applied sciences

MDPI

Article

A New Multi-Person Pose Estimation Method Using the Partitioned CenterPose Network

Jiahua Wu and Hyo-Jong Lee *

Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, Korea; salmon2wu@gmail.com

* Correspondence: hlee@jbnu.ac.kr

Abstract: In bottom-up multi-person pose estimation, grouping joint candidates into the appropriately structured corresponding instance of a person is challenging. In this paper, a new bottom-up method, the Partitioned CenterPose (PCP) Network, is proposed to better cluster the detected joints. To achieve this goal, we propose a novel approach called Partition Pose Representation (PPR) which integrates the instance of a person and its body joints based on joint offset. PPR leverages information about the center of the human body and the offsets between that center point and the positions of the body's joints to encode human poses accurately. To enhance the relationships between body joints, we divide the human body into five parts, and then, we generate a sub-PPR for each part. Based on this PPR, the PCP Network can detect people and their body joints simultaneously, then group all body joints according to joint offset. Moreover, an improved l_1 loss is designed to more accurately measure joint offset. Using the COCO keypoints and CrowdPose datasets for testing, it was found that the performance of the proposed method is on par with that of existing state-of-the-art bottom-up methods in terms of accuracy and speed.

Keywords: multi-person pose estimation; partitioned centerpose network; partition pose representation

Citation: Wu, J.; Lee, H.J. A New Multi-Person Pose Estimation Method Using the Partitioned CenterPose Network. *Appl. Sci.* **2021**, *11*, 4241. https://doi.org/10.3390/app11094241

Academic Editors: Antonio Fernández-Caballero

Received: 8 March 2021
Accepted: 5 May 2021
Published: 7 May 2021

1. Introduction

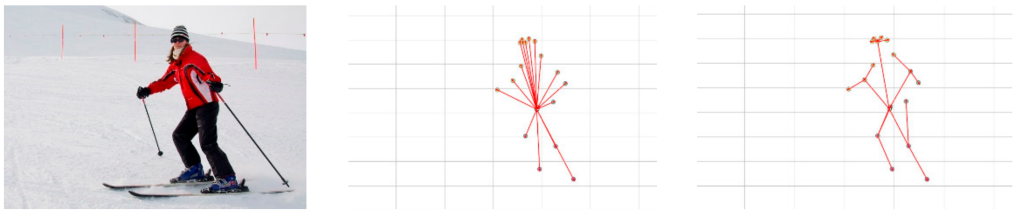
Driven by extensive research efforts, significant progress has been made in human pose estimation. The goal of human pose estimation is to obtain the posture of a human body from monocular images or videos. Pose estimation is a fundamental computer vision task providing vital information for many applications such as action detection and recognition [1], human tracking [2], and medical assistance among others [3].

With the rapid progress in deep learning technology, human pose estimation performance has improved greatly over recent years. However, finding a balance between efficiency and accuracy remains challenging. Multi-person pose estimation methods are generally classified based on their starting point for prediction as either top-down or bottom-up [4]. Top-down methods [5–11] first identify and localize instances of people using an existing person detector system and then conduct pose estimation for each person individually. Generally, top-down methods are effective since these methods profit from advances in person detectors. However, the computational cost of such methods linearly increases with the number of people in an image because single-person pose estimation must be carried out repeatably, in sequence, for each person in the image, as such, such methods are usually too slow to achieve real-time detection.

In contrast, bottom-up strategies [12–16] first identify all the body joints in the entire image, then these joints are grouped into corresponding instances of people. Unlike top-down methods, bottom-up methods avoid higher joint detection and are more robust as the number of people in an image increase. In many cases, performance when clustering the joint candidates determines the final accuracy of detection. Cao et al. [12] proposed the use of Part Affinity Fields (PAFs) to encode the coordinates and angles of limbs to

assist in grouping joints into different people; this approach ignores the relationship between each body joint and instance of a person. Newell et al. [13] constructed associative embedding maps to tag each joint on the corresponding person pose. This method adds a link between each body joint and the corresponding instance of a person, however, it neglects information relevant to adjacent body joints. Consequently, it is difficult to simultaneously maintain relationships between different joints in a single limb and link each joint from the corresponding instance of a person.

To overcome this issue, we first propose a novel pose representation technique, termed Partition Pose Representation (PPR), which combines the position information from instances of people and their body joints. Inspired by [17], we first represent each instance of a person with a single point at the center of their bounding box. Then, the positions of body joints are encoded by their offset from the center point, as shown in Figure 1b. In this way, the relationship between adjacent body joints is severed. To maintain some correlation between adjacent body joints, we further divide the human body into five parts: the head, left arm, right arm, left leg, and right leg, we then extend PPR to sub-PPR for each part. The respective center points of each part are the nose, left elbow, right elbow, left knee, and right knee. With the addition of sub-PPR, human poses generate stable connections with their instance of a person, as shown in Figure 1c.



(**a**) (**b**) (**c**)

Figure 1. Different pose representations captured from image (**a**). (**b**) Traditional pose representation, in which each joint is represented by absolute coordinates. (**c**) Proposed partition pose representation.

To exploit the advantages of PPR, we introduce a new bottom-up model, the Partitioned CenterPose (PCP) Network, to identify the poses of multiple people. The PCP Network can simultaneously locate the position of an instance of a person and identify all joint candidates. Meanwhile, a parallel prediction branch in the PCP Network, called the offset prediction head, builds an associative embedding map to predict the offset for each body center. Here we introduce an improved l_1 loss to obtain more accurate joint offset values. Supported by PPR, the joint candidates can be assigned to the corresponding body center using the offset as a guide.

Experiments on the MS COCO and CrowdPose datasets demonstrate the efficiency and effectiveness of the proposed method. It achieves competitive performance and superior speed versus state-of-the-art methods. Our work makes three main contributions.

