sensors

MDPI

Article

Full-BAPose: Bottom Up Framework for Full Body Pose Estimation

Bruno Artacho and Andreas Savakis *

Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA
* Correspondence: andreas.savakis@rit.edu

Abstract: We present Full-BAPose, a novel bottom-up approach for full body pose estimation that achieves state-of-the-art results without relying on external people detectors. The Full-BAPose method addresses the broader task of full body pose estimation including hands, feet, and facial landmarks. Our deep learning architecture is end-to-end trainable based on an encoder-decoder configuration with HRNet backbone and multi-scale representations using a disentangled waterfall atrous spatial pooling module. The disentangled waterfall module leverages the efficiency of progressive filtering, while maintaining multi-scale fields-of-view comparable to spatial pyramid configurations. Additionally, it combines multi-scale features obtained from the waterfall flow with the person-detection capability of the disentangled adaptive regression and incorporates adaptive convolutions to infer keypoints more precisely in crowded scenes. Full-BAPose achieves state-of-the art performance on the challenging CrowdPose and COCO-WholeBody datasets, with AP of 72.2% and 68.4%, respectively, based on 133 keypoints. Our results demonstrate that Full-BAPose is efficient and robust when operating under a variety conditions, including multiple people, changes in scale, and occlusions.

Keywords: human pose estimation; whole-body pose estimation; deep learning; multi-scale representations; waterfall module; atrous spatial pooling; adaptive convolutions; disentangled keypoint regression

check for updates

Citation: Artacho, B.; Savakis, A. Full-BAPose: Bottom Up Framework for Full Body Pose Estimation. *Sensors* **2023**, *23*, 3725. https://doi.org/10.3390/s23073725

Academic Editor: Sheryl Berlin Brahnam

Received: 10 January 2023
Revised: 14 March 2023
Accepted: 23 March 2023
Published: 4 April 2023

1. Introduction

Full body pose estimation is a challenging task in computer vision that finds applications in action recognition, human–computer interaction, and sign language recognition. Human pose estimation methods have received considerable attention for 2D pose estimation, including multistage architectures, such as stacked hourglass (HG) networks [1] and Convolutional Pose Machines (CPMs) [2], and encoder-decoder architectures, such as UniPose [3] and High-Resolution Network (HRNet) [4]. Multi-person pose estimation is particularly challenging due to joint occlusions and the large number of degrees of freedom in the human body. State-of-the-art (SOTA) methods for multi-person pose estimation can be top-down or bottom-up. Top-down approaches use detectors to localize instances of persons in the image and then perform single-person pose estimation for each instance. While top-down approaches generally achieve higher accuracy, they require external detectors that make the process slower and more costly. Bottom up methods either detect all keypoints and group them for each individual [5] or directly regress the keypoints to each person in the image [6]. Thus, bottom-up approaches require a single processing stage and are more efficient.

Full-BAPose is a full-body version of BAPose [7]. It is a bottom-up framework named after “Basso verso l’Alto” (bottom-up in Italian). The Full-BAPose method deals with the broader task of pose estimation for the whole body and is based on 133 keypoints for the body, hands, feet, and facial landmarks. Examples of Full-BAPose pose estimation are shown in Figure 1, illustrating the method’s effectiveness under various conditions. The

Full-BAPose network is single-stage and end-to-end trainable, building upon the successful approaches by UniPose for single-person 2D pose [3], UniPose+ for single-person 2D and 3D pose [8], and OmniPose for multi-person top-down pose [9]. Full-BAPose achieves SOTA results without requiring post-processing, intermediate supervision, multiple iterations, or anchor poses. The main contributions of Full-BAPose are:

- A single-pass, end-to-end trainable, multi-scale approach for a bottom-up multi-person full-body pose estimation framework that achieves SOTA results;
- Our bottom-up approach operates without requiring a separate detector because it combines multi-scale waterfall features with disentangled adaptive convolutions and includes a decoder to determine the joints of individuals in crowded scenes.

