



Deep High-Resolution Representation Learning for Human Pose Estimation

Abstract

본 논문에서는 이전의 pose estimation의 연구들은 모두 고해상도의 이미지를 저해상도로 낮춘 후, 다시 고해상도로 높이는 방법을 사용하지 않고, 고해상도를 유지하며 다양한 스케일의 정보를 병렬적으로 처리하는 새로운 아키텍처를 제안합니다. 이를 통해서 더욱 정확한 히트맵을 통해서 다양한 밴치마크 데이터셋에서 높은 정확도를 얻을 수 있었다고 합니다.

Introduction

본 논문에서는 단일 사람의 pose estimation에 대해서 다룬다고 합니다. 그리고 최근 다양한 CNN 모델들이 좋은 성능을 달성하고 있으며, 특히 대부분의 모델들은 고해상도를 저해상도로 낮추는 방법으로 진행된다고 합니다. Transpose convolution, dilated convolution, Hourglass Module 등이 그의 예입니다.

본 논문에서는 High-Resolution Net(HRNet)을 제안합니다. HRNet은 초기부터 고해상도 표현을 유지하면서, 점진적으로 저해상도를 다루는 서브 네트워크들을 병렬로 추가합니다.

즉, 동시에, 서로 다른 스케일의 특징들을 반복적으로 융합하여, 고해상도 정보를 보강하고 저해상도 특징의 도움을 받음으로써 최종적으로 고해상도 이미지 기반의 pose estimation을 수행할 수 있는 모델을 제안합니다.

Related work

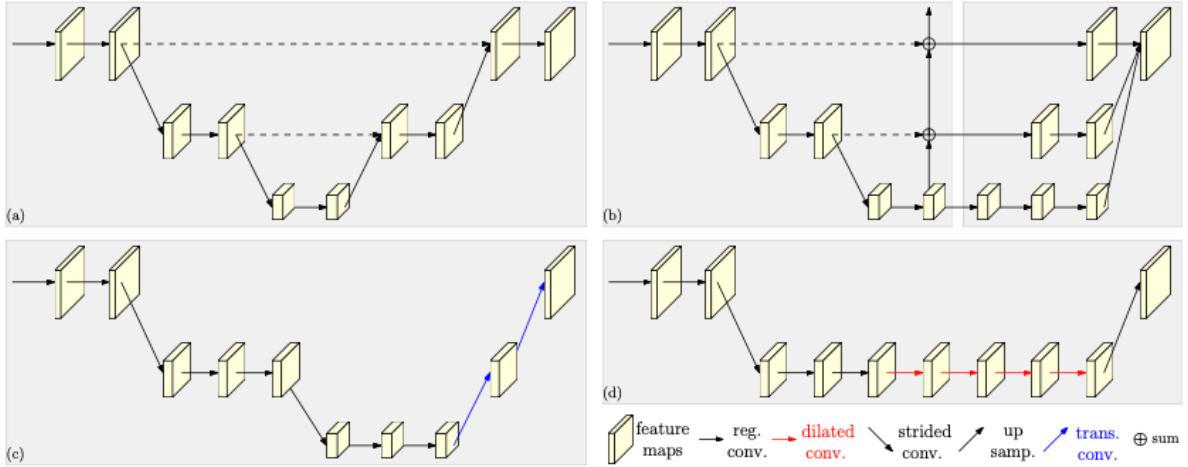


Figure 2. Illustration of representative pose estimation networks that rely on the high-to-low and low-to-high framework. (a) Hourglass [39]. (b) Cascaded pyramid networks [11]. (c) SimpleBaseline [70]: transposed convolutions for low-to-high processing. (d) Combination with dilated convolutions [26]. Bottom-right legend: reg. = regular convolution, dilated = dilated convolution, trans. = transposed convolution, strided = strided convolution, concat. = concatenation. In (a), the high-to-low and low-to-high processes are symmetric. In (b), (c) and (d), the high-to-low process, a part of a classification network (ResNet or VGGNet), is *heavy*, and the low-to-high process is *light*. In (a) and (b), the skip-connections (dashed lines) between the same-resolution layers of the high-to-low and low-to-high processes mainly aim to fuse low-level and high-level features. In (b), the right part, refinenet, combines the low-level and high-level features that are processed through convolutions.