(1) We propose a novel partition pose representation method to construct a relationship between body joints and the body center, while preserving correlations between adjacent body joints.
(2) We propose a new bottom-up model with an improved l_1 loss to efficiently and robustly predict and partition body joints to multiple people.
(3) In experiments, our PCP Network is competitive with state-of-the-art methods using the MS COCO and CrowdPose datasets while achieving a higher inference speed.

2. Related Work

2.1. Multi-Person Pose Estimation

Multi-person pose estimation is a comprehensive task that combines the challenges of person detection and keypoint estimation. With the incredible advancements over recent years in object detection and single-person pose estimation methods [4,5,8,18–24],

the performance of multi-person pose estimation has also improved, getting good results even on some complex datasets. Based on how calculations for a particular method are started, multi-person pose estimation methods are often divided into top-down methods and bottom-up methods.

Top-down methods. Top-down approaches typically first use an object detector to obtain an instance of a person and then independently estimate the pose for each person identified. G-RMI [6] produces a heatmap and offset map for each joint before combining this information using an aggregation procedure. RMPE [11] introduced using a parametric pose NMS for refining pose candidates. He et al. [5] proposed an extension of the Mask R-CNN framework that synchronously predicts keypoints and human masks. In these top-down methods, predicting keypoint heatmaps is made easier by restricting the search to the detected person's bounding box. However, the top-down strategy incurs extra computational costs while initially detecting each person's bounding box.

Bottom-up methods. Bottom-up approaches first detect body joints and then assign these joints to individuals. With the increasing demands to carry out image processing tasks on mobile devices, finding appropriate lightweight methods has become a new research hotspot. Motivated by bottom-up approaches being faster and more capable of achieving real-time estimation, our approach is based on previous bottom-up approaches and aims to obtain better performance while maintaining high computational efficiency.

Existing bottom-up methods mainly focus on how to associate detected keypoints with the corresponding instance of a person. The PersonLab approach [14] introduced a greedy decoding scheme together with Hough voting to determine grouping. CMU-Pose [12] proposed Part Affinity Fields (PAFs) to encode the location and orientation of limbs, this work was further developed in the PifPaf technique [15]. However, the computational efficiency of these two-stage methods is limited by the quality of the greedy algorithm. Newell et al. [13] propose a one-stage method to detect joints and group them in one pipeline. Based on this one-stage strategy and HRNet [8], Cheng et al. [16] presented a Scale-Aware High-Resolution Network (HigherHRNet) to solve the scale variation challenge. However, existing research only focuses on the features of joints (like in PifPaf), or only uses the connection between joints and an instance of a person to cluster (like in AssocEmbedding and HigherHRNet). The novelty of our method is to use Partition Pose Representation (PPR) to combine position information from instances of a person with structure information about body joints. In PPR, we utilize tailored semantic information and information on the offset of joints from the body center to replace information from tags in associative embedding maps. Moreover, we divide the human body into five parts, define the pivot joint in these parts as the part's center. Assisted by these part centers, the relationships between different joints in a single limb become enhanced by the offset of the body joint to the part center.

2.2. Backbone Network

The backbone networks of multi-person pose estimation methods are designed to extract keypoint features and instances of people; the accuracy with which they do so largely determines the quality of the prediction results. To ensure the effectiveness of the proposed method, three different backbones architectures, Hourglass [4], Deep Layer Aggregation (DLA) [25], and HRNet [8], are comprehensively considered.

Hourglass: The stacked Hourglass Network [4] consists of overlapping residual blocks [26], each of which is linked by a skip connection to effectively process and consolidate multi-scale features. With an encoder–decoder architecture and an intermediate supervision process, the Hourglass network shows robust performance in some complex environments, such as in cases with occlusion or cases where similar parts from nearby people are present [27]. The size of this network is quite large, which results in graceful keypoint estimation performance. The structure of an hourglass module is illustrated in Figure 2a.