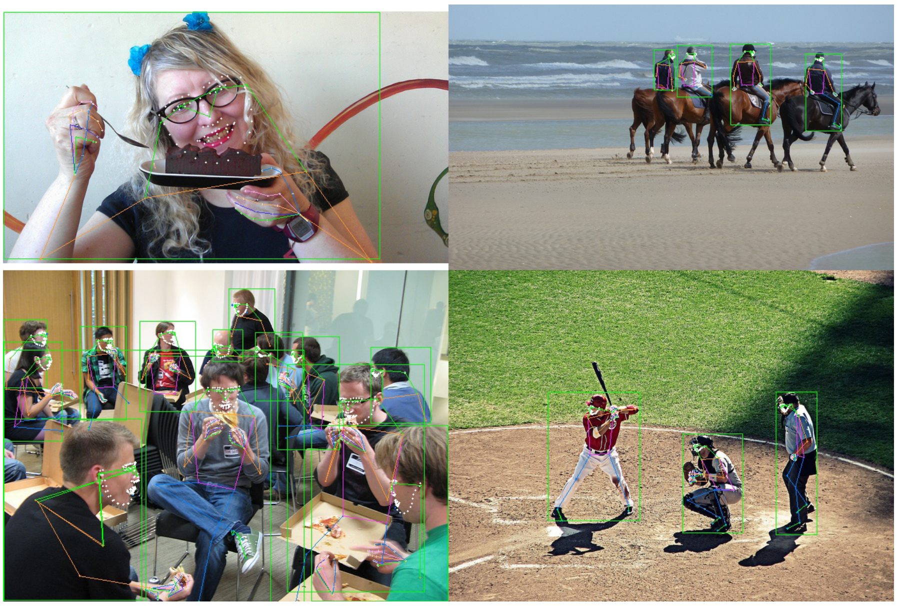

Figure 1. Full-body pose estimation examples with our Full-BAPose method showing a single person and multiple people at various scales and occlusions. The bottom-up approach determines the bounding boxes (in green) and the person pose estimation.

2. Related Work

Convolutional Neural Networks (CNNs) have enabled impressive advances for human pose estimation [5,8,10,11]. Convolutional pose machines [2] utilize a CNN with stages that refine joint detection. Part Affinity Fields (PAFs) in OpenPose [5] capture relationships between joints for improved human pose estimation. The stacked hourglass network [1] cascades hourglass structures for pose estimation refinement. The multi-context approach in [12] augments the backbone with residual units to increase the receptive Field-of-View (FOV) but increases complexity due to postprocessing with Conditional Random Fields (CRFs). Aiming to develop a unified framework for single-person pose estimation, UniPose [3] and UniPose+ [8] combine bounding box generation and joint detection in a unified, one-pass network. This is made feasible by the Waterfall Atrous Spatial Pooling (WASP) module [13], which allows a larger field-of-view and multi-scale representations to determine context.

The multi-scale approach of HRNet includes both high and low resolutions to obtain a larger FOV, while higher HRNet [14] combined the HRNet structure with multi-resolution pyramids. The Multi-Stage Pose Network (MSPN) [15] combines the cross-stage feature aggregation and coarse-to-fine supervision. The Distribution-Aware coordinate Representation of Keypoints (DARK) method [16] refines their decoder in order to reduce the inference error at the decoder stage.

The work in [17] utilized graphs to extract contextual information for pose. Cascade Feature Aggregation (CFA) [18] employed the cascade approach for semantic information in pose estimation. Generative Adversarial Networks (GANs) were used in [19] to learn context for poses. More recently, methods such as TokenPose [20] are investigating transformer networks to determine global dependencies for pose estimation. Neural architecture search was explored in ZoomNAS [21] to obtain separate networks for pose estimation on the body, hands, and face.

A limitation of top-down approaches is the requirement of an independent module for the detection of instances of humans in the frame. LightTrack [22], for instance, applies YOLOv3 [23] to detect subjects prior to the detection of joints for pose estimation, while LCR-Net [10] applies multiple branches for detection by using the Detectron [24] and the arrangement of joints during classification.

2.1. Bottom-Up Approaches

Bottom-up approaches face the bigger challenge of detecting the joints of multiple people without an external person detector used by top-down methods. Bottom-up methods try to associate detected keypoints with persons in the image. This problem was cast in [25,26] as integer linear programming, but this type of optimization requires a significant processing time. OpenPose [5] used part affinity fields in a breakthrough approach to grouping keypoints for each person. This approach was extended by Pif-Paf [27] and associative embedding [28]. PersonLab [29] adopted Hough voting and [30] Hierarchical Graphical Clustering (HGG). The works in [31,32] consider dense regression of pose candidates but face the limitation of lower localization accuracy that requires post-processing to improve the results. In a related approach, Ref. [33] considered a mixture density network before regression. Alternatively, the Disentangled Keypoint Regression (DEKR) [6] method learns disentangled representations for each keypoint and utilizes adaptively activated pixels so that each representation focuses on the corresponding keypoint area.

2.2. Multi-Scale Feature Representations

The pooling operations in CNNs present a challenge for pose estimation due to the resolution reduction. To overcome this problem, Fully Convolutional Networks (FCNs) [34] utilize upsampling to increase the resolution of the feature maps to the size of the input image. DeepLab [35] adopted atrous convolutions in the multi-scale Atrous Spatial Pyramid Pooling (ASPP) framework that maintains the size of the receptive fields. DeepLab applies atrous convolutions in four parallel branches with different rates and combines them at the original image resolution.

