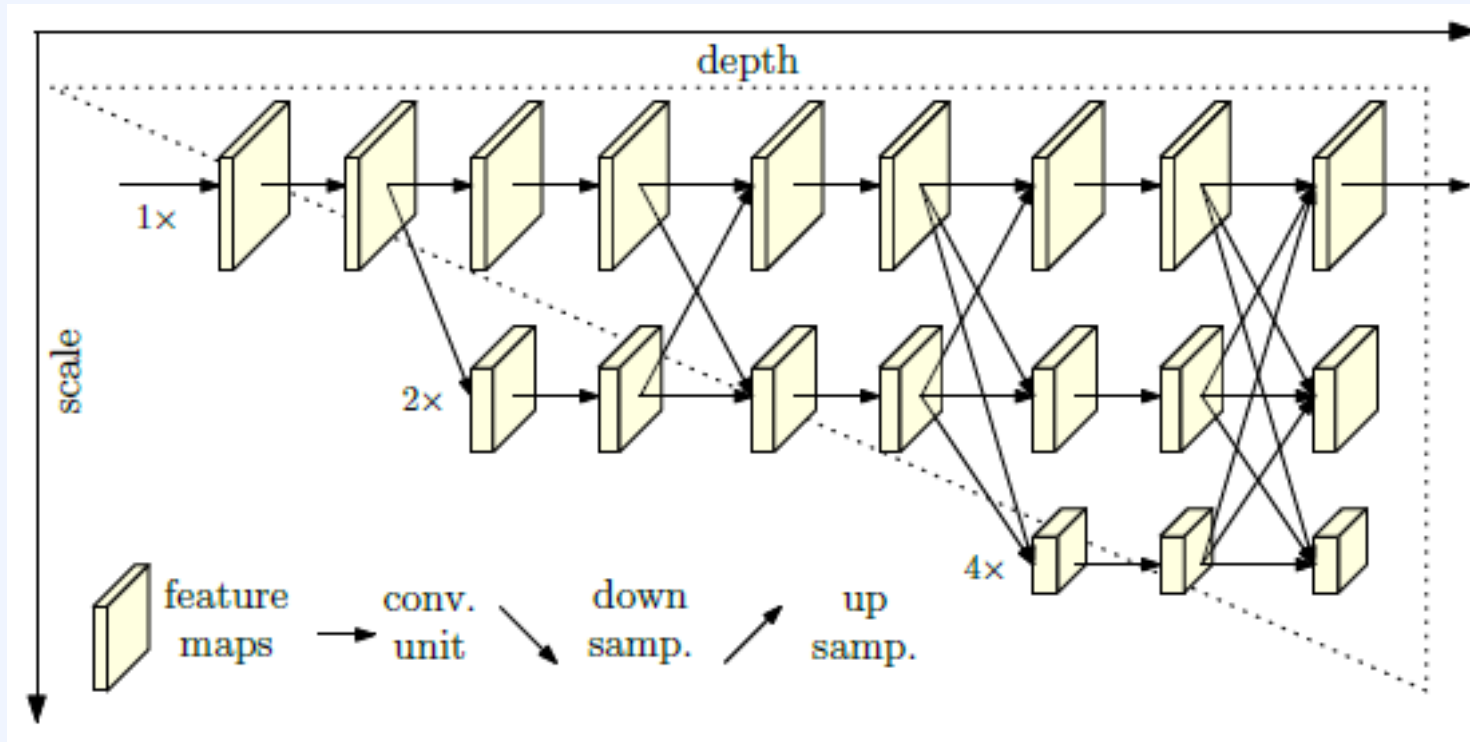
HR-Net

- Fusion of Multi-scale resolution
 - 병렬적인 하위네트워크들 간 학습 정보를 공유하여 Multi-scale의 spatial의 정보를 더욱 풍부하게 할 수 있음