최근 컨볼루션 네트워크의 발전으로 인해 키포인트를 직접 회귀하거나 키포인트 히트맵을 추정하는 2가지 접근 방법이 주로 연구되고 있습니다. 그리고 대부분의 모델의 경우 이미지를 저해상도로 낮추어 고수준의 특징을 추출하고 다시 복원함으로써 해당 위치를 추정하는 방법을 따르고 있습니다.

- (1) 좌표 예측 : 단순히 좌표값을 예측하기에 convolution 을 사용하는 경우 공간 정보를 무시하는 경향. 하지만 출력시 후처리 필요없이 바로 예측 가능
- (2) Heat Map : 공간 정보를 최대한 유지하며 공간적 지역 특징을 반영할 수 있습니다. 또한 출력이 확률 분포이기에 불확실성을 자연스럽게 모델링 가능합니다. 하지만 후처리를 통해서 좌표를 예측 해야한다 + 히트맵의 해상도에 따라 정밀도가 제한될수 있다.

+) Hourglass 논문에서는 위의 2가지를 자연스럽게 섞어서 사용하고 있다.

Approach

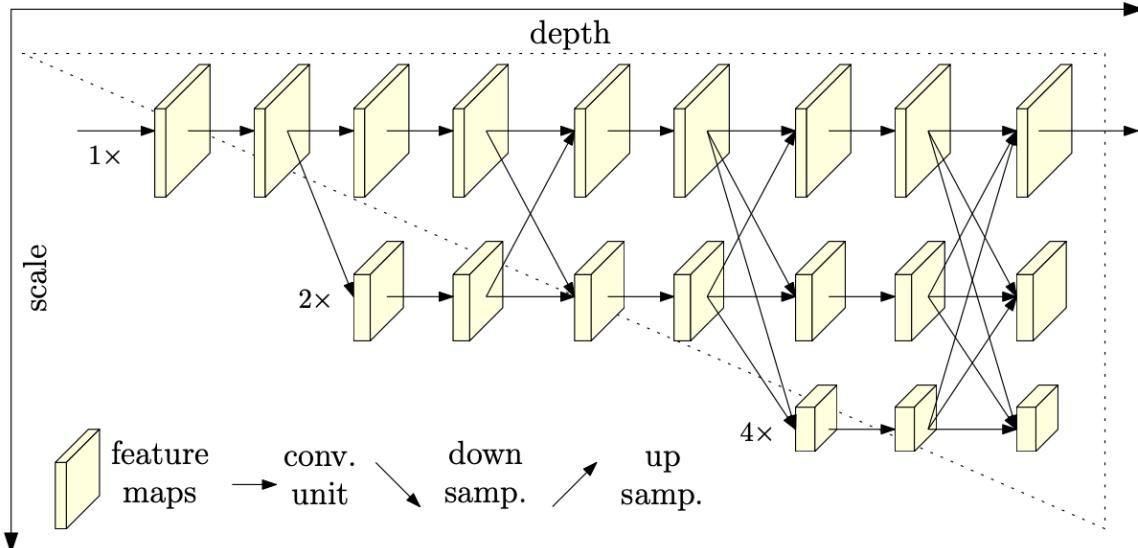


Figure 1. Illustrating the architecture of the proposed HRNet. It consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion). The horizontal and vertical directions correspond to the depth of the network and the scale of the feature maps, respectively.