Presenting an improvement over ASPP, the WASP module [3,13] generates multi-scale features in an efficient manner by creating a waterfall flow. Instead of immediately parallelizing all branches of the input stream, the WASP module first processes through a filter and then creates a new branch. The waterfall flow extends the cascade approach by combining the streams from all its branches to create a multi-scale representation.

3. Full-BAPose Architecture

The Full-BAPose bottom-up architecture, as illustrated in Figure 2, consists of a single-pass, end-to-end trainable network with HRNet backbone and our Disentangled Waterfall Atrous Spatial Pyramid (D-WASP) module with an integrated decoder for full-body multi-person pose estimation. The input image is fed in the HRNet backbone and the extracted feature maps from the HRNet levels are processed by the D-WASP module to obtain enhanced multi-scale representations. The D-WASP module includes an integrated decoder to obtain 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 two offset maps for the identification of person instances and the 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 throughout the network.

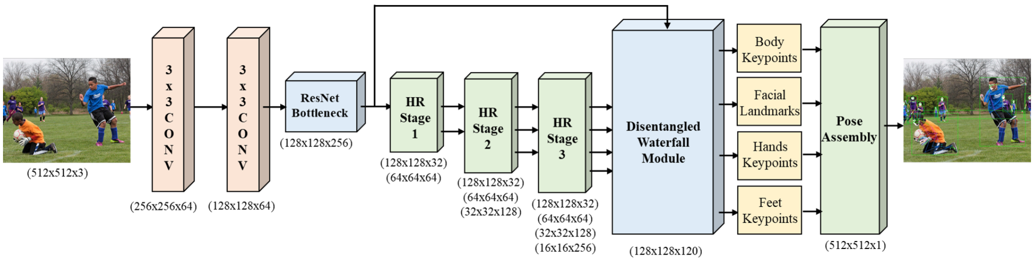

Figure 2. Full-BAPose architecture for whole-body multi-person pose estimation. The input color image is fed through the HRNet backbone for initial feature extraction. The feature sizes are denoted by the two spatial dimensions first and the channel dimension last, e.g., (128 × 128 × 32) denotes feature size of 128 × 128 with 32 channels. The HRNet features are combined by the D-WASP module and a decoder utilizing adaptive convolutions generates the detection bounding boxes and the keypoints for the hand, head, feet, and entire body, i.e., 133 keypoints and 4 bounding boxes for each person instance.

Our approach includes several innovations that contribute to increased accuracy. The D-WASP module combines atrous convolutions and the waterfall architecture to improve multi-scale representations with contextual information by the processing of feature maps at multiple rates of dilation. Our architecture allows a larger FOV in the encoder and integrates disentangled adaptive convolutions in the decoder, facilitating the detection of multi-person instances and their keypoints in a single-pass. Further, our network demonstrates the ability to deal with a large number of persons in the image due to feature extraction at multiple scales. Finally, the modular nature of the architecture facilitates easy implementation and training of the network.

Full-BAPose utilizes a bottom-up approach without relying on external detectors to locate faces or people. The D-WASP module combines 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 enhanced "Disentangled Waterfall Atrous Spatial Pyramid" module 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 for multi-scale representations. Low-level and high-level features are represented at the same resolution, achieving refined joint estimation. Furthermore, the D-WASP module uses adaptive convolution blocks to infer heatmaps for joint localization and offset maps for person instance regression.

The design of the D-WASP module relies on a combination 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 [36] 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 previous versions of the waterfall module [3,9] by adopting all 480 feature maps from all levels of the HRNet feature extractor. Adaptive convolutions are used for improved estimation of individual keypoints and offset heatmaps during the regression process by providing context around each detected joint and associated joints.

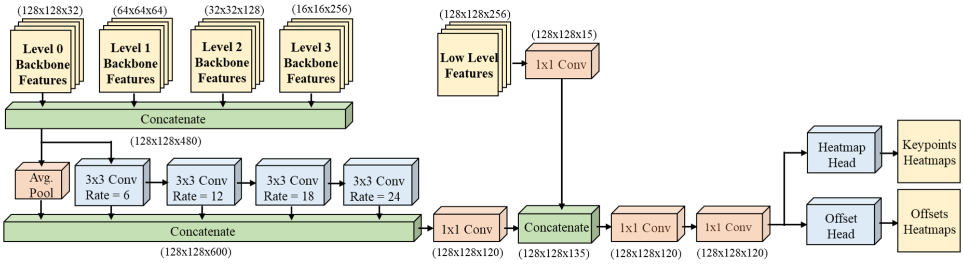

Figure 3. The D-WASP disentangled waterfall module. The feature sizes are denoted by the two spatial dimensions, followed by the channel dimension. The inputs are 32, 64, 128, and 256 feature maps from all four levels of the HRNet backbone, as illustrated in Figure 2, and low-level features from the initial layers of the framework. The module processes the backbone features at different rates of dilation in a waterfall fashion and outputs both the keypoints and offset heatmaps for each person instance.

