2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) | 979-8-3503-2056-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/WACVW58289.2023.00059

# BAPose: Bottom-Up Pose Estimation with Disentangled Waterfall Representations

Bruno Artacho
Rochester Institute of Technology
bmartacho@mail.rit.edu

Andreas Savakis
Rochester Institute of Technology
andreas.savakis@rit.edu

## Abstract

*We propose BAPose, a novel bottom-up approach that achieves state-of-the-art results for multi-person pose estimation. Our end-to-end trainable framework leverages a disentangled multi-scale waterfall architecture and incorporates adaptive convolutions to infer keypoints more precisely in crowded scenes with occlusions. The multi-scale representations, obtained by the disentangled waterfall module in BAPose, leverage the efficiency of progressive filtering in the cascade architecture, while maintaining multi-scale fields-of-view comparable to spatial pyramid configurations. Our results on the challenging COCO and CrowdPose datasets demonstrate that BAPose is an efficient and robust framework for multi-person pose estimation, significantly improving state-of-the-art accuracy. Human Pose Estimation, Multi-Scale Representations*

## 1. Introduction

Estimating human pose in crowded scenes is a challenging task of high interest in computer vision research and applications such as action recognition, sports analysis, human-computer interactions, and sign language recognition. Various methods have focused on specific aspects of human pose estimation, including 2D pose estimation [40], [28], [43], [3], [39]; 3D pose estimation [37], [47], [1], [5]; single frame detection [6]; pose detection in videos [12]; dealing with a single person [43] or multiple people [7].

Multi-person pose estimation is very challenging due to joint occlusions and the large number of degrees of freedom in the human body. Common approaches include the deployment of statistical and geometric models to estimate occluded joints [33], [29] and the use of anchor poses [37], [42], although the latter is limited by the number of poses in its library and has difficult generalizing to unforeseen poses.

State-of-the-art (SOTA) methods for multi-person pose estimation can be divided in two distinct approaches: top-down and bottom-up. The former detects instances of persons in the image and then perform single person pose estimation for each individual. The latter either detect all keypoints and group them by affinity relations [7], [21], or directly regress the keypoints to each person in the image [14]. Overall, top-down approaches achieve high accuracy, albeit they require an extra step for detection, resulting in a slower and more costly process. On the other hand, bottom-up approaches are based on a single-stage for multi-person pose estimation that is generally more efficient.

Figure 1. Pose estimation examples with our BAPose method.

We propose BAPose, a bottom up framework that is named after "Basso verso l'Alto" (bottom-up in Italian). The BAPose method is a single-stage, end-to-end trainable network that improves recent successful approaches by UniPose [3] , UniPose+ [5], and OmniPose [4] to bottom-up multi-person 2D pose estimation. BAPose achieves SOTA results in two large datasets without requiring post-processing, intermediate supervision, multiple iterations or anchor poses. The main contributions of BAPose are:

- We propose BAPose, a novel single-pass, end-to-end trainable, multi-scale approach for bottom-up multi-person 2D pose estimation, that achieves SOTA results on the COCO and CrowdPose benchmarks.
- Our bottom-up approach combines multi-scale waterfall features with disentangled adaptive convolutions

2690-621X/23/$31.00 ©2023 IEEE
DOI 10.1109/WACVW58289.2023.00059

and an integrated multi-scale decoder to disambiguate the joints of individuals in crowded scenes without requiring a separate detector.
- The enhanced multi-scale capability of BAPose is suitable for human pose estimation in images with a large number of person instances, drastically increasing the SOTA performance for the CrowdPose dataset.

## 2. Related Work

The use of Convolutional Neural Networks (CNNs) for deep learning methods enabled leaping advances for the task of human pose estimation [40], [7], [37], [5]. The Convolutional Pose Machines (CPM) [43] approach uses a sequence of CNN stages in the network to refine joint detection. Expanding [43] integrated Part Affinity Fields (PAF) in their OpenPose [7] framework to better capture relationships between joints for improved human pose estimation.

The Stacked Hourglass (HG) network [28] utilizes a multi-stage approach by cascading hourglass structures through the network to refine the resulting pose estimation. This work was further expanded to incorporate the multi-context approach in [13] by augmenting the backbone with residual units in order to increase the receptive Field-of-View (FOV). A downside of this approach is the increase in complexity by the addition of another stage of postprocessing with Conditional Random Fields (CRFs) and the associated increase in computational load.

Aiming to offer a multi-scale approach to feature representations, the High-Resolution Network (HRNet) includes both the high and low resolutions to obtain a larger FOV. The Multi-Stage Pose Network (MSPN) [23] follows a similar approach to HRNet by combining the cross-stage feature aggregation and coarse-to-fine supervision, while Distribution-Aware coordinate Representation of Keypoints (DARK) method [44] refines their decoder in order to reduce the inference error at the decoder stage. In further work, [11] combined the HRNet structure with multi-resolution pyramids to obtain multi-scale features.