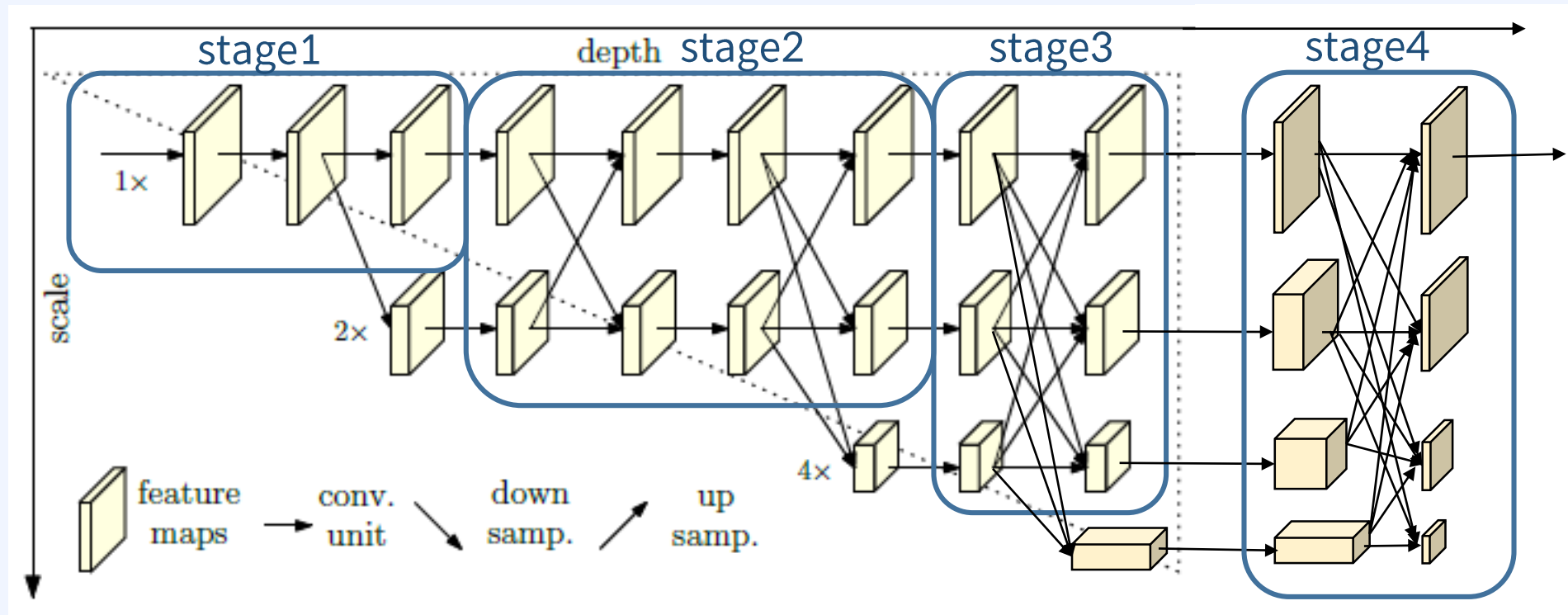
HRNet

- Fusion of Multi-scale resolution
 - 병렬적으로 구성된 Sub-네트워크간 Fusion을 통해, 상단의 High Resolution을 유지하며 Global, Local 정보 학습



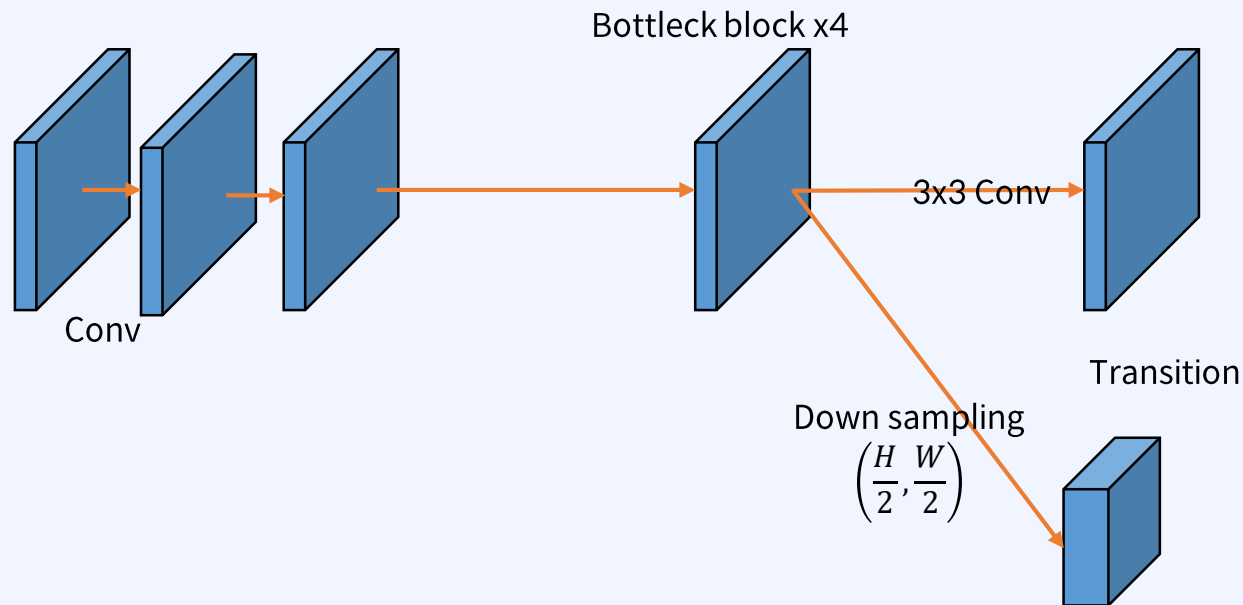
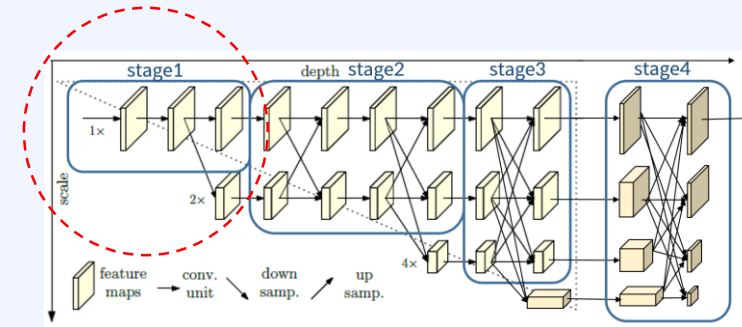
HRNet structure