Human pose estimation은 이미지 내에서 K개의 관절 혹은 특정 부위의 위치를 찾는 문제입니다. 최근 연구들에서는 각 좌표를 직접 예측하는 대신 $H * W$ 크기의 2차원 히트맵을 K 개 예측하는 방식으로 접근되어 왔습니다. 즉, 각각의 히트맵은 각 키포인트가 존재할 가능성을 확률 분포로 나타내게 됩니다.

그래서 본 논문에서도 HRNet을 기존의 방식과 아키텍처 부분에서는 거의 동일하게 구성된다고 합니다.

(1) Stem : 첫 이미지를 2번의 convolution을 거쳐 해상도를 줄여서 파라미터 수를 줄이고, 특징맵을 추출합니다.

(2) Main Body : 입력 특징맵과 동일한 크기의 출력 특징맵을 추출하여 공간 정보의 손실을 최소화 합니다. 동시에 여러 스케일의 특징맵을 융합하여 특징을 보강합니다.

(3) Regressor : 히트맵으로부터 최종 좌표를 결정합니다. 필요시 다시 원본 해상도로 복구합니다.

HRNet은 이러한 구조를 가지고 있고, 여기서 특히 Main Body 모델 설계에 초점을 두고 있다고 주장합니다.

Sequential multi-resolution subnetworks

기존의 Pose estimation 네트워크들은 해상도를 점진적으로 낮추는 서브네트워크들을 직렬로 연결하는 구조를 가지고 있습니다. 즉, 각 스테이지마다 하나의 네트워크가 구성되며, 각 스테이지마다 이전 특징맵의 해상도의 $1/2$ 가 됩니다. 이를 수식으로 나타내면 아래와 같이 나타낼 수 있습니다. (단 , 해당 서브네트워크의 해상도는 처음 서브네트워크의 해상도의 $\frac{1}{2^r}$ 입니다)

$$N_{1,1} \rightarrow N_{2,2} \rightarrow N_{3,3} \rightarrow N_{4,4}$$

이러한 방법을 사용하게 되면 어쩔 수 없이 고해상도의 공간적 정보가 손실되게 됩니다.

Parallel multi-resolution subnetworks

기존의 순차적 방법과 다르게 병렬적 방법은 해상도를 유지하는 stream이 있음과 동시에 다음 스테이지 해상도가 낮은 또 다른 서브 스트림을 생성합니다. 그리고 이들을 병렬로 연결하는 구조를 가지고 있습니다. 병렬적 표현을 식으로 표현하면 아래와 같이 표현할 수 있습니다 ((단 , 해당 서브네트워크의 해상도는 처음 서브네트워크의 해상도의 $\frac{1}{2^r}$ 입니다)

$$\begin{aligned} N_{11} &\rightarrow N_{21} \rightarrow N_{31} \rightarrow N_{41} \\ &\quad \swarrow N_{22} \rightarrow N_{32} \rightarrow N_{42} \\ &\quad \swarrow N_{33} \rightarrow N_{43} \\ &\quad \swarrow N_{44}. \end{aligned}$$

여기서 N_{ij} 로 표현되는데, j 가 동일한 경우 동일한 해상도를 갖는 subnetwork임을 의미합니다. 그리고 i 는 각각의 스테이지를 의미합니다. 이러한 구조를 통해서 고해상도의 공간적 정보를 유지하면서도 동시에 저해상도의 특징을 융합하여 정밀도를 높히는데 기여할 수 있습니다.

Repeated multi-scale fusion

본 논문에서는 다양한 스케일의 특징맵을 융합하는 **exchange units**을 소개합니다. 이를 아래와 같이 표현하고 있습니다.

$$\begin{array}{ccccccc}
 C_{31}^1 & \searrow & C_{31}^2 & \searrow & C_{31}^3 & \searrow & \\
 C_{32}^1 & \rightarrow & \mathcal{E}_3^1 & \rightarrow & C_{32}^2 & \rightarrow & \mathcal{E}_3^2 & \rightarrow & C_{32}^3 & \rightarrow & \mathcal{E}_3^3, \\
 C_{33}^1 & \nearrow & C_{33}^2 & \nearrow & C_{33}^3 & \nearrow & \\
 \end{array} \tag{3}$$

그리고 각 C의 구성을 보면 $c_{s,r}^b$ 여기서 b는 s 스테이지의 블럭을 의미하며, r은 해상도를 의미합니다.