3.1.1. Waterfall Features and Adaptive Convolutions

The D-WASP module operation starts 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 indicates concatenation:

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

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 ith 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 × 1 and 3 × 3 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 × 1 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:

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

where c is the center pixel of the convolution, $y(c)$ represents the output of the convolution for input x, w_i are the kernel weights for the center pixel and its neighbors, and g_i^c is the offset of the ith activated pixel. In the adaptive convolutions, the offsets g_i^c are adopted in a parametric manner as an extension of spatial transformer networks [37].

3.1.2. Disentangled Adaptive Regression

The regression stage for multi-person pose estimation is a bottleneck in performance for bottom-up methods. To address this limitation, additional processing may utilize pose candidates, post-processing matching, proximity matching, or statistical methods. However, these may be computationally expensive or limited in effectiveness.

D-WASP leverages the idea of regression by focus by learning disentangled representations for each of the K joints and using multiple scales to infer each representation for all keypoints from multiple adaptively activated pixels. This configuration offers more robust contextual information of the keypoint region and results in a more accurate spatial representation. 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.

4. Datasets

We evaluated BAPose on two datasets for 2D multi-person pose estimation: Common Objects in Context (COCO) [38] and CrowdPose [39]. The large and most commonly adopted COCO dataset [38] 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, including nose, eyes, and ears. The dataset is challenging due to the large number of images in a diverse set of scales and occlusion for poses in the wild.

The CrowdPose dataset [39] 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 joint annotations, it also contains body part occlusions. We follow evaluation procedures from [6,14].

The COCO-WholeBody dataset consists of images from the large COCO dataset labeled to contain facial landmarks, feet and hands keypoints, and the original body keypoints, totaling 133 keypoints to be extracted by the network for multiple people in each image. The dataset contains over 130K instances of facial landmarks, hands, and feet for a total of over 800K hand keypoints and 1.4M facial landmarks.

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 output 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 avoidance of large overlapping of keypoints.

5. Experiments

BAPose experiments followed standard metrics set by each dataset and the same procedures applied by [6,14].

5.1. Metrics

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

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

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 [9], 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 [6]. All experiments were performed using PyTorch on Ubuntu 16.04. The workstation has an Intel i5-2650 2.20 GHz CPU with 16 GB of RAM and an NVIDIA Tesla V100 GPU.

6. Results

This section presents body pose results with BAPose [7] and full-body pose results with Full-BAPose, as well as comparisons with SOTA methods.

6.1. Experimental Results on the CrowdPose Dataset

We performed training and testing of BAPose (for body pose based on 17 keypoints) on the CrowdPose dataset, a difficult challenge due to the high occurrence of crowds in the images. The results are shown in Table 1. Our BAPose-W32 method uses the HRNet-W32 backbone [6]. Figure 4 illustrates successful detections of multi-person pose for the CrowdPose test set. The examples demonstrate how effectively BAPose deals with occlusions, the close proximity of individuals, as well as detections at different scales.

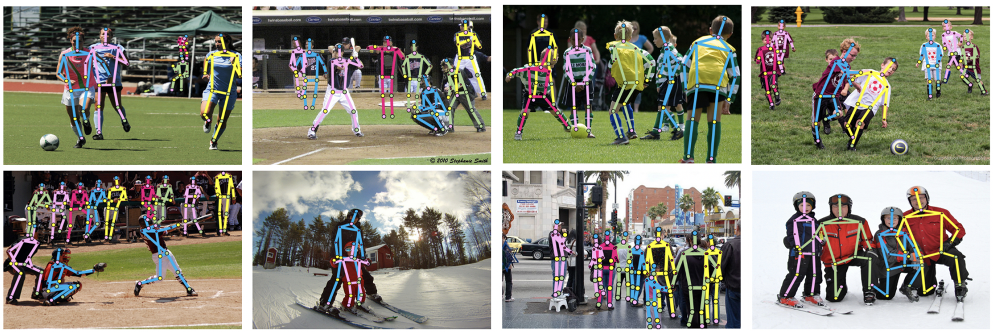

Figure 4. Pose estimation examples using BAPose with the CrowdPose dataset.

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. Best results are in bold.