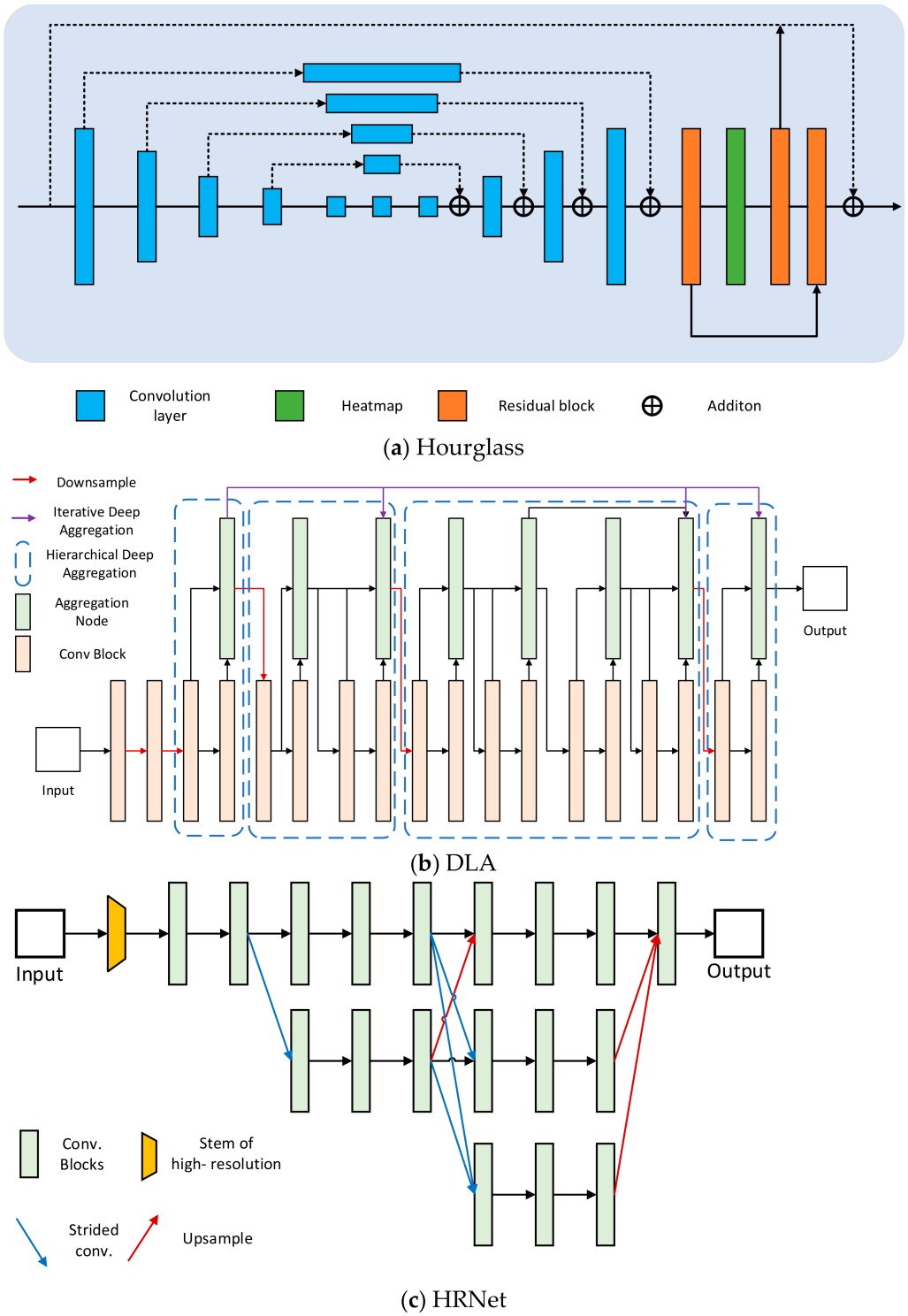


Figure 2. Structures of three state-of-the-art backbone networks for human pose estimation.

DLA: DLA [25] is an image classification network with hierarchical skip connections, in which aggregation is defined as the combination of different layers throughout a network. DLA uses iterative deep aggregation to symmetrically increase feature map resolution, preventing loss of information in dense predictions. Moreover, DLA hierarchically merges features to create networks with better accuracy and fewer parameters. The structure of a DLA network is illustrated in Figure 2b.

HRNet: HRNet [8] aims to maintain high-resolution features throughout the entire network. This network can be divided into parallel multi-resolution convolutions and

repeated multi-resolution fusions. High- to low-resolution convolution streams generate multi-scale feature maps in parallel. The goal of the fusion module is to merge information across multi-resolution representations. The structure of HRNet with three parallel branches is illustrated in Figure 2c.

3. Partition Pose Representation

In this section, we describe the proposed PPR in detail. Unlike traditional grouping methods, PPR is committed to generating connections between each body joint and instance of a person while simultaneously strengthening the correlations between different body joints. Let $I \in R^{W \times H \times 3}$ denote an input image of width W and height H and $p^k = \left\{p_1^k, p_2^k, \ldots, p_N^k\right\}$ denote N joint candidates from the kth persons in I. $\left(x_n^k, y_n^k\right)$ is the spatial coordinate of p^k, and $\left(x_{lt}^k, y_{lt}^k, x_{rb}^k, y_{rb}^k\right)$ is the bounding box of the kth instance of a person. Inspired by CenterNet [17], the body center is denoted by $\left(\hat{x}_0^k, \hat{y}_0^k\right) = \left(x_{lt}^k + x_{rb}^k, y_{lt}^k + y_{rb}^k\right)/2$.

PPR aims to aggregate the instance of a person and body pose with an offset to the body center. So, the coordinates of the nth joint of person k can be defined as:

$$\left(x_n^k, y_n^k\right) = \left(\hat{x}_0^k + \delta x_n^k, \hat{y}_0^k + \delta y_n^k\right) \tag{1}$$

where $\left(\delta x_n^k, \delta y_n^k\right)$ is the offset of the nth joint to the body center.

However, Equation (1) only considers unification of an instance of a person and body pose; it ignores the relationship between adjacent joints. Using additional information from correlated joints, the offset vector can be more accurately mapped to the position of the pose by the prediction model. Naturally, PPR divides the human body into five parts: (1) head, including nose, left eye, right eye, left ear, and right ear; (2) left arm, including left shoulder, left elbow, and left wrist; (3) right arm, including right shoulder, right elbow, and right wrist; (4) left leg, including left hip, left knee, and left ankle; and (5) right leg, including right hip, right knee, and right ankle. Then, we use the same approach as used in Equation (1) to represent the joints in each part. Here, the center points of each part p_c^k are no longer the body center, but the nose, left elbow, right elbow, left knee, and right knee are taken as the centers of the five respective body parts. Some complex environments may mean a part center is not visible; this will affect encoding by PPR. In this situation, we calculate the center of the remaining joints in this part to replace the part center; we call this point the illusion center. Thus, the complete PPR can be formulated as:

$$\left(x_n^k, y_n^k\right) = \begin{cases} \left(\hat{x}_0^k + \delta x_n^k, \hat{y}_0^k + \delta y_n^k\right) \text{ if} p_n^k \in p_c^k \\ \left(\hat{x}_m^k + \delta \hat{x}_n^k, \hat{y}_m^k + \delta \hat{y}_n^k\right) \text{ otherwise} \end{cases} \tag{2}$$

when the part center is visible, $\left(\hat{x}_m^k, \hat{y}_m^k\right)$ is the coordinates of the center point of the mth part and $\left(\delta \hat{x}_n^k, \delta \hat{y}_n^k\right)$ is the offset of the nth joint from the corresponding part center. When the part center is not visible, $\left(\hat{x}_m^k, \hat{y}_m^k\right)$ is the coordinate of the illusion center of the mth part and $\left(\delta \hat{x}_n^k, \delta \hat{y}_n^k\right)$ is the offset of the nth joint from the corresponding illusion center.