(추가로 스테이지와 블럭의 개념이 헷갈려서 figure1의 그림을 빌려 설명하면 가장 큰 단위가 스테이지이며, 스테이지는 몇개의 스케일을 다루는지로 구별이 가능하며 해당 그림에서는 다음과같이 3개의 스테이지로 구별이 가능하다. 그리고 각 스테이지에서 몇번째 블럭인지를 b로 표시할 수 있는 것이다.)

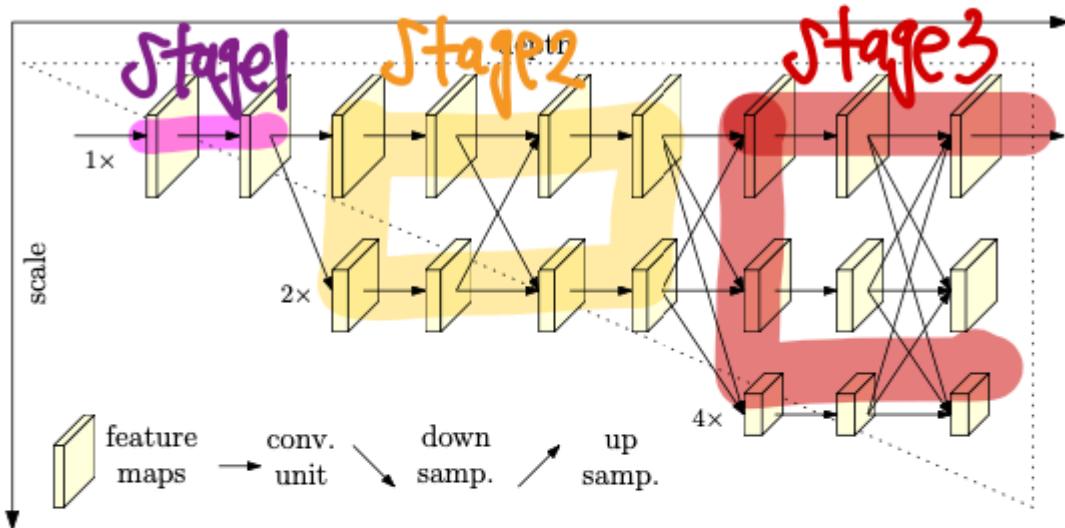


Figure 1. Illustrating the architecture of the proposed HRNet. It consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion). The horizontal and vertical directions correspond to the depth of the network and the scale of the feature maps, respectively.

그래서 (3) 에서 표현하는 것은 결국 figure1의 stage3를 의미하는 것이다. 그리고 결국 서로다른 3개의 스케일을 입력으로 받아 융합한 후, 다음 블럭의 입력으로 전달해주는 역할을 하는 것이 Exchange Unit 이다.

해당 유닛의 경우 서로 다른 스케일의 스케일의 입력을 업샘플링 혹은 다운 샘플링을 진행한 후 합치게 됩니다. 즉 출력의 경우 입력과 동일한 크기를 갖게 됩니다.