Method	Input Size	Approach	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
BAPose-W32 (ours)	512	BU	**72.2%**	**89.6%**	**78.0%**	**79.9%**	**73.4%**	**61.3%**
MIPNet [40]	512	TP	70.0%	-	-	-	-	-
HRNet-W48 [6]	640	BU	67.3%	86.4%	72.2%	74.6%	68.1%	58.7%
JC SPPE [39]	-	TP	66.0%	84.2	71.5	75.5%	66.3%	57.4%
HigherHRNet-W48 [14]	640	BU	65.9%	86.4%	70.6%	73.3%	66.5%	57.9%
HRNet-W32 [6]	512	BU	65.7%	85.7%	70.4%	73.0%	66.4%	57.5%
Mask R-CNN [41]	-	BU	60.3%	-	-	69.4%	57.9%	45.8%

BAPose significantly improves upon SOTA methods for 512 × 512 input resolution, achieving an 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, which is a 19.0% reduction in error (from 34.3% to 27.8%). The capabilities of the multi-scale approach of BAPose are further illustrated by observing more precise joint estimations with a 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 [6].

Additionally, BAPose outperforms networks that utilize top-down approaches by a significant margin, increasing from 70.0% to 72.2%. In contrast to 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 performance is superior to networks utilizing higher-resolution inputs of 640 × 640 [6,14] while processing the less computationally expensive 512 × 512 resolution.

We observe that the BAPose framework was able to achieve this significant increase in AP for the CrowdPose dataset while utilizing a backbone that is smaller (HRNet-W32) compared to the previous SOTA deploying a larger backbone (HRNet-W48 [6]), reducing the number of parameters by 54.9% and GFLOPs by 67.9%.

6.2. Experimental Results on the COCO Dataset

We next performed training and testing of BAPose on the COCO dataset, which is challenging due to the large number of diverse images with multiple people in close proximity and 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 2 show that BAPose achieves significant improvement over the previous SOTA for both input resolutions. Our BAPose-W32 and BAPose-W48 methods use HRNet-W32 and HRNet-W48 backbones respectively [6]. The BAPose results at the former resolution are obtained with a significantly lower computational cost compared to methods with a higher resolution, as shown in Table 3, while achieving comparable results to a higher resolution.

Table 2. BAPose results and comparison with SOTA methods for the COCO dataset for validation. Best results are in bold.

Method	**Input Size**	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Single-Scale Testing							
BAPose-W48 (ours)	640	**71.6%**	**88.6%**	**78.3%**	**67.3%**	**78.7%**	**76.5%**
HRNet-W48 [6]	640	71.0%	88.3%	77.4%	66.7%	78.5%	76.0%
HigherHRNet-W48 [14]	640	69.9%	87.2%	76.1%	-	-	-
BAPose-W32 (ours)	512	**69.1%**	**87.0%**	**75.6%**	**63.1%**	**78.6%**	**73.7%**
HRNet-W32 [6]	512	68.0%	86.7%	74.5%	62.1%	77.7%	73.0%
HigherHRNet-W32 [14]	512	67.1%	86.2%	73.0%	-	-	-
HGG [30]	512	60.4%	83.0%	66.2%	-	-	64.8%
Multi-Scale Testing							
BAPose-W48 (ours)	640	**72.7%**	**88.6%**	**79.1%**	**69.3%**	78.4%	**77.9%**
HRNet-W48 [6]	640	72.3%	88.3%	78.6%	68.6%	**78.6%**	77.7%
HigherHRNet-W48 [14]	640	72.1%	88.4%	78.2%	-	-	-
BAPose-W32 (ours)	512	**71.9%**	**88.3%**	**77.8%**	**67.2%**	**79.1%**	**76.6%**
HRNet-W32 [6]	512	70.7%	87.7%	77.1%	66.2%	77.8%	75.9%
HigherHRNet-W32 [14]	512	69.9%	87.1%	76.0%	-	-	-
HGG [30]	512	68.3%	86.7%	75.8%	-	-	72.0%

Table 3. GFLOPs and number of parameters comparison.

Method	**Input Size**	**GFLOPs**	**Params (M)**
HRNet-W32 [6]	512	45.4	29.6
BAPose-W32 (ours)	512	56.8	30.3
HRNet-W48 [6]	640	141.5	65.7
HigherHRNet-W48 [14]	640	154.3	63.8
BAPose-W48 (ours)	640	183.2	67.4

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%, which 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 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 2, 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 a 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 [6]. However, larger resolution models require much higher computational resources, as illustrated by the GFLOPs and memory requirements in Table 3. Compared to BAPose-W32, HRNet-W48 requires a 249.1% increase in the number of GFLOPs, and HigherHRNet-W48 requires a 271.7% increase in 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 a high overlapping ratio combined with shadows or darker images or partial poses present in the image.