- Fusion of Multi-scale resolution
 - 병렬적으로 구성된 Sub-네트워크간 Fusion을 통해, 상단의 High Resolution을 유지하며 Global, Local 정보 학습



HRNet structure

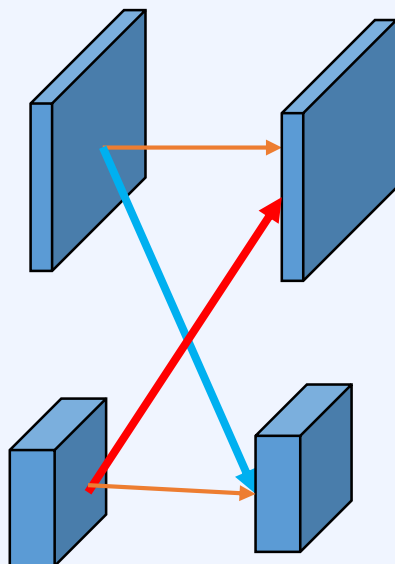
- Stage 1



HRNet structure



• Stage 2

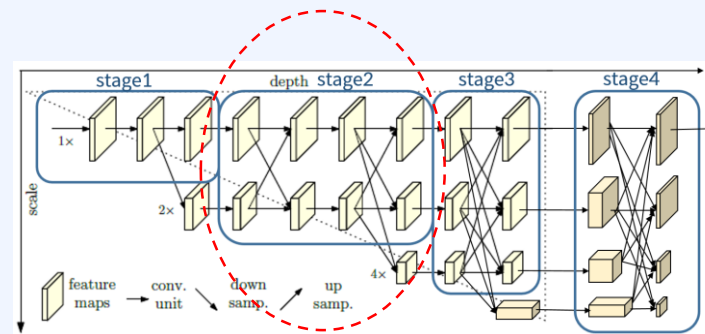
Residual block x4



• Exchange unit(Fusion)

1. Residual Block x 4
2. Exchange

-  Down-sampling(halve) 3x3 Conv(stride=2, Padding=1)
-  Up-sampling(double) Nearest-neighbor Up-sampling(x2)