Using the offset from the part center to the body center, PPR establishes the connection between a body pose and the instance of a person. At the same time, PPR retains global information related to the limbs and generates correlations between body joints in one part through the offset of other joints to the part center.

4. Partitioned CenterPose Network

In conjunction with PPR, we propose the box-free bottom-up PCP Network to detect body joints of multiple people. Motivated by the recent success of keypoint-based object detection approaches [17,28], we implement the PCP Network with a simple one-stage model. Below, we will describe the network architecture, training, and inference details of the PCP Network. The overall pipeline for the proposed network is shown in Figure 3.

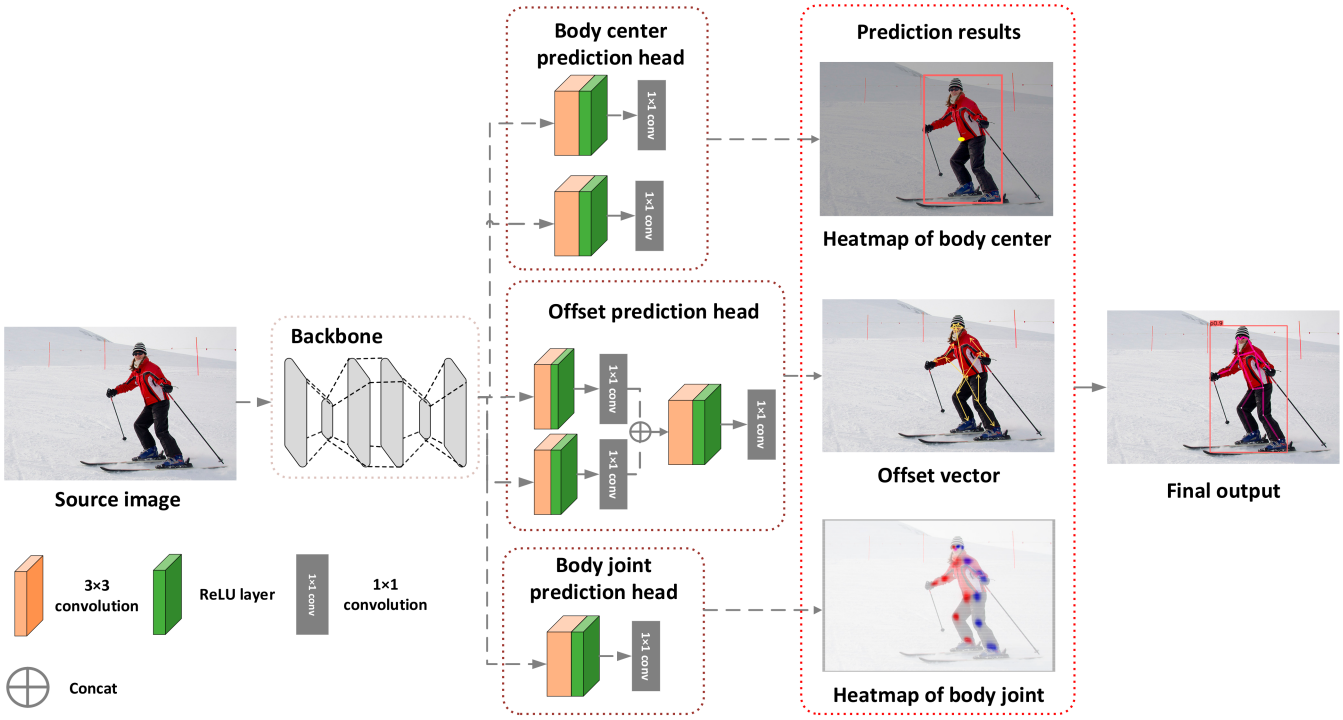


Figure 3. The architecture of the Partitioned CenterPose Network. A convolutional backbone network applies three sets of prediction heads to predict instance location, joint offset, and joint heatmap. The final output is generated by combining these three prediction results.

4.1. Network Architecture

In the PCP Network, a convolutional backbone network is applied for feature extraction. Then, we use three sets of prediction heads (body center prediction head, offset prediction head, and body joint prediction head) to process the output features. First, we will discuss the structure of the offset prediction head. In PPR, the offset vector is the key to connecting an instance of a person with their body joints; as such, it is very important to obtain an accurate offset vector. Directly regressing the value of an offset vector is inefficient as it is a highly non-linear task and difficult to learn the mapping [3]. Inspired by [13], we use two associative embedding maps to record the vector value of each offset. As shown in Figure 3, the output of the backbone is passed through two parallel branches. The output channel of the first branch is twice the number of part centers, which focus on the 2D vector value of the offset from the body center to the part centers. The second branch looks at the offset of the remaining joints to the part center. Then, we concatenate the output of these two branches and pass it through a simple convolutional module to acquire the final embedding maps. When the coordinates of the body center or part center are obtained, the feature value of the embedding map at this position can be regarded as the corresponding offset vector value. In the body center prediction head, follow the approach used by CenterNet [17], we use a simple convolutional module, which contains only a separate 3×3 convolution, ReLU, and a 1×1 convolution, to predict the body center and the bounding box using two parallel branches. The body joint prediction head

estimates a heatmap of each body joint $\left(x_n^k, y_n^k\right)$ using the same structure as used for the body center prediction head to reduce computational complexity.