Developments in graphical components for CNNs inspired [45] by applying graphs to further extract the contextual information for pose, while Cascade Feature Aggregation (CFA) [38] applied the cascade approach into the semantic information for pose estimation. Generative Adversarial Networks (GANs) were used in [8] to learn dependencies and contextual information for pose. Transformer-based networks were also investigated by TokenPose [24] to better asses the global dependencies of the pose estimation.

A limitation of top-down approaches is the requirement of an independent module for the detection of instances of humans in the frame. LightTrack [32], for instance, applies YOLOv3 [36] to detect subjects prior to the detection of joints for pose estimation, while LCR-Net [37] applies multiple branches for detection by using Detectron [15] and the arrangement of joints during classification.

With the goal of developing a unified framework to overcome the limitation of top-down approaches, UniPose [3] combines the bounding box generation and pose estimation in a single, one-pass network. This approach is possible due to the larger FOV and significant increase in the multi-scale representation obtained by the Waterfall Atrous Spatial Pooling (WASP) module [2], allowing greater FOV and results in better representation of contextual information.

### 2.1. Bottom-Up Approaches

Bootom-up methods face the additional challenge of detecting the joints of multiple people without an external person detector that is commonly used in top-down methods. The most common approach for bottom-up estimation is to associate detected keypoints with each person in the image. The problem was cast in terms of integer linear programming in [35], [17], but a clear drawback is the high processing time inhibiting real-time performance.

OpenPose [7] is considered a breakthrough approach for grouping keypoints by introducing PAF. Other methods further developed PAF, such as Pif-Paf [21] and associative embedding [27]. Similarly, PersonLab [34] adopted Hough voting, and [19] hierarchical graphical clustering.

The dense regression of pose candidates is adopted by several recent works [31], [30]. A limitation of this approach is the lower regression accuracy in the localization process, that usually requires an additional post-processing step in order to improve the regression results. Aiming to bridge the gap, [41] applied a mixture density network to better handle uncertainty in the network before regression. The recent Disentangled Keypoint Regression (DEKR) method [14], on the other hand, learns disentangled representations for each keypoint and utilizes adaptively activated pixels, ensuring that each representation focuses on the corresponding keypoint area.

### 2.2. Multi-Scale Feature Representations

The reduction of resolution that takes place in CNN methods is a challenge for pose estimation or semantic segmentation methods. Fully Convolutional Networks (FCN) [26] addressed resolution reduction by adopting upsampling strategies to increase the size of the features maps, reverting it to the original input dimensions. Further, DeepLab [9] deployed atrous convolutions to achieve a multi-scale framework and increase the size of the receptive fields, avoiding downsampling with the introduction of the Atrous Spacial Pyramid Pooling (ASPP). The DeepLab applies atrous convolutions in four parallel branches with different rates, and combines them at the original image resolution.

Improving upon ASPP [9], the WASP module incorporates multi-scale features without immediately parallelizing the input stream [2], [3]. The WASP module creates a wa-

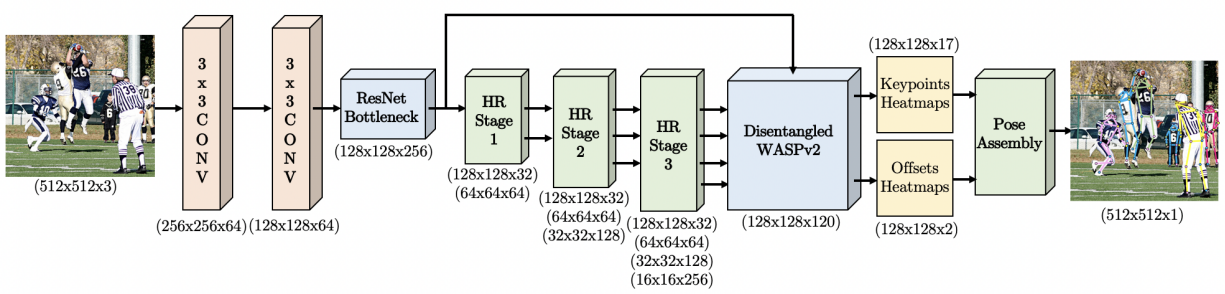


Figure 2. BAPose architecture for bottom-up multi-person pose estimation. The input color image is fed through the HRNet backbone for initial feature extraction, followed by the D-WASP module and an adaptive convolution based decoder to generate one heatmap per joint (17 joints in the figure) and offset regression for the localization of each person instance.

terfall flow by initially processing through a filter and later creating a new branch, and extends the cascade approach by combining the streams from all its branches reaching a multi-scale representation. The OmniPose framework [4] introduced the enhanced WASPv2 module, improving upon the multi-scale feature extraction from the backbone and includes the decoder features of the network.