Nearest Neighbor

1	2
3	4

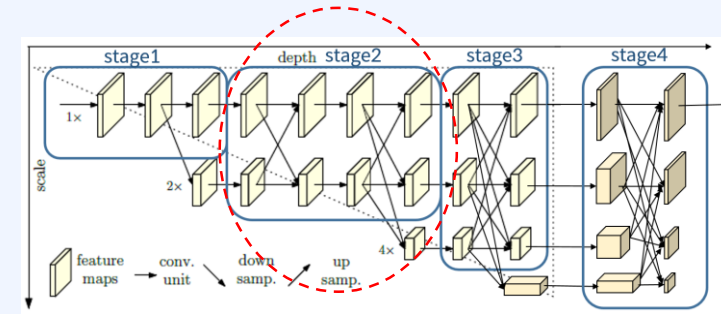
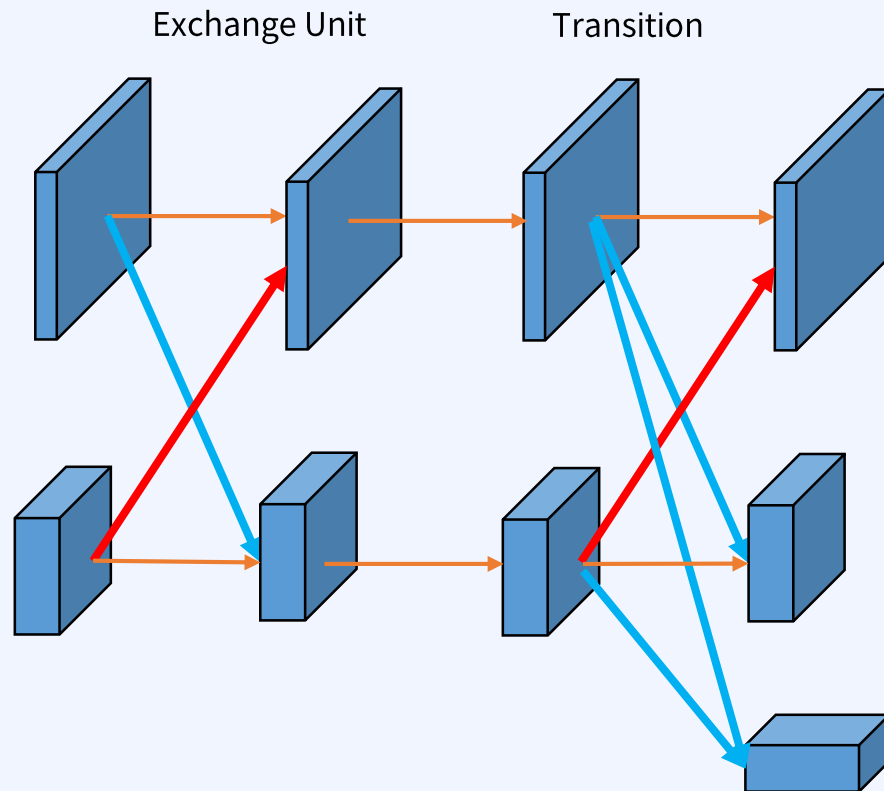
Input: 2 x 2

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

Output: 4 x 4

HRNet structure

• Stage2

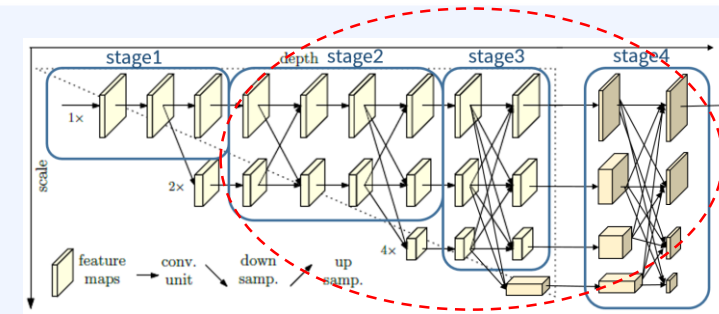


• Transition

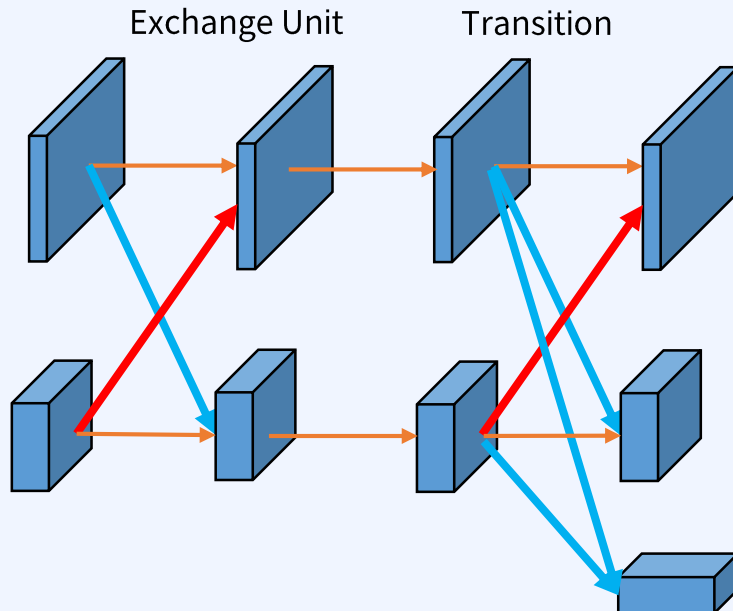
- Down-sampling(halve) 3x3 Conv(Stride=2, Padding=1)
- Up-sampling(double) Nearest-neighbor Up-sampling(x2)