4.2. Training and Inference

Training. An improved l_1 loss was designed for the PCP Network to better train the system to identify the offset between the joint and part center. As shown in Figure 4, the lengths of the offset vectors in the head part are short, but the structures of the offset vectors in different people are relatively similar. Thus, enhancing the weight of offset length in the loss function allows the network to understand small differences in head structure more accurately. Conversely, the offset vectors of the limbs of different people differ more in terms of angle while the lengths tend to be quite similar. Accordingly, based on the l_1 loss, we designed two different loss functions for the offset vector in the head and in the limbs:

$$L_{off}^{head} = \frac{1}{N}\sum_{i}^{N}\left(||\vec{O}_i - \vec{O'}_i||_1 + 0.5||\vec{O}_i - \vec{O'}_i||_2\right) \tag{3}$$

$$()()L_{off}^{limb} = \frac{1}{N}\sum_{i}^{N}\left(||\vec{O}_i - \vec{O'}_i||_1 + \left|arctan\frac{\vec{O}_i}{||\vec{O}_i||_2} - arctan\frac{\vec{O'}_i}{||\vec{O'}_i||_2}\right|\right) \tag{4}$$

where $\vec{O}$ is the predicted offset vector and $\vec{O'}$ is the corresponding ground truth. N is the number of body joints in the body part. $|\cdot|$ is the absolute value, and $||\cdot||_1$ and $||\cdot||_2$ are the l_1-norm and l_2-norm, respectively. In Section 5.4, we discuss an ablation experiment to demonstrate the effect of the improved l_1 loss.

The total loss of the improved l_1 loss is shown below:

$$L = L_{bct} + \alpha L_{bsize} + L_{off}^{pct} + L_{off}^{pj} + L_{bj} \tag{5}$$

$$L_{off}^{pct} = L_{off}^{limb} \tag{6}$$

$$L_{off}^{pj} = \left(L_{off}^{head} + 4 * L_{off}^{limb}\right)/5 \tag{7}$$

where L_{bct} and L_{bj} denote the focal losses [29], which are used to train the network to detect the body center and body joint heatmaps, respectively. The focal loss is defined as:

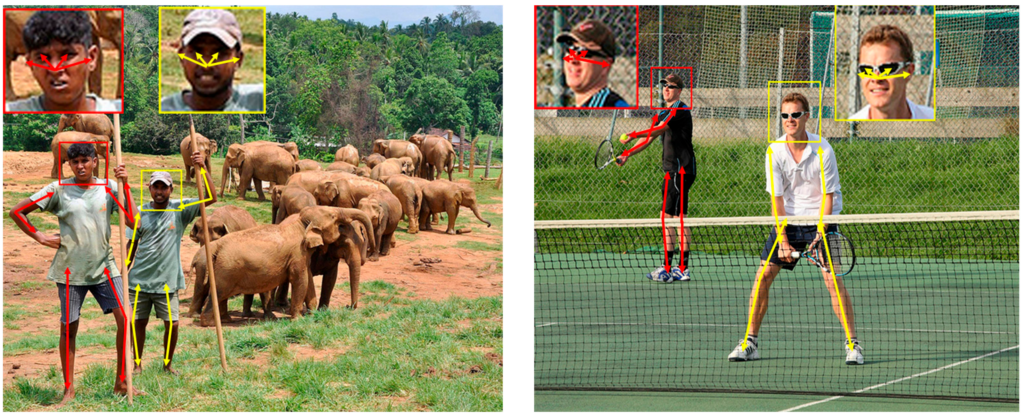
$$L_{focal} = \frac{-1}{N}\sum_{n}\begin{cases} (1-\hat{H}_p)^{\beta}\log(\hat{H}_p) & \text{if } H_p = 1 \\ (1-H_p)^{\gamma}(\hat{H}_p)^{\beta}\log(1-\hat{H}_p) & \text{otherwise} \end{cases} \tag{8}$$

where β and γ are hyper-parameters used to reduce the imbalance between an easy example and a hard example. H_p is the ground truth heatmap and $\hat{H}_p$ is the heatmap of p^k. Following [28], β is set to 2 and γ is set to 4. L_{bsize} is the l_1 loss [30] used to regress the size of the bounding box. L_{off}^{pct} is the loss function used to train the offset between the part center and body center, while L_{off}^{pj} is the loss function used to train the offset between the joint and part center. α is a constant weight parameter that is set to 0.1.

Inference. Following PPR, we group the detected keypoints by offset vector. Given a test image of width W and height H, the outputs of the PCP Network include a body center heatmap $H_{bc} \in R^{W\times H\times 1}$, bounding box maps $H_{bb} \in R^{W\times H\times 2}$, offset maps $H_{off} \in R^{W\times H\times 34}$, joint heatmaps $H_{bj} \in R^{W\times H\times 17}$. We first choose the top N_η high-confidence instances of people (100 was used in our implementation) and extract their body centers $\left(\hat{x}_0^k, \hat{y}_0^k\right)$ from H_{bc}. With the coordinates of body centers, the size of the bounding box $\left(w^k, h^k\right)$ and the offset of the part centers to the body center $\vec{O}_n^k$ can be extracted from $H_{bb}\left(\hat{x}_0^k, \hat{y}_0^k\right)$ and $H_{off}^{n,n+1}\left(\hat{x}_0^k, \hat{y}_0^k\right)$, respectively. For the kth body center, we extract the

coordinates of part center candidates $\left(x_n^k, y_n^k\right)$ from the joint heatmaps H_{bj}, where the candidates are selected from inside the bounding box of the kth body center. Then, the offset $\vec{O}_{ht}$ from the part center candidates to the body center are calculated by:

$$\vec{O}_{ht} = \left(x_n^k - \hat{x}_0^k, y_n^k - \hat{y}_0^k\right) \tag{9}$$



(**a**) (**b**)

Figure 4. PPR of the head (magnified area) and PPR of the limbs in different people. (**a**,**b**) are two samples from COCO dataset [31]. The lengths of the offset vectors in the head part are much shorter than those in the limb parts.