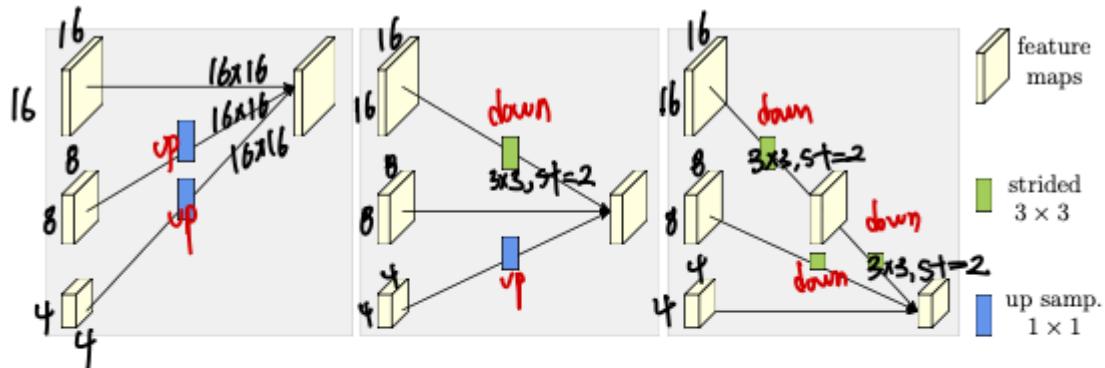


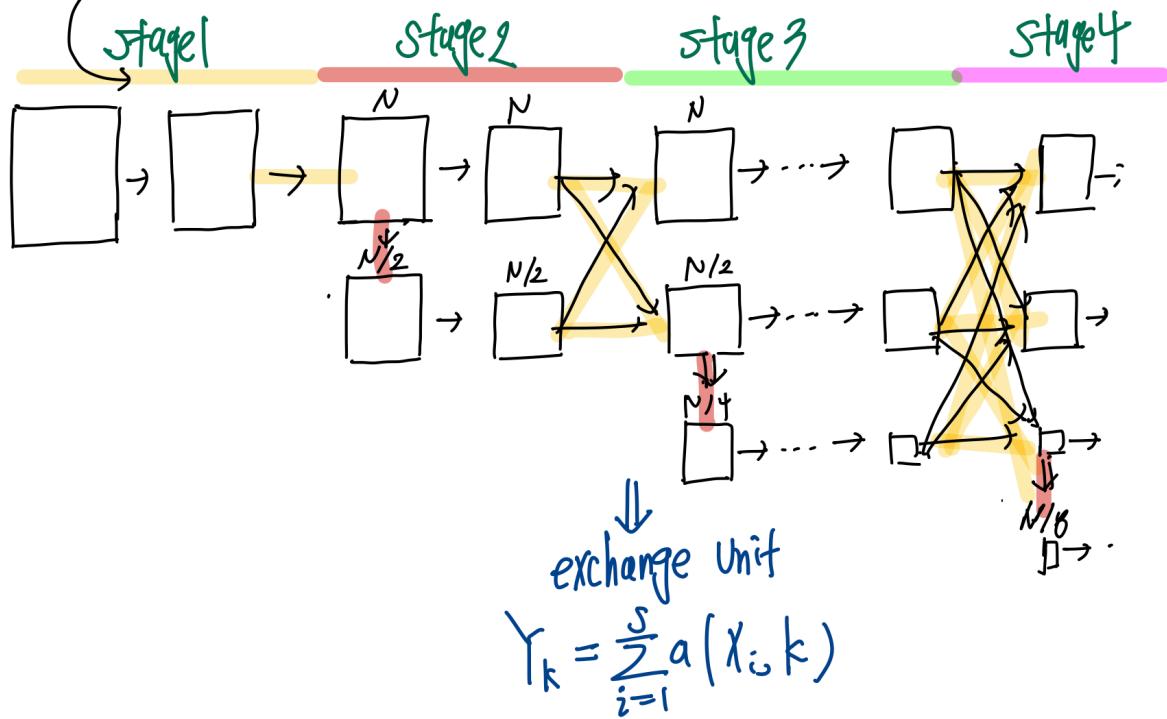
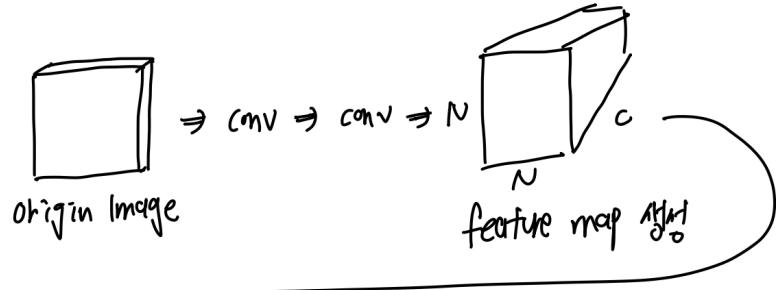
Figure 3. Illustrating how the exchange unit aggregates the information for high, medium and low resolutions from the left to the right, respectively. Right legend: strided 3×3 = strided 3×3 convolution, up samp. 1×1 = nearest neighbor up-sampling following a 1×1 convolution.

