A-HRNet: Attention Based High Resolution Network for Human pose estimation

Ying Li
Department of Computer Science
University of Massachusetts Lowell
Lowell, USA
ying_li@student.uml.edu

Chenxi Wang
Department of Electrical and Computer Engineering
University of Massachusetts Lowell
Lowell, USA
chenxi_wang1@student.uml.edu

Yu Cao
Department of Computer Science
University of Massachusetts Lowell
Lowell, USA
yu_cao@uml.edu

Benyuan Liu
Department of Computer Science
University of Massachusetts Lowell
Lowell, USA
benyuan_liu@uml.edu

Yan Luo
Department of Electrical and Computer Engineering
University of Massachusetts Lowell
Lowell, USA
yan_luo@uml.edu

Honggang Zhang
Engineering Department
University of Massachusetts Boston
Boston, USA
Honggang.Zhang@umb.edu

Abstract—Recently, human pose estimation has received much attention in the research community due to its broad range of application scenarios. Most architectures for human pose estimation use multiple resolution networks, such as Hourglass, CPN, HRNet, etc. High Resolution Network (HRNet) is the latest SOTA architecture improved from Hourglass. In this paper, we propose a novel attention block that leverages a special Channel-Attention branch. We use this attention block as the building block and adopt the architecture of HRNet to build our Attention Based HRNet (A-HRNet). Experiments show that our model can consistently outperform HRNet on different datasets. Moreover, our model achieves the state-of-the-art performance on the COCO keypoint detection val2017 dataset (77.7 AP).

Index Terms—human pose estimation, attention block

I. INTRODUCTION

Human pose estimation (HPE) focuses on detecting the positions of human body key joint points, such as ears, shoulders, elbows, wrists, waist, knees and ankles, etc, from a single image. The techniques developed for this task provide an important foundation for a wide range of practical applications, such as human-computer interaction [1], human action recognition [2] [3] and human pose tracking [4]. Sub-areas of this research topic include single-person and multi-person pose estimation. In this paper, we focus on the single-person task. An example is shown in Fig. 1.

In recent years, convolutional neural networks (CNNs) have attracted intensive attention from the computer vision community across a variety of tasks [5], [6]. This tendency is also witnessed in human pose estimation [4], [7]–[9]. Nearly all the current state-of-art solutions on public datasets are CNN based. Consequently, deeper and more complicated CNN models are being developed to further improve the performance, such as [10] and [11].

To help CNN models better focus on relevant features, attention models are leveraged in many computer vision tasks [11] [12] [13]. In particular, DNANet [11] adopts HRNet as the backbone, outperforms HRNet on the COCO benchmark by adding attention modules. All these previous work demonstrates that extra attention-aware features help to train better models.

Our work is inspired by the combination of HRNet and the attention mechanism. However, two differences exist between our work and [11]. First, [11] adopts three different types of attention modules in the model while we only use one unified attention module, which makes the model simpler and easier to train. Second, our proposed attention module is less complex with fewer parameters and lower computation costs than that in [11]. More details will be described in Section 3 and 4.

In summary, our main contributions in this paper include:

• We design a novel attention block that introduces a channel attention branch in the basic residual block, so as to generate the extra attention-aware features with inconspicuous computation cost.
• We propose an Attention Based High Resolution Network...
Fig. 2. Architecture of A-HRNet. It keeps the multiple resolution branches, down samplings and up samplings from the original HRNet. But the deep convolution blocks are replaced by our proposed attention blocks shown in Fig. 3. Better viewed in color.

(A-HRNet) that replaces the residual basic blocks in HRNet with our novel attention blocks, to provide the easy accessibility for implementation.

- Our model is trained from scratch only on COCO train2017 dataset but achieves state-of-the-art performance on COCO val2017 dataset.

II. RELATED WORK

Attention Mechanism. The attention mechanism has been widely adopted in natural language processing (NLP) tasks, to achieve the state-of-the-art performance in the machine translation task [14] and language understanding task [15]. Recently, attention-aware features are also found to work very well in computer vision tasks. For example, [12] proposed a powerful attention module that combines an attention branch with an hourglass block, which was stacked multiple times to form a deep convolutional neural network for image classification. Based on the self-attention mechanism, the network proposed in [13] captured rich contextual dependencies for the scene segmentation task. [16] [11] incorporated the attention mechanism with different convolutional neural networks for human pose estimation.

Human Pose Estimation. Human pose estimation has been an active research topic for decades. Before the advent of CNN, traditional methods rely on a variety of features designed by researchers [17] [18]. After the AlexNet [19] won the ImageNet challenge in 2012, significant improvements have been achieved in human pose estimation tasks benefiting from the representation capability of CNNs. For example, [20] first adopted CNN for human pose estimation and formulated this task as a CNN-based regression problem towards body joints. [9] proposed a Stacked Hourglass Network, which is the first multi-resolution network for human pose estimation. [21] proposed a Cascade Pyramid Network (CPN) to integrate all levels of feature representations to refine the process of pose estimation, which won the COCO 2017 keypoint challenge. [4] provided simple and effective baseline methods by adding a few deconvolutional layers on a backbone network.

[10] proposed a novel High Resolution Network (HRNet), which maintained high-resolution representations throughout the whole architecture. [11] improved HRNet by increasing different types of attention modules.

III. A-HRNet

We adopt High Resolution Network (HRNet) as the backbone in our A-HRNet, and replace the basic deep convolutional blocks by our attention blocks. In this section, we first review HRNet and then explain the details of the proposed method.