Next, each part center candidate is assigned by argmin $i \in N_c \left(\vec{O}_{ht} - \vec{O}_n^k\right)$ to identify the closest predicted offset vector $\vec{O}_n^k$. Here, N_c is the total number of part centers. After grouping the part centers, we can extract the offset of the remaining joints to the part center $\vec{O}_m^k$ from $H_{off}^{m,m+1}\left(x_n^k, y_n^k\right)$. If the part center is not visible, $\left(x_n^k, y_n^k\right)$ will be replaced by $\left(\hat{x}_0^k, \hat{y}_0^k\right) + \vec{O}_n^k$. Using the same strategy, we can group the remaining body joints to corresponding instances of a person. Finally, the complete human skeletons of multiple people are formed using the default connections between the predicted body joints.

The network structure of the prediction heads is simple and lightweight, the body centers are obtained directly from keypoint estimation without the need for IoU-based non-maxima suppression or other greedy algorithms. In the inference post-processing, due to the constraints of the bounding box, the number of joint candidates can be reduced greatly to only in the candidates in small areas of the image, this not only improves accuracy it also reduces computing time. Therefore, in our method, post-processing does not take too long while the computational efficiency is similar to one-stage methods.

5. Experiments

5.1. Dataset

The experiments were performed using the MS-COCO dataset [31]. This dataset contains more than 250,000 instances of people with 17 body joints, the dataset is divided into *train*, *val* and *test-dev* sets with 57 k, 5 k, and 20 k images, respectively. We use the *train* set for training and test the results on the *test-dev* set. The *val* set is used to perform ablation studies and visualization experiments.

The MS-COCO dataset uses Object Keypoint Similarity (OKS)-based AP (average precision) and AR (average recall) metrics to evaluate the performance of a detector. OKS is inspired by the IoU index in object detection, this calculates the distance between predicted

body joints and the ground truth, normalized to the scale of the person [32]. OKS can be defined as:

$$OKS_p = \frac{\sum_i exp\left\{-d_{pi}^2/2S_p^2\sigma_i^2\right\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)} \quad (10)$$

where p denotes the pth person in an image and i is the ith keypoint of this person. d_{pi} is the Euclidian distance between the ground truth keypoint and predicted keypoint. S_p is the scale factor of the person, which is equal to the square root of the object segment area. σ_i is the normalization factor of the ith keypoint, which reflects the difficulty of labeling this keypoint. $v_{pi} = 1$ indicates that the ith keypoint of the pth person is visible.

In this section, we mainly use AP (mean AP score in OKS $= 0.5,\ 0.55, \ldots, 0.90,\ 0.95$), $AP^{0.5}$, $AP^{0.75}$, AP^M, AP^L, and AR as metrics, where 0.5 and 0.75 are the threshold values for OKS, M and L represent medium objects (32^2 < area < 96^2) and large objects (area > 96^2), respectively [33,34].

5.2. Experimental Setup

We experimented on using four backbones in our method: DLA-34 [25], ResNet-101 [26], Hourglass-104 [4], and HRNet-w32 [8]. All these models were written using PyTorch software [35]. The resolution of the input image was 512 × 512, leading to heatmaps with a size of 128 × 128. The ground-truth heatmap was constructed by applying a Gaussian kernel with the same parameters as used in [36] to filter all body joints. Each sample was augmented by rotating, scaling, and flipping. We utilized Adam [37] as the optimizer and trained the PCP Network on a RTX2080ti GPU. For the DLA-34 backbone, we trained with a batch size of 48 and a learning rate of 3×10^{-4} for 300 epochs; the learning rate was decreased by 0.1 in epochs 250 and 280. For the ResNet-101 backbone, we trained with a batch size of 48 and a learning rate of 1×10^{-3} for 300 epochs; the learning rate was decreased by 0.1 in epochs 250 and 280. For the Hourglass-104 backbone, we trained with a batch size of 24 and a learning rate of 2.5×10^{-4} for 150 epochs; the learning rate was decreased by 0.1 in epochs 110 and 130. For the HRNet-w32 backbone, we trained with a batch size of 32 and a learning rate of 2×10^{-4} for 320 epochs; the learning rate decreased by 0.1 in epochs 270 and 300.

5.3. Experimental Results

To assess the performance of our PCP Network, we compared the results of our method with those of six current mainstream bottom-up pose estimation methods, including CMU-Pose [12], Mask-RCNN [5], G-RMI [6], AssocEmbedding [13], PifPaf [15], PersonLab [14], and HigherHRNet [16]. Table 1 summarizes the experimental results on the test-dev dataset. The differences between HigherHRNet-1 and HigherHRNet-2 are the backbone and input size. As shown in Table 1, our method is slightly inferior to PersonLab and HigherHRNet-2, which both use a more powerful backbone and larger training images. However, when using the same backbone and same input size, the performance of our method is better than Mask-RCNN, G-RMI, AssocEmbedding, PifPaf, and HigherHRNet-1. In addition to performance, we also consider the inference time of each method.

As shown in Table 1, the speed of our PCP Network is outstanding, especially when DLA is used as the backbone. Even with the HRNet backbone, the inference speed of our PCP Network was 5× faster than that of PersonLab. These results verify that our method has superior efficiency due to its excellent inference speed while maintaining very competitive performance for multi-person pose estimation tasks.

Table 1. Comparisons of our model to other state-of-the-art models on the MSCOCO test-dev daTable 2080. ti GPU. Superscripts M, L of AP stand for medium and large objects. The highest values are indicated in bold.