## 3. BAPose Architecture

The proposed BAPose bottom-up method, illustrated in Figure 2, consists of a single-pass, single output branch network that is effective for multi-person 2D pose estimation in crowded scenes. BAPose integrates improvements in multi-scale feature representations [4], [14], an encoder-decoder structure combined with the spatial pyramid pooling of the waterfall configuration, and disentangled adaptive regression for person localization and parts association.

The processing pipeline of the BAPose architecture is shown in Figure 2. The input image is initially processed by the HRNet feature extractor. The extracted multi-scale feature maps are then processed by the D-WASP module with integrated decoder, that extracts the location of keypoints and contextual information for the localization regression. The network generates $K$ heatmaps, one for each joint, with the corresponding confidence maps as well as 2 offset maps for the identification of person instances and association of keypoints to each instance. The integrated D-WASP decoder generates detections from all scales of the feature extraction for both visible and occluded joints while maintaining the image resolution through the network.

Our architecture includes several innovations that contribute to increased accuracy. In the D-WASP module, BAPose combines atrous convolutions and the waterfall architecture to increase the network's capacity to represent multi-scale contextual information by the probing of feature maps at multiple rates of dilation. This configuration achieves a larger FOV in the encoder. Our architecture also integrates disentangled adaptive convolutions in the decoding process, enabling the single-pass detection of multiple person instances and their keypoint estimation. Additionally, our network demonstrates superior ability to deal with a large number of subjects by the enhanced extraction of features at multiple scales, as indicated by SOTA results for the CrowdPose dataset presented in Section 6. Finally, the modular nature of BAPose facilitates the easy implementation and training of the network.

Our work on BAPose introduces the D-WASP module for the more complex task of bottom-up, multi-person pose estimation. Top down methods utilize detectors for identifying individual poses in multi-person settings, which requires an additional stage of processing. BAPose utilizes a bottom-up approach, without relying on an additional person detector to locate different instances of persons in the image, which is different and more efficient than top-down approaches, e.g.OmniPose [4]. The D-WASP module proposed in BAPose combines, for the first time, the multi-scale approach of the waterfall atrous convolutions with disentangled adaptive convolutions to better estimate the joints and effectively detect multiple person instances.

### 3.1. Disentangled Waterfall Module

The proposed enhanced "Disentangled Waterfall Atrous Spatial Pyramid" module, or D-WASP, is shown in Figure 3. The D-WASP module processes all four levels of feature maps from the backbone through the waterfall branches with different dilation rates. Low-level and high-level features are represented at the same resolution, achieving a refined localization for joint estimation. Furthermore, the D-WASP module uses adaptive convolution blocks to infer the final heatmaps for joint localization and offset maps for person instance regression. The module generates both the keypoints and offset heatmaps for each person, through their respective heads illustrated in Figure 3. The D-WASP architecture helps to more effectively discern multiple people in a crowded setting due to its multi-level and multi-scale representations, contributing to SOTA performance.

The design of the D-WASP module relies on a combina-

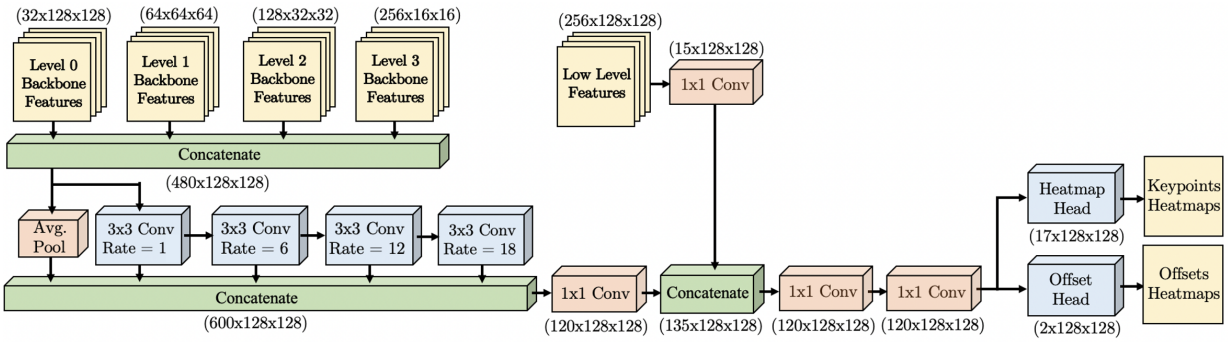


Figure 3. The D-WASP disentangled waterfall module. The inputs are 32, 64, 128, and 256 features maps from all four levels of the HRNet backbone and low-level features from the initial layers of the framework. The module outputs both the keypoints and offsets heatmaps.

tion of atrous and adaptive convolutions. Atrous convolutions are utilized in the initial stages to expand the FOV by performing a filtering cascade at increasing rates to gain efficiency. The waterfall modules are designed to create a waterfall flow, initially processing the input and then creating a new branch. D-WASP goes beyond the cascade approach of [10] by combining all streams from all its branches and the average pooling layer from the original input. Additionally, our module incorporates a larger number of scales compared to WASPv2 [4] by adopting all 480 feature maps from all levels of the HRNet feature extractor. Adaptive convolutions are used to better infer the individual keypoints and offset heatmaps during the regression process by providing context around the vicinity of each detected joint and strengthening the relationship between associated joints.