다음과 같이 $a(X_i, k)$ 함수를 통해서 up, down sampling이 진행됩니다. a 함수의 경우 X_i 입력을 k 해상도로 변경해주는 역할을 수행합니다. 그래서 최종적으로 식으로 표현하면 아래와 같은 식으로 표현이 가능합니다.

$$Y_k = \sum_{i=1}^s a(X_i, k).$$

그리고 다음 스테이지로 넘어가는 경우 아래와 같은 수식을 통해서 저해상도 브랜치를 생성하게 됩니다.

$$Y_{s+1} = a(Y_s, s+1).$$



Heatmap estimation

그리고 본 논문에서는 마지막 exchange unit을 통해서 나온 고해상도 이미지를 기반으로 단순하게 Regress 하는 것이 가장 좋다고 합니다. 이렇게 해서 나온 히트맵과 정답과의 MSE 손실함수를 적용하게 된다고 합니다.

Network instantiation

본 논문의 아키텍처 또한 ResNet 기반으로 설계되었습니다. Residual block은 $1 * 1 \rightarrow 3 * 3 \rightarrow 1 * 1$ 로 구성 되며 이때 $3 * 3$ 시 너비를 64로 낮춘다고 합니다. 그리고 첫 스테이지에서는 총 4개의 Residual block을 사용하여 이미지의 특징을 추출한다고 합니다.

그리고 다음 2,3,4 스테이지의 경우 Exchange block으로 구성되는데, 각 Exchange block의 경우 총 4개의 Residual block으로 구성된 후 exchange unit을 실행합니다. 그

리고 2,3,4 스테이지는 각각 1,4,3 개의 Exchange block으로 구성됩니다. 이를 통해서 총 8번 exchange unit을 통해서 서로 다른 스케일의 정보가 섞이게 됩니다.

IV Residual Block R

HRNet의 경우
3x3x1 C=6으로 설정.

$$\text{Input} \rightarrow |X| \Rightarrow 3 \times 3 \Rightarrow |X| \Rightarrow \text{최종 해상도 변화} X.$$

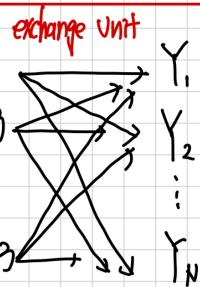
V Exchange Unit E

$$\text{Input 1} \rightarrow R \Rightarrow R \Rightarrow R \Rightarrow R \Rightarrow 3 \times 3 \rightarrow 3 \times 3$$

$$\text{Input 2} \rightarrow R \Rightarrow R \Rightarrow R \Rightarrow R \Rightarrow 3 \times 3 \rightarrow 3 \times 3$$

:

$$\text{Input N} \rightarrow R \Rightarrow R \Rightarrow R \Rightarrow R \Rightarrow 3 \times 3 \rightarrow 3 \times 3$$



VI Model Architecture (Y_{ij} i: stage j: index)