Method	Backbone	Input Size	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{M}	AP^{L}	AR	Time [s]
CMU-Pose [12]	-	-	0.618	0.849	0.675	0.571	0.682	0.665	0.5
Mask-RCNN [5]	ResNet-101	-	0.631	0.873	0.687	0.578	0.714	-	0.2
G-RMI [6]	ResNet-101	353	0.649	0.855	0.713	0.623	0.700	0.697	-
AssocEmbedding [13]	Hourglass	512	0.655	0.868	0.723	0.606	0.726	0.710	0.19
PifPaf [15]	-	-	0.667	-	-	0.624	0.729	-	-
PersonLab [14]	ResNet-152	1401	0.687	0.890	0.754	0.641	0.755	**0.754**	0.381
HigherHRNet-1 [16]	HRNet-W32	512	0.664	0.875	0.728	0.612	0.742	-	0.052
HigherHRNet-2 [16]	HRNet-W48	640	**0.705**	**0.893**	**0.772**	**0.666**	**0.758**	0.749	0.142
Ours (DLA)	DLA-34	512	0.634	0.864	0.693	0.575	0.739	0.698	**0.039**
Ours (ResNet)	ResNet-101	512	0.651	0.868	0.703	0.642	0.737	0.721	0.073
Ours (Hourglass)	Hourglass	512	0.663	0.881	0.731	0.662	0.747	0.748	0.132
Ours (HRNet)	HRNet-W32	512	0.668	0.883	0.740	0.665	0.748	0.751	0.078

To further prove that the performance of the proposed method is satisfactory, we also show some results from the proposed method that show intuitively that our approach is able to identify joints on a human skeleton accurately. Figure 5 shows qualitative examples from the MSCOCO dataset, including the intermediate body joint heatmaps and final predicted human poses. It is clear that our method performs well even on scenes with some challenging attributes such as sub-optimal scale, appearance variation, occlusion, or crowding.

Figure 5. Qualitative results on the MSCOCO dataset. For each pair, we show the predicted human pose (**left**) and intermediate heatmap (**right**). In the predicted human pose, each color corresponds to a particular human instance.

5.4. Ablation Analysis

We perform several ablation experiments on the COCO *val* set to better understand the gain of the proposed PPR and improved l_1 loss. Here, HRNet is used as the backbone of our network.

First, to demonstrate the effect of the proposed PPR, we trained the PCP Network with traditional pose representations (Figure 1b). Here, the body center prediction head was removed. As shown in Table 2, this network achieved an AP of 0.648. Using the proposed PPR, our PCP Network outperformed the above network by +0.12 AP (AP = 0.660). Table 3 shows the performance results from using the original l_1 loss and the improved l_1 loss. When the improved l_1 loss was used, the performance of our model increased from AP = 0.657 to 0.660. These results verify the effectiveness of the proposed PPR and improved l_1 loss. Table 3 also shows that the increase in AP for poses of large people is significantly higher than for other methods. This indicates that the improved l_1 loss works better on instances of large people.

Table 2. Ablation study results: traditional pose representation (TPR) vs. proposed PPR on the COCO2017 val dataset. Superscripts M, L of AP stand for medium and large objects. The highest values are indicated in bold.

Method	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{M}	AP^{L}	AR
PCP Network (TPR)	0.648	0.854	0.715	0.603	0.700	0.697
PCP Network (PPR)	**0.660**	**0.869**	**0.725**	**0.608**	**0.742**	**0.704**

Table 3. Ablation study results: original l_1 loss vs. improved l_1 loss on the COCO2017 val dataset. Superscripts M, L of AP stand for medium and large objects. The highest values are indicated in bold.

Method	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{M}	AP^{L}	AR
PCP Network (original loss)	0.657	0.867	0.722	0.607	0.728	0.701
PCP Network (improved loss)	**0.660**	**0.869**	**0.725**	**0.608**	**0.742**	**0.704**

5.5. CrowdPose

We demonstrated the proposed method has a state-of-the-art human pose estimation performance on the CrowdPose [38] dataset, which contained crowd scenes to make it more challenging. The training, validation, and testing subset contained 10K, 2K, and 8K images, respectively. The CrowdPose dataset also used the AP from the COCO dataset as an evaluation metric and split it into three crowding levels: easy, medium, hard. In this section, for metrics, we mainly use AP, $AP^{0.5}$, $AP^{0.75}$, AP^{E} (for easy images), AP^{M} (for medium images), and AP^{H} (for hard images). We trained the models on the training and validation subsets and reported the results achieved on the testing subset. The experimental setup follows that of COCO exactly.

The experimental results are shown in Table 4. Our method outperforms traditional top-down methods (Mask-RCNN and AlphaPose) and bottom-up method (CMU-Pose) by a large margin in terms of AP. SPPE is an efficient crowded scene pose estimation method which is a global refinement of AlphaPose; the performance of our method is comparable to AlphaPose without additional optimization. Multi-scale testing can improve the precision of predictions for small people, especially in crowd scenes. After multi-scale testing, HigherHRNet achieves the best performance on the CrowdPose dataset. While, without the optimization of multi-scale testing, the performance of our method is on par with HigherHRNet even the latter significant advantages in terms of the backbone used and the input size. The experimental results in Table 4 show the great potential of our method in complex environments and challenging scenes.

Table 4. Comparisons of our model to other state-of-the-art models on the CrowdPose test dataset. Superscripts E, M, H of AP stand for easy, medium and hard. * indicates multi-scale testing. The highest values are indicated in bold.