#### 3.1.1 Waterfall Features and Adaptive Convolutions

The D-WASP module operation begins with the concatenation $g_0$ of all feature maps $f_i$ from the HRNet feature extractor, where $i = 0, 1, 2, 3$ indicates the levels at different scales and summation is overloaded for concatenation:

$$g_0 = \sum_{i=0}^{3}(f_i) \tag{1}$$

Following the concatenation of all feature maps, the waterfall processing is described as follows:

$$f_{Waterfall} = W_1 \circledast (\sum_{i=1}^{4}(W_{d_i} \circledast g_{i-1}) + AP(g_0)) \tag{2}$$

$$f_{maps} = W_1 \circledast (W_1 \circledast (W_1 \circledast f_{LLF} + f_{Waterfall}) \tag{3}$$

where $\circledast$ represents convolution, $g_0$ is the input feature map, $g_i$ is the feature map from the $i^{th}$ atrous convolution, $AP$ is the average pooling operation, $f_{LLF}$ are the low-level feature maps, and $W_1$ and $W_{d_i}$ represent convolutions of kernel size $1\times1$ and $3\times3$ with dilations of $d_i = [1, 6, 12, 18]$, as shown in Figure 3. After concatenation, the feature maps are combined with low level features. The last $1\times1$ convolution reduces the number of feature maps down to one quarter of the number in the combined input feature maps.

Finally, the D-WASP module output $f_{D-WASP}$ is obtained from the multi-scale adaptive convolutional regression, where adaptive convolution is defined as:

$$\mathbf{y}(c) = \sum_{i=1}^{9}(\mathbf{w_i}\mathbf{x}(g_i^c + c)) \tag{4}$$

where $c$ is the center pixel of the convolution, $\mathbf{y}(c)$ represents the output of the convolution for input $\mathbf{x}$, $\mathbf{w_i}$ are the kernel weights for the the center pixel its neighbors, and $g_i^c$ is the offset of the $i^{th}$ activated pixel. In the adaptive convolutions, the offsets $g_i^c$ are adopted in a parametric manner as an extension of spatial transformer networks [18].

#### 3.1.2 Disentangled Adaptive Regression

The regression stage for multi-person pose estimation is considered the most challenging and a bottleneck in performance for bottom-up methods. To address the limitation of regression, additional processing may utilize pose candidates, post-processing matching schemes, proximity matching, and statistical methods, however these may be computationally expensive or limited in effectiveness.

D-WASP expands on the idea of regression by focus, by not only learning disentangled representations for each of the $K$ joints, but also using multiple scales to infer each representation for all keypoints from multiple adaptively activated pixels. This configuration gives each regression a more robust contextual information of the keypoint region, and results in a more accurate spatial representation.

The multi-scale approach proposed by the D-WASP module, allows BAPose to regress person detections and keypoints with a larger FOV, increasing the network capability to infer joints association through the use of adaptive convolutions. Differently than the WASPv2 [4] decoder stage that only extracts the heatmaps for joints, the

D-WASP multi-scale disentangled adaptive regression determines both the keypoint heatmaps and the final offset heatmaps that are used to regress the position of each individual in the image and their respective joints.

In addition, the integration of the multi-scale feature maps in the disentangled adaptive regression utilizes multiple resolutions at the regression stage, allowing the network to better infer the locations of people and their joints. As a consequence, BAPose demonstrates superior performance (see Section 6), especially in challenging scenarios that include large numbers of people in close proximity.

# 4. Datasets

We evaluated the BAPose method on two datasets for 2D multi-person pose estimation: Common Objects in Context (COCO) [25] and CrowdPose [22]. The large and most commonly adopted COCO dataset [25] consists of over 200K images with more than 250K instances of labelled people keypoints. The keypoint labels consist of 17 keypoints including all major joints in the torso and limbs, as well as facial landmarks of nose, eyes, and ears. The dataset is challenging dataset due to the large number of images in a diverse set of scales and occlusion for poses in the wild.

The CrowdPose dataset [22] is more challenging due to crowds and low separation among individuals. The dataset contains 10K images for training, 2K images for validation, and 20K images for testing. In addition to joints annotations, it also contains body part occlusions. We follow evaluation procedures from [11] and [14].

We generated ideal Gaussian maps for the joints ground truth locations during training, which is a more effective strategy for training loss assessment compared to single points at joint locations. As a consequence, the BAPose was outputs heatmap locations for each joint. The value of $\sigma = 3$ was adopted, generating a well defined Gaussian response for both the ground truth and keypoint predictions, with a decent separation of keypoints and avoiding large overlapping of keypoints.