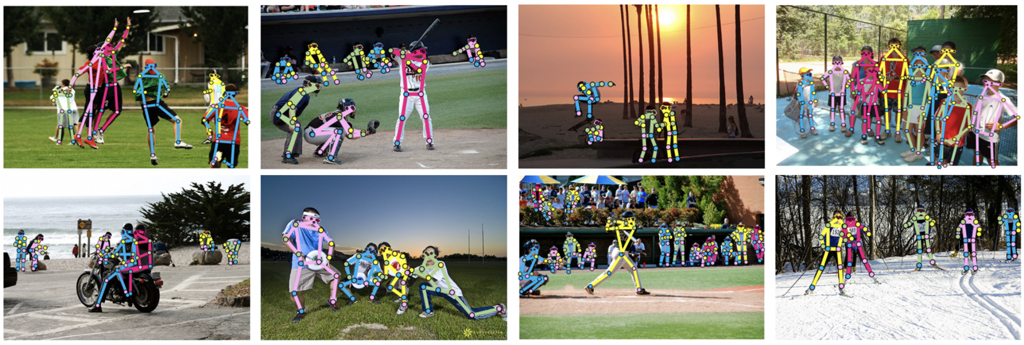

Figure 5. Pose estimation results using BAPose with the COCO dataset.

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-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's most significant improvements are in smaller and harder targets consistent with the findings from the validation dataset.

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

Method	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
			Single-Scale Testing				
BAPose-W48 (ours)	640	**70.3%**	**89.6%**	**77.5%**	**65.9%**	**77.1%**	**75.4%**
HRNet-W48 [6]	640	70.0%	89.4%	77.3%	65.7%	76.9%	75.4%
HigherHRNet-W48 [14]	640	68.4%	88.2%	75.1%	64.4	74.2	-
BAPose-W32 (ours)	512	**68.0%**	**88.0%**	**74.8%**	**62.4%**	**76.6%**	**72.9%**
HRNet-W32 [6]	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 [42]	512	63.0%	86.8%	69.6%	58.9%	70.4%	-
OpenPose [5]	-	61.8%	84.9%	67.5%	57.1%	68.2%	66.5%
			Multi-Scale Testing				
BAPose-W48 (ours)	640	**71.2%**	**89.4%**	**78.1%**	**67.4%**	76.8%	**76.8%**
HRNet-W48 [6]	640	71.0%	89.2%	78.0%	67.1%	**76.9%**	76.7%
HigherHRNet-W48 [14]	640	70.5%	89.3%	77.2%	66.6%	75.8%	-
Point-set Anchors [43]	640	68.7%	89.9%	76.3%	64.8%	75.3%	74.8%
BAPose-W32 (ours)	512	**70.4%**	**89.3%**	**77.4%**	**66.0%**	**76.9%**	**75.6%**
HRNet-W32 [6]	512	69.6%	89.0%	76.6%	65.2%	76.5%	75.1%
HGG [30]	512	67.6%	85.1%	73.7%	62.7%	74.6%	71.3%

6.3. Experimental Results on the COCO-WholeBody Dataset

We trained and tested Full-BAPose on the COCO-WholeBody dataset [44] for the larger task of estimating a full set of keypoints including all body pose keypoints, all joints of feet and hands, and facial landmarks.

The comparison of both the Full-BAPose and OmniPose framework to state-of-the-art methods for the validation dataset is shown in Table 5. The results demonstrate that both architectures present a significant increase compared to the previous state-of-the-art, especially the Full-BAPose framework, achieving an accuracy increase of 13.1% in the overall accuracy compared to the previous state-of-the-art. In addition, it is important to notice that the large increase observed by Full-BAPose utilizes a shared and unique backbone to detect all 133 keypoints in contrast to previous work that deploy different backbones for different tasks (face, body, hands, feet) on the COCO-WholeBody dataset.

Table 5. Full-BAPose results and comparison with SOTA methods for the COCO-WholeBody dataset for validation. Best results are in bold.

Method	Backbone	Approach	Single Task	Whole Body AP	Body AP	Foot AP	Face AP	Hand AP
Full-BAPose	HRNet-W48	Top-Down	N	**68.4%**	**74.4%**	76.4%	**86.8%**	**64.6%**
OmniPose [9]	HRNet-W48	Top-Down	N	65.8%	73.8%	66.6%	86.8%	60.2%
Zauss et al. [45]	ShuffleNetV2k168	Bottom-Up	N	60.4%	69.6%	63.4%	85.0%	52.9%
ZoomNet [44]	2 × HRNet (W32 + W18)	Top-Down	Y	54.1%	74.3%	**79.8%**	62.3%	40.1%
HRNet [4]	HRNet-W32	Top-Down	N	43.2%	65.9%	31.4%	52.3%	30.0%
HPRNet [46]	HG	Bottom-Up	N	34.8%	59.4%	53.0%	75.4%	50.4%
OpenPose [5]	-	Bottom-Up	N	33.8%	56.3%	53.2%	48.2%	19.8%
AE [28]	HG	Bottom-Up	N	27.4%	40.5%	7.7%	47.7%	34.1%

Figure 6 presents sample pose estimation for the COCO-WholeBody dataset, exemplifying the high accuracy of Full-BAPose for the complete human pose estimation task.