HRNet structure

- Fusion of Multi-scale resolution



Stage2



Stage3

Exchange Unit x 4 Transition

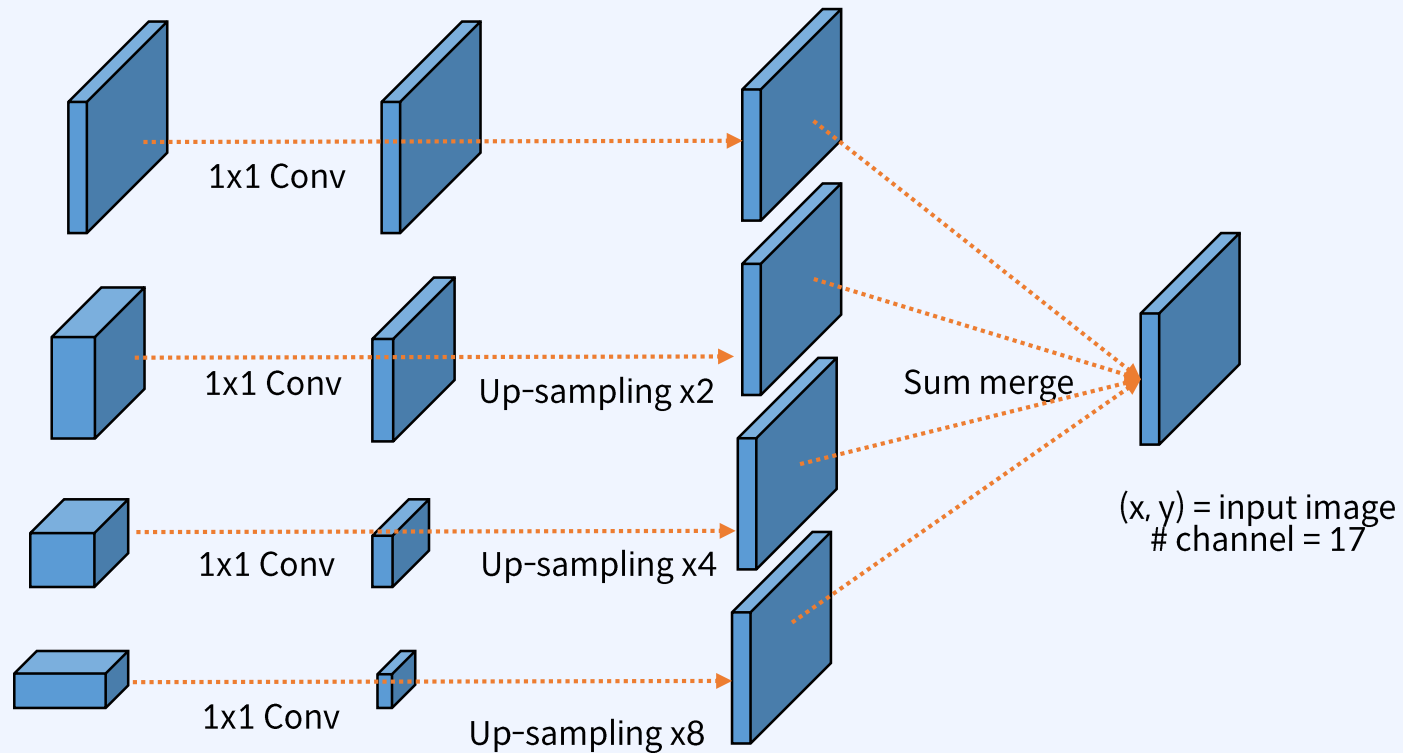
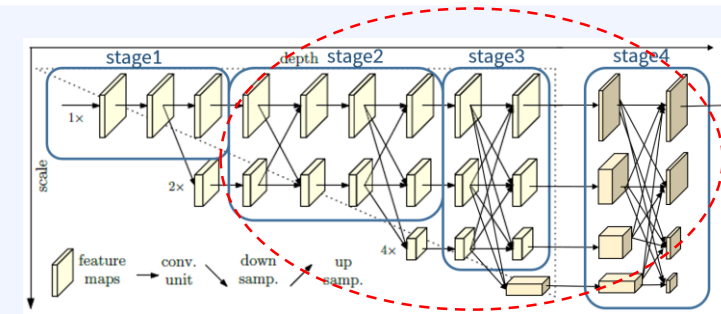
Stage4

Exchange Unit x 2

- Stage 2~4
 - Multi-resolution 간 Fusion
 - Resolution scale 확장하는 Transition

HRNet structure

- Stage 4



Summary

- Human pose estimation
 - Single Person Pose Estimation
 - Muti Person Pose Estimation
- HRNet Structure
 - Stage1,2,3,4