# 5. Experiments

BAPose experiments followed standard metrics set by each dataset, and same procedures applied by [11], and [14].

## 5.1. Metrics

For the evaluation of BAPose, the evaluation is done based on the Object Keypoint Similarity metric (OKS).

$$OKS = \frac{(\sum_i e^{-d_i^2/2s^2k_i^2})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{5}$$

where, $d_i$ is the Euclidian distance between the estimated keypoint and its ground truth, $v_i$ indicates if the keypoint is visible, $s$ is the scale of the corresponding target, and $k_i$ is the falloff control constant. Since the OKS measurement is adopted by both datasets and is similar to the intersection over the union (IOU), we report our OKS results as the Average Precision (AP) for the IOUs for all instances between 0.5 and 0.95 ($AP$), at 0.5 ($AP^{50}$) and 0.75 ($AP^{75}$), as well as instances of medium ($AP^M$) and large size ($AP^L$) for the COCO dataset. For the CrowdPose dataset, we report easy ($AP^E$), medium ($AP^M$,) and hard size ($AP^H$) instances, as well as the overall Average Recall ($AR$), including for medium ($AR^M$) and large ($AR^L$) instances.

## 5.2. Parameter Selection

We use a set of dilation rates of $r = \{1, 6, 12, 18\}$ for the D-WASP module, similar to [4], and train the network for 140 epochs. The learning rate is initialized at $10^{-3}$ and is reduced by an order of magnitude in two steps at 90 and 120 epochs. The training procedure includes random rotation $[-30°, 30°]$, random scale $[0.75, 1.5]$, and random translation $[-40, 40]$, mirroring procedures followed by [14]. All experiments were performed using PyTorch on Ubuntu 16.04. The workstation has an Intel i5-2650 2.20GHz CPU with 16GB of RAM and an NVIDIA Tesla V100 GPU.

# 6. Results

This section presents BAPose results and compares on two large datasets with SOTA methods.

## 6.1. Experimental results on the CrowdPose dataset

We performed training and testing on the CrowdPose dataset, a difficult challenge due to the high occurrence of crowds in the images. Results are shown in Table 1.

Our BAPose method significantly improves upon the performance of SOTA methods for 512×512 input resolution, achieving AP of 72.2% BAPose outperforms other bottom-up approaches by a wide margin, even those that utilized higher input resolutions. BAPose increased the AP of previous bottom-up methods from 65.7% to 72.2% (relative increase of 9.9%) when compared to previous SOTA at the same resolution, that is a 19.0% reduction in error (from 34.3% to 27.8%). The capabilities of the multi-scale approach of BAPose are further exemplified by observing more precise joint estimations with threshold of 75% ($AP^{75}$), drastically reducing the error by 25.7% (from 29.6% to 22.0%) and increasing the previous SOTA AP from 70.4% to 78.0% (relative increase of 10.8%) when compared to the previous SOTA, HRNet-W32 [14].

Additionally, BAPose outperforms networks that utilize top-down approaches by a significant margin increasing from 70.0% to 72.2%. Differently than top-down methods, BAPose does not rely on ground truth for person detection and has to infer the location of all individuals in a modular, single-pass process. For the CrowdPose dataset, BAPose's

Figure 4. Pose estimation examples using BAPose with the CrowdPose dataset.

| Method | Input Size | Approach | AP | $AP^{50}$ | $AP^{75}$ | $AP^{E}$ | $AP^{M}$ | $AP^{H}$ |
|---|---|---|---|---|---|---|---|---|
| **BAPose (W32)** | 512 | BU | **72.2%** | **89.6%** | **78.0%** | **79.9%** | **73.4%** | **61.3%** |
| MIPNet [20] | 512 | TP | 70.0% | - | - | - | - | - |
| HRNet-W48 [14] | 640 | BU | 67.3% | 86.4% | 72.2% | 74.6% | 68.1% | 58.7% |
| JC SPPE [22] | - | TP | 66.0% | 84.2 | 71.5 | 75.5% | 66.3% | 57.4% |
| HigherHRNet-W48 [11] | 640 | BU | 65.9% | 86.4% | 70.6% | 73.3% | 66.5% | 57.9% |
| HRNet-W32 [14] | 512 | BU | 65.7% | 85.7% | 70.4% | 73.0% | 66.4% | 57.5% |
| Mask R-CNN [16] | - | BU | 60.3% | - | - | 69.4% | 57.9% | 45.8% |

Table 1. BAPose results and comparison with SOTA methods for the CrowdPose dataset for testing. TP and BU represent the Top-Down and Bottom-Up approaches, respectively.

| Method | Input Size | GFLOPs | Params (M) |
|---|---|---|---|
| HRNet-W32 [14] | 512 | 45.4 | 29.6 |
| **BAPose-W32** | 512 | 56.8 | 30.3 |
| HRNet-W48 [14] | 640 | 141.5 | 65.7 |
| HigherHRNet-W48 [11] | 640 | 154.3 | 63.8 |
| **BAPose-W48** | 640 | 183.2 | 67.4 |

Table 2. GFLOPs and number of parameters comparison.

performance is superior to networks utilizing higher resolution inputs of 640×640 [14], [11] while processing the less computationally expensive 512×512 resolution.