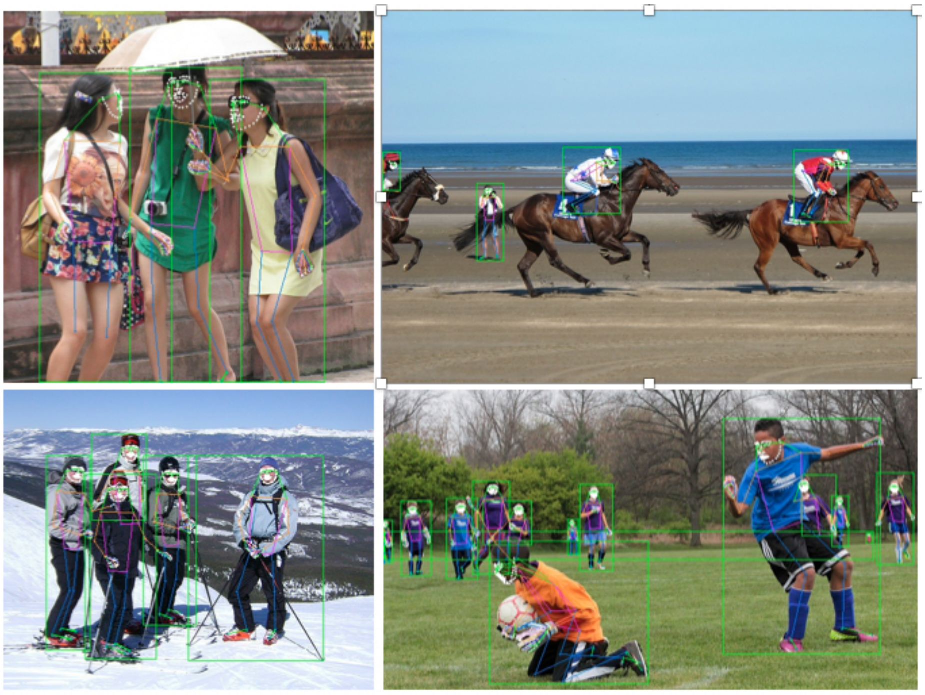

Figure 6. Pose estimation results using Full-BAPose with the COCO-WholeBody dataset.

7. Conclusions

We presented the Full-BAPose framework for bottom-up multi-person, full-body pose estimation. The Full-BAPose method addressed the broader task of full-body pose estimation including hands, feet, and facial landmarks. The Full-BAPose network includes the D-WASP module that combines multi-scale features obtained from the waterfall flow with the person-detection capability of disentangled adaptive regression. The results demonstrate SOTA performance on body pose for both the COCO and CrowdPose datasets in all metrics, as well as superior capability of person detection and pose estimation in densely populated images. Future work will extend our framework with vision transformers for improved performance.

Author Contributions: Conceptualization, B.A. and A.S.; methodology, B.A.; algorithm and experiments, B.A. and A.S.; original draft preparation, B.A. and A.S.; review and editing, B.A. and A.S.; supervision, A.S.; project administration, A.S.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Science Foundation grant number 1749376.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average Precision
AR	Average Recall
ASPP	Atrous Spacial Pyramid Pooling
CFA	Cascade Feature Aggregation
CNN	Convolutional Neural Networks
COCO	Common Objects in Context
CPM	Convolutional Pose Machines
CRF	Conditional Random Fields
DARK	Distribution-Aware coordinate Representation of Keypoints
DEKR	Disentangled Keypoint Regression

FCN	Fully Convolutional Networks
FOV	Field-of-View
GAN	Generative Adversarial Networks
HG	Hourglass
HGG	Hierarchical Graphical Clustering
HRNet	High-Resolution Network
IOU	Intersection Over the Union
MSPN	Multi-Stage Pose Network
OKS	Object Keypoint Similarity
PAF	Part Affinity Fields
SOTA	State-of-the-art
WASP	Waterfall Atrous Spatial Pooling