<Stage 1>

$$\text{Input Image} \rightarrow R \Rightarrow R \Rightarrow R \Rightarrow R \Rightarrow 3 \times 3 \Rightarrow Y_{11}$$

해상도 1/2 & 너비 X2

$$Y_{12}$$

<Stage 2>

$$Y_{11} \rightarrow E_1 \rightarrow Y_{21}$$

$$Y_{12} \rightarrow E_2 \rightarrow Y_{22}$$

$$Y_{23}$$

<Stage 3>

$$Y_{21} \rightarrow E_3 \rightarrow E_3 \rightarrow E_3 \rightarrow E_3 \rightarrow Y_{31}$$

$$Y_{22} \rightarrow E_3 \rightarrow E_3 \rightarrow E_3 \rightarrow E_3 \rightarrow Y_{32}$$

$$Y_{23} \rightarrow E_3 \rightarrow E_3 \rightarrow E_3 \rightarrow E_3 \rightarrow Y_{33}$$

$$Y_{34}$$

<Stage 4>

$$Y_{31} \rightarrow E \rightarrow E \rightarrow E \rightarrow Y_{41} \rightarrow |X| \Rightarrow \text{Final heatmap.}$$

$$Y_{32} \rightarrow E \rightarrow E \rightarrow E \rightarrow Y_{42}$$

$$Y_{33} \rightarrow E \rightarrow E \rightarrow E \rightarrow Y_{43}$$

$$Y_{34} \rightarrow E \rightarrow E \rightarrow E \rightarrow Y_{44}$$

그리고 최종 히트맵의 경우 고해상도 특징맵에 1×1 conv를 적용하여 k채널의 히트맵을 생성한다고 합니다.

Experiments and Result



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.

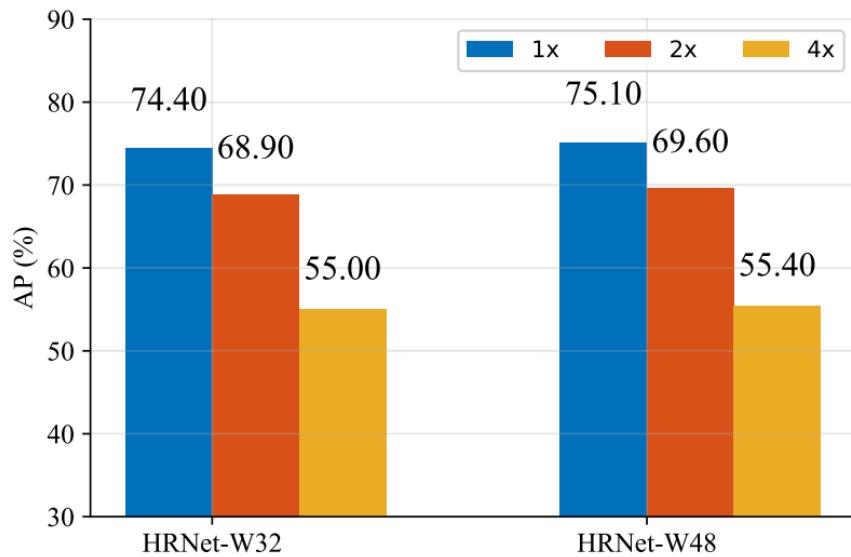


Figure 5. Ablation study of high and low representations. $1\times$, $2\times$, $4\times$ correspond to the representations of the high, medium, low resolutions, respectively.