It is important to observe that the BAPose framework was able to achieve this significant increase in AP for the CrowdPose dataset while utilizing a backbone smaller (HRNet-W32 [47]) compared to the previous SOTA deploying a larger backbone (HRNet-W48 [47]), reducing the number of parameters by 54.9% and GFLOPs by 67.9%.

Figure 4 illustrates successful detections of multi-person pose for the CrowdPose test set. The examples demonstrate how effectively BAPose deals with occlusions, close proximity of individuals, as well as detections at different scales.

### 6.2. Experimental results on the COCO dataset

We next performed training and testing on the COCO dataset, which is challenging due to the large number of diverse images with multiple people in close proximity, and additionally includes images lacking a person instance.

We first compared BAPose with SOTA methods for the COCO validation and test-dev datasets. The validation results in Table 3 show that BAPose achieves significant improvement over the previous SOTA for both input resolutions. The BAPose results at the former resolution are obtained with a significantly lower computational cost compared to methods with higher resolution, as shown in Table 2, while achieving comparable results to higher resolution.

The incorporation of the D-WASP module achieves an increased overall accuracy of 69.1% when using single-scale testing significantly increasing the AP accuracy at 512×512 resolution by 1.6%. For multi-scale testing BAPose achieves 71.9% improving upon previous SOTA of 70.7%, that is an increase in accuracy of 1.7%. This performance increase represents an error reduction of 3.4% (from 32.0% to 30.9%) for single-scale and 4.1% error reduction for multi-scale (from 29.3% to 28.1%).

BAPose improves the accuracy of the previous SOTA in all keypoint estimation metrics and IOU for the COCO

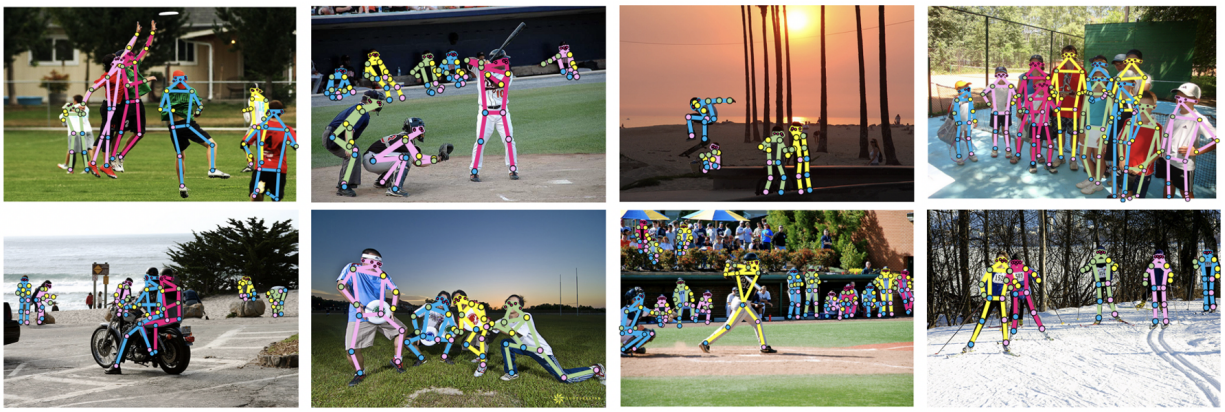

Figure 5. Pose estimation results using BAPose with the COCO dataset.

| Method | Input Size | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ | $AR$ |
|---|---|---|---|---|---|---|---|
| Single-Scale Testing | | | | | | | |
| **BAPose (W48)** | 640 | **71.6%** | **88.6%** | **78.3%** | **67.3%** | **78.7%** | **76.5%** |
| HRNet-W48 [14] | 640 | 71.0% | 88.3% | 77.4% | 66.7% | 78.5% | 76.0% |
| HigherHRNet-W48 [11] | 640 | 69.9% | 87.2% | 76.1% | - | - | - |
| **BAPose (W32)** | 512 | **69.1%** | **87.0%** | **75.6%** | **63.1%** | **78.6%** | **73.7%** |
| HRNet-W32 [14] | 512 | 68.0% | 86.7% | 74.5% | 62.1% | 77.7% | 73.0% |
| HigherHRNet-W32 [11] | 512 | 67.1% | 86.2% | 73.0% | - | - | - |
| HGG [19] | 512 | 60.4% | 83.0% | 66.2% | - | - | 64.8% |
| Multi-Scale Testing | | | | | | | |
| **BAPose (W48)** | 640 | **72.7%** | **88.6%** | **79.1%** | **69.3%** | 78.4% | **77.9%** |
| HRNet-W48 [14] | 640 | 72.3% | 88.3% | 78.6% | 68.6% | **78.6%** | 77.7% |
| HigherHRNet-W48 [11] | 640 | 72.1% | 88.4% | 78.2% | - | - | - |
| **BAPose (W32)** | 512 | **71.9%** | **88.3%** | **77.8%** | **67.2%** | **79.1%** | **76.6%** |
| HRNet-W32 [14] | 512 | 70.7% | 87.7% | 77.1% | 66.2% | 77.8% | 75.9% |
| HigherHRNet-W32 [11] | 512 | 69.9% | 87.1% | 76.0% | - | - | - |
| HGG [19] | 512 | 68.3% | 86.7% | 75.8% | - | - | 72.0% |