References

1. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 483–499.
2. Wei, S.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
3. Artacho, B.; Savakis, A. UniPose: Unified Human Pose Estimation in Single Images and Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
4. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
5. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
6. Geng, Z.; Sun, K.; Bin Xiao, Z.Z.; Wang, J. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
7. Artacho, B.; Savakis, A. BAPose: Bottom-Up Pose Estimation with Disentangled Waterfall Representations. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 528–537.
8. Artacho, B.; Savakis, A. UniPose+: A unified framework for 2D and 3D human pose estimation in images and videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 9641–9653. [CrossRef] [PubMed]
9. Artacho, B.; Savakis, A. OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation. *arXiv* **2021**, arXiv:2103.10180.
10. Rogez, G.; Weinzaepfel, P.; Schmid, C. LCR-Net: Localization-Classification-Regression for Human Pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
11. Chen, H.; Feng, R.; Wu, S.; Xu, H.; Zhou, F.; Liu, Z. 2D Human pose estimation: A survey. In *Multimedia Systems*; Springer: Cham, Switzerland, 2022; pp. 1432–1882. [CrossRef]
12. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1840.
13. Artacho, B.; Savakis, A. Waterfall Atrous Spatial Pooling Architecture for Efficient Semantic Segmentation. *Sensors* **2019**, *19*, 5361. [CrossRef] [PubMed]
14. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
15. Li, W.; Wang, Z.; Yin, B.; Peng, Q.; Du, Y.; Xiao, T.; Yu, G.; Lu, H.; Wei, Y.; Sun, J. Rethinking on multi-stage networks for human pose estimation. *arXiv* **2019**, arXiv:1901.00148.
16. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-Aware Coordinate Representation for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
17. Zhang, H.; Ouyang, H.; Liu, S.; Qi, X.; Shen, X.; Yang, R.; Jia, J. Human pose estimation with spatial contextual information. *arXiv* **2019**, arXiv:1901.01760.
18. Su, Z.; Ye, M.; Zhang, G.; Dai, L.; Sheng, J. Cascade Feature Aggregation for Human Pose Estimation. *arXiv* **2019**, arXiv:1902.07837.
19. Cao, Z.; Wang, R.; Wang, X.; Liu, Z.; Zhu, X. Improving Human Pose Estimation With Self-Attention Generative Adversarial Networks. In Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 8–12 July 2019; IEEE: New York, NY, USA, 2019; pp. 567–572.
20. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2019; pp. 11313–11322.
21. Xu, L.; Jin, S.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P.; Wang, X. ZoomNAS: Searching for Whole-Body Human Pose Estimation in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 5296–5313. [CrossRef] [PubMed]
22. Ning, G.; Huang, H. LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking. *arXiv* **2019**, arXiv:1905.02822.

23. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Girshick, R.; Radosavovic, I.; Gkioxari, G.; Dollár, P.; He, K. Detectron. 2018. Available online: https://github.com/facebookresearch/Detectron (accessed on 2 February 2023).
25. Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.V.; Schiele, B. Deepcut: Joint subset partition and labeling for multi person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4929–4937.
26. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 34–50.
27. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite Fields for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
28. Newell, A.; Huang, Z.; Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv* **2016**, arXiv:1611.05424.
29. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. 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), Munich, Germany, 8–14 September 2018; pp. 269–286.
30. Jin, S.; Liu, W.; Xie, E.; Wang, W.; Qian, C.; Ouyang, W.; Luo, P. Differentiable Hierarchical Graph Grouping for Multi-person Pose Estimation. *arXiv* **2019**, arXiv:2007.11864.
31. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-stage multi-person pose machines. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6951–6960.
32. Nie, X.; Feng, J.; Xing, J.; Yan, S. Pose partition networks for multi-person pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.
33. Varamesh, A.; Tuytelaars, T. Mixture dense regression for object detection and human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13086–13095.
34. Long, J.; Shelhamer, E.; Darrel, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
35. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution and Fully Connected CFRs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–845. [CrossRef] [PubMed]
36. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
37. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2017–2025.
38. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014. Available online: http://xxx.lanl.gov/abs/1405.0312 (accessed on 10 March 2023).
39. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.; Lu, C. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
40. Khirodkar, R.; Chari, V.; Agrawal, A.; Tyagi, A. Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2019; pp. 3122–3131.
41. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
42. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
43. Wei, F.; Sun, X.; Li, H.; Wang, J.; Lin, S. Point-set anchors for object detection, instance segmentation and pose estimation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 527–544.
44. Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P. Whole-Body Human Pose Estimation in the Wild. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
45. Zauss, D.; Kreiss, S.; Alahi, A. Keypoint Communities. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 11057–11066.
46. Samet, N.; Akbas, E. HPRNet: Hierarchical point regression for whole-body human pose estimation. *Image Vis. Comput.* **2021**, *115*, 104285. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.