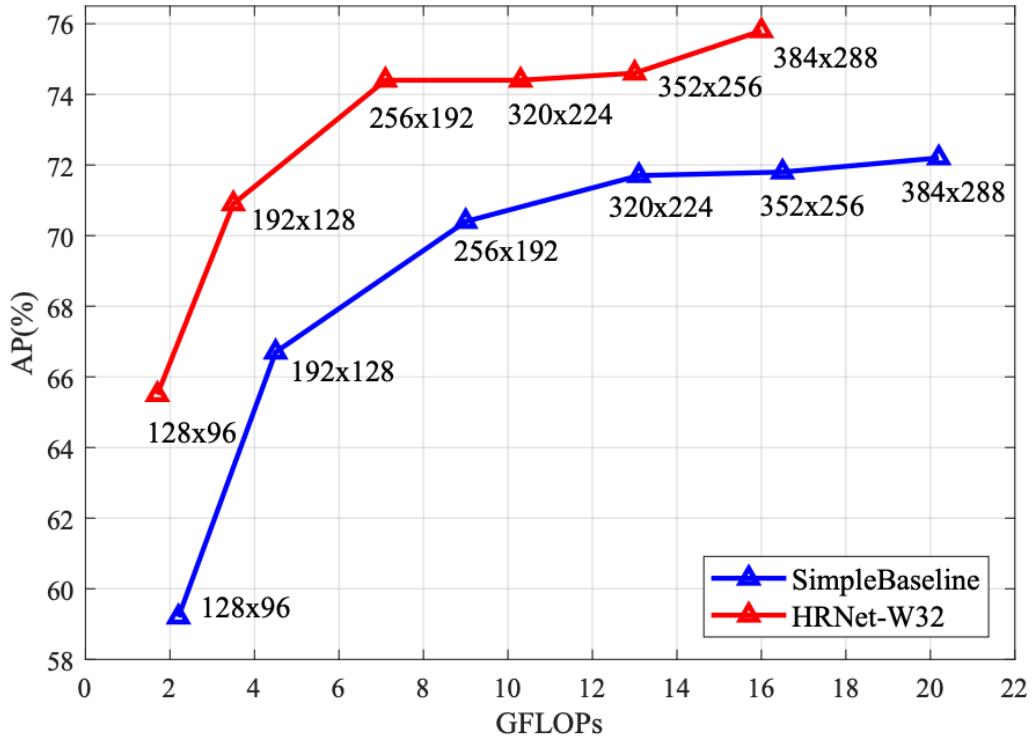


Figure 6. Illustrating how the performances of our HRNet and SimpleBaseline [72] are affected by the input size.

다양한 실험을 통해서 다른 모델보다 높은 성능을 달성하였고, 특히 융합 회수가 많을수록 더 높은 AP를 달성했다고 실험적으로 증명하였습니다. 추가로 고해상도의 히트맵이 저해상도 히트맵 보다 높은 성능을 달성함을 통해 고해상도의 공간적 정보 유지의 중요성을 주장합니다.

정리

Pose Estimation 분야는 최근 다양한 스케일과 히트맵을 통한 예측이 높은 성능을 달성하고 있었다. 그래서 본 논문은 히트맵의 성능을 위해 고해상도 정보를 유지한체, 다양한 스케일의 이미지를 병렬적으로 처리할 수 있는 HRNet 모델을 제안하며 높은 성능을 달성하였습니다.

나의 생각

지금까지의 pose estimation 연구를 보면 결국 히트맵을 기반으로 최대한 다양한 스케일의 이미지를 융합하고자 하는 연구들이 진행중인 것 같다. 하지만 모델의 구조를 그리면서도 느꼈지만 모델의 굉장히 큰것 같다. 그래서 학습이 잘되더라도 실시간으로 추론이 가능하는데 한계가 있다고 생각되었다. 그리고 exchange unit에서 단순히 가중합을 하는것이 아니라 attention 을 적용하면 더욱 성능 향상에 도움이 되지 않을까? 생각이 들었다. 추가로 마지막 고해상도 이미지만 히트맵으로 생성하기 보다 각 스케일별 히트맵을 생성하고 히트맵을 다시 융합하여 하나의 최종 히트맵을 생성하면 Hourglass처럼 Heatmap을 재사용하여 2 차원에 더욱 잘 매핑되지 않을까? 생각하였다.