Table 3. BAPose results and comparison with SOTA methods for the COCO dataset for validation.

dataset. Most of the performance improvements of BAPose are attributed to performing better on harder detections and more refined predictions at $AP^{75}$. The results on the COCO validation dataset, in Table 3, show the greater capability of BAPose to detect more complex and harder poses while still using a smaller resolution in the input image.

We also trained and tested BAPose-W48 at 640×640 resolution, achieving 71.6% accuracy for the COCO validation set with single scale testing and 72.7% with multi-scale testing, improving the previous SOTA by 0.8% and 0.6%, respectively. This improvement represents an error reduction of 2.1% and 1.4% compared to HRNet-w48 [14]. However, larger resolution models require much higher computational resources, as illustrated by the GFLOPs and memory requirements in Table 2. Compared to BAPose-W32, HRNet-W48 requires 249.1% the number of GFLOPs and HigherHRNet-W48 requires 271.7% the number of GFLOPs, demonstrating that BAPose-W32 results in a better trade-off between accuracy and computational cost.

Figure 5 presents examples of pose estimation results for the COCO dataset. BAPose effectively locates symmetric body joints and avoids confusion due to occlusion between individuals. This is illustrated in harder to detect joints such as ankles and wrists. Overall, the BAPose results demonstrate robustness for pose estimation in challenging conditions, such as images that include multiple individuals with high overlapping ratio combined with shadows or darker images, or partial pose present in the image.

For the larger COCO test-dev dataset, shown in Table 4, BAPose achieves again new SOTA performance over methods using input resolutions of 512×512. Our method obtained an overall precision of 68.0% when using single-

| Method | Input Size | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ | $AR$ |
|---|---|---|---|---|---|---|---|
| Single-Scale Testing | | | | | | | |
| **BAPose (W48)** | 640 | **70.3%** | **89.6%** | **77.5%** | **65.9%** | **77.1%** | **75.4%** |
| HRNet-W48 [14] | 640 | 70.0% | 89.4% | 77.3% | 65.7% | 76.9% | 75.4% |
| HigherHRNet-W48 [11] | 640 | 68.4% | 88.2% | 75.1% | 64.4 | 74.2 | - |
| **BAPose (W32)** | 512 | **68.0%** | **88.0%** | **74.8%** | **62.4%** | **76.6%** | **72.9%** |
| HRNet-W32 [14] | 512 | 67.3% | 87.9% | 74.1% | 61.5% | 76.1% | 72.4% |
| SPM [31] | - | 66.9% | 88.5% | 72.9% | 62.6% | 73.1% | - |
| CenterNet-HG [46] | 512 | 63.0% | 86.8% | 69.6% | 58.9% | 70.4% | - |
| OpenPose [7] | - | 61.8% | 84.9% | 67.5% | 57.1% | 68.2% | 66.5% |
| Multi-Scale Testing | | | | | | | |
| **BAPose (W48)** | 640 | **71.2%** | **89.4%** | **78.1%** | **67.4%** | 76.8% | **76.8%** |
| HRNet-W48 [14] | 640 | 71.0% | 89.2% | 78.0% | 67.1% | **76.9%** | 76.7% |
| HigherHRNet-W48 [11] | 640 | 70.5% | 89.3% | 77.2% | 66.6% | 75.8% | - |
| Point-set Anchors [42] | 640 | 68.7% | 89.9% | 76.3% | 64.8% | 75.3% | 74.8% |
| **BAPose (W32)** | 512 | **70.4%** | **89.3%** | **77.4%** | **66.0%** | **76.9%** | **75.6%** |
| HRNet-W32 [14] | 512 | 69.6% | 89.0% | 76.6% | 65.2% | 76.5% | 75.1% |
| HGG [19] | 512 | 67.6% | 85.1% | 73.7% | 62.7% | 74.6% | 71.3% |

Table 4. BAPose results and comparison with SOTA methods for the COCO dataset for test-dev.

scale testing and 70.4% when using multi-scale testing, which are relative improvements over SOTA of 1.0% for single (from 67.3% to 68.0%) and 1.1% (from 69.6.% to 70.4%) for multi scale testing. BAPose reduced the error at the 512×512 resolution by 2.1% (from 32.7% to 32.0%) for single-scale and 2.6% (from 30.4% to 29.6%) for multi-scale testing. When training and testing at the 640×640 resolution, BAPose-W48 achieved accuracies of 70.3% for single-scale testing and 71.2% when using single-scale multi-scale testing, an improvement of 0.4% for single-scale testing and 0.3% for multi-scale testing compared to the previous SOTA, reducing the error by 1.0% and 0.7%, respectively. These results further demonstrate BAPose most significant improvements are in smaller and harder targets consistent with the findings from the validation dataset.