A. Review High Resolution Network

High Resolution Network (HRNet) is a multi-resolution network improved from Hourglass [9]. As shown in Fig. 2, HRNet starts from a high resolution subnetwork and keeps this high resolution branch throughout the whole architecture. Starting from the second stage, HRNet adds a lower resolution branch at the beginning of the stage and merges features from all lower resolution branches to the high resolution branch at the end of each stage. The subsequent stages follow the same manner and keep adding branches of lower resolution. Let $N_{sr}$ denote the subnetwork in the $s$-th stage of $r$ resolution index. The resolution in this branch is $1/2^{r-1}$ of that in the high resolution branch. An example of HRNet, containing 4 different resolution branches, is illustrated as follows,

$$
N_{11} \rightarrow N_{21} \rightarrow N_{31} \rightarrow N_{41} \\
\downarrow \quad \downarrow \quad \downarrow \\
N_{22} \rightarrow N_{32} \rightarrow N_{42} \\
\downarrow \quad \downarrow \\
N_{33} \rightarrow N_{43} \\
\downarrow \\
N_{44}.
$$

(1)
Benefiting from the high resolution features and the fusion of features from all lower resolution branches, HRNet achieves excellent performance on human pose estimation tasks.

### B. Our Method

#### Attention Block

We design a novel attention block to extract extra attention-aware features on the basis of the residual convolutional block in HRNet. As shown in Fig.3, our attention block contains a channel attention branch, and the rest is the same as the basic residual block in HRNet. Suppose that the size of the input feature map is $D_w \times D_h \times c$, in which $c$ is the number of channels and $D_w \times D_h$ is the feature map size. After the attention block, the resolution remains the same. The computing cost for the channel attention branch is

$$2 \times D_w \times D_h \times c \times \frac{c}{s} \times \frac{c}{s} \times D_w \times D_h,$$

(2)

where $s$ is the scale parameter for the channel attention branch. We adopt $s = 4$ in our network. The computing cost for the main convolutional branch is

$$2 \times 3 \times 3 \times D_w \times D_h \times c \times c.$$

(3)

Comparing equations (2) and (3), we have

$$\frac{2 \times D_w \times D_h \times c \times \frac{c}{s} \times \frac{c}{s} \times D_w \times D_h}{2 \times 3 \times 3 \times D_w \times D_h \times c \times c} = \frac{1}{32},$$

(4)

which shows that the computing cost for the channel attention branch is much smaller than that of the main branch.

#### A-HRNet

We leverage our attention blocks to improve the feature representation capability of HRNet. As shown in Fig. 2, we replace all the basic residual convolutional blocks in HRNet with our attention blocks. The only difference lies in the channel attention branch. As aforementioned, this branch will only introduce little extra computational cost while we will show in Section 4 that our A-HRNet can consistently outperform HRNet at a considerable scale.

### IV. EXPERIMENTS

#### A. COCO Keypoint Detection

**Dataset.** COCO keypoint detection dataset [22] contains over 200K images and 250K person instances labeled with 17 keypoints. We train our model on COCO train2017 dataset that consists of 57K images and 150K person instances. We evaluate our model on the val2017 set with 5000 images.

**Evaluation metric.** The keypoint evaluation metrics used by COCO is Object Keypoint Similarity (OKS): $\text{OKS} = \Sigma_i [\exp(-d_i^2/2\sigma^2)] / \Sigma_i [\delta(v_i > 0)]$, in OKS, $d_i$ is the Euclidean distance between each corresponding ground truth and the detected keypoint, $v_i$ is the visibility flag of the ground truth, $s$ is the object scale, and $k_i$ is a per-keypoint constant that controls falloff. We report standard average precision and recall scores [23]: $\text{AP}$ (AP at OKS=.50:.05:.95, primary challenge metric), $\text{AP}^{50}$ (AP at OKS=.50, loose metric), $\text{AP}^{75}$ (AP at OKS=.75, strict metric); $\text{AP}^L$ (AP for medium objects), $\text{AP}^S$ (AP for large objects); and AR at OKS=.50:.05:.95.

**Training.** We use the same human detector, input size, data augmentation, and Adam optimizer as in HRNet [10]. Since we train our models from scratch, a different learning schedule is adopted, which sets the initial learning rate at $1e-2$, and reduces it to $1e-3$, $1e-4$, and $1e-5$ at the 20th, 170th, and 200th epoch respectively. The training process is terminated within 210 epochs.

**Results on the validation set.** We compare the results of our method and other state-of-the-art methods in Table I. Our A-HRNet achieves an AP score of 77.7, outperforming all other methods. Note that we train the model on COCO train2017 set from scratch without extra data. Compared to DNNAnet [11], our model improves the AP score by 0.8 points with fewer number of parameters and GFLOPs.

### B. Rehabilitation Activities Keypoint Detection

**Dataset.** The lower body rehabilitation activities keypoint detection dataset [3] is organized following the COCO dataset but with 30 joint points annotations. The dataset is split into a training set and a validation set of 500,000 images and 10,000 images, respectively. And the images in the training set and the validation set are from different volunteers. The ground truth results are stored in JSON files in the same structure as in the COCO dataset. There is only one person in each image.

**Training.** Firstly, we modify the COCO API codes to process the images in our dataset and calculate the standard average precision and recall scores. Then we follow the same data
As shown in Table III, all the models using attention branches outperform the original HRNet on COCO keypoint detection dataset. Future work includes improving the model for human pose estimation and designing a new model for multi-person pose estimation.

VI. ACKNOWLEDGEMENT

The authors would like to thank the anonymous referees for their reviews and insightful suggestions to improve this paper. This work is partially supported by the National Science Foundation of USA (Award No. 1547428 and No. 1738965).

REFERENCES