Method	Backbone	Input Size	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{E}	AP^{M}	AP^{H}
Top-down methods								
Mask-RCNN [5]	ResNet-101	-	0.572	0.835	0.603	0.694	0.579	0.458
AlphaPose [11]	-	-	0.610	0.813	0.660	0.712	0.614	0.511
SPPE [38]	ResNet-101	-	0.660	0.842	0.715	0.755	0.663	0.574
Bottom-up methods								
CMU-Pose [12]	-	-	-	-	-	0.627	0.487	0.323
HigherHRNet [16]	HRNet-W48	640	0.659	0.864	0.706	0.733	0.665	0.579
HigherHRNet * [16]	HRNet-W48	640	**0.676**	**0.874**	**0.726**	**0.758**	**0.681**	**0.589**
Ours (HRNet)	HRNet-W32	512	0.657	0.855	0.705	0.742	0.668	0.574

6. Conclusions

In this paper, we proposed a new bottom-up multi-person pose estimation method which strikes a balance between efficiency and accuracy. The grouping of candidate joints into a corresponding pose in a limited amount of time is the main challenge in bottom-up multi-person pose estimation. To solve this problem, we first introduced Partition Pose Representation (PPR) for multi-person pose estimation. PPR builds relationships between each joint and the corresponding instance of a person using the offset between the joint and the body center. Moreover, PPR further divides the human body into five constituent parts and utilizes another offset to the center of these parts to rebuild relationships between adjacent joints. With PPR, it is possible to group candidate joints simply and quickly without the need for any additional complex algorithms.

To leverage the advantages of PPR, we proposed the Partitioned CenterPose (PCP) Network to estimate instances of people and their body joints, PCP then groups all body joints by joint offset. By considering the different characteristics of the offsets of joints on different parts of the human body, we proposed an improved l_1 loss to enhance the accuracy of the predicted joint offsets. Extensive experiments and subjective evaluation of predictions on the COCO and CrowdPose datasets demonstrate that our method performs well both in terms of efficiency and prediction accuracy. A future study that extends PPR to 3D human pose estimation is planned. Considering the complexity of human poses in 3D space, we must reconsider how we define the center of the human body and design different loss functions to obtain more accurate offsets.

Author Contributions: Conception and design of the proposed method: H.J.L. and J.W.; performance of the experiments: J.W.; writing of the paper: J.W.; paper review and editing: H.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the NRF of Korea funded by the Ministry of Education (GR 2019R1D1A3A03103736). This research was also conducted by the 'Leaders in Industry-university Cooperation +', supported by the Ministry of Education and National Research Foundation of Korea

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

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

References

1. Dong, J.; Gao, Y.; Lee, H.J.; Zhou, H.; Yao, Y.; Fang, Z.; Huang, B. Action Recognition Based on the Fusion of Graph Convolutional Networks with High Order Features. *Appl. Sci.* **2020**, *10*, 1482. [CrossRef]
2. Insafutdinov, E.; Andriluka, M.; Pishchulin, L.; Tang, S.; Levinkov, E.; Andres, B.; Schiele, B. Arttrack: Articulated mul-ti-person tracking in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2017; pp. 6457–6465.
3. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
4. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
5. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
6. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards Accurate Multi-Person Pose Estimation in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3711–3719.
7. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision, GASTEIG Cultural Center, Munich, Germany, 10–13 September 2018; pp. 472–487.
8. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696.
9. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-person Pose Estimation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7103–7112.
10. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral Human Pose Regression. In Proceedings of the European Conference on Computer Vision, GASTEIG Cultural Center, Munich, Germany, 10–13 September 2018; pp. 536–553.
11. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2353–2362.
12. Cao, Z.; Šimon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
13. Newell, A.; Huang, Z.; Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, USA, 4–9 December 2017; pp. 2274–2284.
14. 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, GASTEIG Cultural Center, Munich, Germany, 10–13 September 2018; pp. 282–299.
15. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite Fields for Human Pose Estimation. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11969–11978.
16. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bot-tom-up human pose estimation. In Proceedings of the International Conference on Computer Vision and Pattern Recogni-tion (CVPR), Seattle, WA, USA, 16–28 June 2020; pp. 5386–5395.
17. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
18. Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; Bregler, C. Efficient object localization using Convolutional Networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 648–656.
19. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
20. Bulat, A.; Tzimiropoulos, G. Human Pose Estimation via Convolutional Part Heatmap Regression. In Proceedings of the Haptics: Science, Technology, Applications, London, UK, 4–7 July 2016; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9911, pp. 717–732.
21. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context Attention for Human Pose Estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5669–5678.
22. Lifshitz, I.; Fetaya, E.; Ullman, S. Human Pose Estimation Using Deep Consensus Voting. In *European Conference on Computer Vision*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 9906, pp. 246–260.
23. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human Pose Estimation with Iterative Error Feedback. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4733–4742.
24. Hu, P.; Ramanan, D. Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5600–5609.

25. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2403–2412.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
27. Kim, S.-T.; Lee, H.J. Lightweight Stacked Hourglass Network for Human Pose Estimation. *Appl. Sci.* **2020**, *10*, 6497. [CrossRef]
28. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision, GASTEIG Cultural Center, Munich, Germany, 10–13 September 2018; pp. 734–750.
29. Lin, T.-Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
30. Girshick, R. Fast R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
31. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV ECCV Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 6–12 September 2014; pp. 740–755.
32. Li, Y.; Wang, X.; Liu, W.; Feng, B. Pose Anchor: A Single-stage Hand Keypoint Detection Network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1. [CrossRef]
33. Xia, H.; Zhang, T. Self-Attention Network for Human Pose Estimation. *Appl. Sci.* **2021**, *11*, 1826. [CrossRef]
34. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6568–6577.
35. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the NeurIPS Workshop, Long Beach, CA, USA, 4–9 December 2017.
36. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference Learn, Represent, (ICLR), San Diego, CA, USA, 5–8 May 2015.
38. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; Lu, C. CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, 16–20 June; pp. 10855–10864.