## 7. Ablation Study on Waterfall Modules

We performed an ablation study comparing the D-WASP module vs. WASP [3] vs. WASPv2 [4], and vs. the baseline without the waterfall module [39]. Since HRNet and UniPose are methods for single person and top-down pose estimation, we adapted all models to the same decoding procedures used by BAPose. We adopted a HRNet-W32 for all backbones to have a direct comparison. Table 5 demonstrate the results. For both datasets, we found that D-WASP obtained a significant and consistent improvement over SOTA.

The improvement obtained with BAPose using the D-WASP module was more significant for the CrowdPose dataset due to the more complex settings with multiple people and bigger need for people detection over multiple scales and multiple occlusions, further demonstrating the higher robustness of the D-WASP module for more complex tasks of bottom-up pose estimation and more keypoints.

| Method | Waterfall Module | COCO AP | CrowdPose AP |
|---|---|---|---|
| HR-Net[39] | None | 68.0% | 65.7% |
| UniPose[3] | WASP | 68.2% | 67.2% |
| OmniPose[4] | WASPv2 | 68.5% | 69.0% |
| **BAPose** | **D-WASP** | **69.1%** | **72.2%** |

Table 5. Results on the COCO and CrowdPose datasets. All networks use HRNet-W32 backbone and different waterfall modules.

## 8. Conclusion

We presented the BAPose method for bottom-up multi-person pose estimation. The BAPose network includes the novel D-WASP module that combines multi-scale features obtained from the waterfall flow with the person detection capability of the disentangled adaptive regression. BAPose is a end-to-end trainable, single-pass architecture that without anchors, prior person detections, or postprocessing.

The results demonstrate SOTA performance for both the COCO and CrowdPose datasets in all metrics, superior capability of person detection and pose estimation in densely populated images, and the robustness of the BAPose framework, when estimating a large number of pose instances in crowds, allowing the expansion of our framework to the broader task of complete body pose estimation including hands, feet, and facial landmarks.

## References


[1] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018.

[2] Bruno Artacho and Andreas Savakis. Waterfall atrous spatial pooling architecture for efficient semantic segmentation. *Sensors*, 19(24):5361, 2019.

[3] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Bruno Artacho and Andreas Savakis. Omnipose: A multi-scale framework for multi-person pose estimation. 2021.

[5] Bruno Artacho and Andreas Savakis. Unipose+: A unified framework for 2d and 3d human pose estimation in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[6] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 468–475, 2017.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-Person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[8] Zhongzheng Cao, Rui Wang, Xiangyang Wang, Zhi Liu, and Xiaoqiang Zhu. Improving human pose estimation with self-attention generative adversarial networks. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 567–572. IEEE, 2019.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution and fully connected cfrs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–845, 2018.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*, pages 801–818, 2018.

[11] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T. Tan. Occlusion-aware networks for 3d human pose estimation in video. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[13] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1831–1840, 2017.

[14] Zigang Geng, Ke Sun, Zhaoxiang Zhang Bin Xiao, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021.

[15] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron, 2018.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[17] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[19] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *European Conference on Computer Vision (ECCV)*, 2019.

[20] Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3122–3131, October 2021.

[21] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Haoshu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and A new benchmark. *CoRR*, abs/1812.00324, 2018.

[23] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[24] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11313–11322, October 2021.

[25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[26] J. Long, E. Shelhamer, and T. Darrel. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016.

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.

[29] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 3d human pose estimation with 2D marginal heatmaps.

In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1477–1485, 2019.

[30] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. Pose partition networks for multi-person pose estimation. In *Proceedings of the european conference on computer vision (eccv)*, pages 684–699, 2018.

[31] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6951–6960, 2019.

[32] Guanghan Ning and Heng Huang. Lighttrack: A generic framework for online top-down human pose tracking. *arXiv preprint arXiv:1905.02822*, 2019.

[33] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *European Conference on Computer Vision (ECCV)*, 2018.

[34] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[35] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[37] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[38] Zhihui Su, Ming Ye, Guohui Zhang, Lei Dai, and Jianda Sheng. Cascade feature aggregation for human pose estimation. *arXiv preprint arXiv:1902.07837*, 2019.

[39] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014.

[41] Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13086–13095, 2020.

[42] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *European Conference on Computer Vision*, pages 527–544. Springer, 2020.

[43] Shih-En Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[45] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*, 2019.

[46] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.

[47] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. MonoCap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):901–914, 2018.