Article

Semantic–Structural Graph Convolutional Networks for Whole-Body Human Pose Estimation

Weiwei Li ^{1,2} , Rong Du ^{1,2} and Shudong Chen ^{1,2,*}

¹ Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China; liweiwei@ime.ac.cn (W.L.); durong@ime.ac.cn (R.D.)

² University of Chinese Academy of Sciences, Beijing 100864, China

* Correspondence: chenshudong@ime.ac.cn; Tel.: +86-010-82995548

Abstract: Existing whole-body human pose estimation methods mostly segment the parts of the body's hands and feet for specific processing, which not only splits the overall semantics of the body, but also increases the amount of calculation and the complexity of the model. To address these drawbacks, we designed a novel semantic–structural graph convolutional network (SSGCN) for whole-body human pose estimation tasks, which leverages the whole-body graph structure to analyze the semantics of the whole-body keypoints through a graph convolutional network and improves the accuracy of pose estimation. Firstly, we introduced a novel heat-map-based keypoint embedding, which encodes the position information and feature information of the keypoints of the human body. Secondly, we propose a novel semantic–structural graph convolutional network consisting of several sets of cascaded structure-based graph layers and data-dependent whole-body non-local layers. Specifically, the proposed method extracts groups of keypoints and constructs a high-level abstract body graph to process the high-level semantic information of the whole-body keypoints. The experimental results showed that our method achieved very promising results on the challenging COCO whole-body dataset.

Keywords: human pose estimation; graph convolutional networks; non-local mechanics; feature embedding

Citation: Li, W.; Du, R.; Chen, S. Semantic–Structural Graph Convolutional Networks for Whole-Body Human Pose Estimation. *Information* **2022**, *13*, 109. <https://doi.org/10.3390/info13030109>

Academic Editor: Zoran H. Peric

Received: 31 January 2022

Accepted: 21 February 2022

Published: 25 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Human pose estimation is a challenging computer vision task, which aims to locate the human body keypoints in images and videos. Different from traditional human pose estimation, whole-body pose estimation aims at localizing the keypoints of the body, face, hand, and foot simultaneously. This task is important for the development of downstream applications, such as virtual reality [1,2], augmented reality [3], human mesh recovery [4–7], and action recognition [8–12]. The difficulty of this task is that it requires the estimation of differently sized parts of the body simultaneously. For example, it is difficult to process both the larger body torso and the relatively small hands, face, and feet in the whole-body pose estimation task.

However, we observed two main problems with existing methods:

- The existing methods do not sufficiently utilize the positional and semantic information of whole-body keypoints to integrate human whole-body pose estimation;
- Previous works did not make full use of the semantic relationship among the body, hand, and face keypoints to guide pose estimation.

First, the existing methods do not sufficiently utilize the positional and semantic information of whole-body keypoints to integrate human whole-body pose estimation. The current approaches lack the exploration of the relationship between positional and semantic features based on the keypoint heat map. Coordinate position embedding and feature map fixed cropping of keypoint representations may lose contextual semantic information. To

leverage the joint point information represented by the keypoint heat map, we propose a novel 1D heat-map-based keypoint embedding method, which encodes the position information and feature information of the keypoints of the human-body-based keypoint heat map context. The keypoint embeddings are then fed into the GCN and non-local layers for the final whole-body pose estimation.

Second, previous works did not make full use of the semantic relationship among the body, hand, and face keypoints to guide pose estimation. Hands, face, and feet are the ends of the body's torso and are closely related to the semantics of the torso. For example, a running torso pose often corresponds to a specific arm swing hand gesture and face pose. The postures of the whole-body parts form an interrelated overall semantics to improve the accuracy of whole-body posture estimation. Most of these methods use separate models to detect the keypoints of the torso, hands, and face separately and then stitch them together to estimate the whole-body pose, which not only splits the overall semantics of the body, but also increases the amount of calculation and complexity of the model. As it is shown in Figure 1, by introducing a whole-body graph convolutional network, a whole-body pose semantic map is constructed to estimate the overall pose of the human body. The proposed full-body pose estimation scheme is more lightweight and has high accuracy at the same time.

Figure 1. Illustration of the proposed semantic-structural graph convolutional network. Existing whole-body pose estimation methods tend to separate parts of the body for pose estimation, while our proposed method leverages the whole-body graph structure to analyze the semantics of each part of the body through the graph convolutional network and improves the accuracy of pose estimation.

To the best of our knowledge, we are the first work to leverage the heat-map-based graph convolutional network to calibrate human whole-body human pose estimation. Our main contributions are summarized as follows:

- This work presents a novel graph convolutional network framework for whole-body human pose estimation tasks, which leverages the whole-body graph structure to analyze the semantics of each part of the body through the graph convolutional network;
- We propose a novel heat-map-based keypoint embedding module, which encodes the position information and feature information of the keypoints of the human body;
- The proposed semantic-structural graph convolutional network consists of a structure-based graph layer to capture skeleton structure information and a data-dependent non-local layer to analyze the long-range grouped joint features;
- We represent groups of keypoints and construct a high-level abstract body graph to process the high-level semantic information of the whole-body keypoints.

In this paper, we discuss recent work on whole-body pose estimation in Section 2, detail our proposed SSGCN method in Section 3, and finally, illustrate the experimental procedure and analysis of the results in Section 4.

2. Related Work

2.1. Human Pose Estimation

Human pose estimation (HPE) [13] is the estimation of the human body part configuration from input data captured by sensors, especially in images and videos. Significant progress and significant performance have been achieved by using deep learning techniques in HPE tasks.

Two-dimensional single-person pose estimation is used to localize human body joint positions when the input is a single-person image. The HPE body part detection method is designed to train the body part detector to predict the position of body joints. Some recent methods treat pose estimation as a heat map prediction problem. The ground truth heat map is generated from a 2D Gaussian distribution centered on the ground truth joint position [14,15]. Compared to joint coordinates, heat maps facilitate the training of convolutional networks by retaining spatial location information, thus providing richer supervised information. Therefore, there is a growing interest in using heat maps to represent joint locations and to develop effective CNN architectures for HPEs, such as [15–19].

Two-dimensional hand pose estimation methods typically use RGB images as the input, while 3D hand pose estimation is more challenging. Three-dimensional hand pose estimation methods can be divided into three categories: depth images, multiple RGB images, and monocular RGB images. Monocular RGB images are easier to obtain than depth images and multiple RGB images. Some early works [20,21] proposed complex model fitting methods that are based on dynamics and multiple assumptions and depend on the constrained requirements.

2.2. Whole-Body Pose Estimation

Whole-body posture estimation requires accurate localization of keypoints throughout the body, including on the body, face, hands, and feet. The detection of keypoints was studied independently for each of these body parts, including face alignment [22–25], facial landmark detection [26,27], hand pose estimation [28,29], hand tracking [30,31], and foot keypoint detection [32]. However, there is not much work related to whole-body pose estimation due to the lack of large-scale annotated datasets. OpenPose [32] attempted whole-body keypoint estimation prior to the release of the COCO WholeBody Dataset [33]. OpenPose integrates five separately trained models, namely human pose estimation, hand detection, face detection, hand pose estimation, and face pose estimation. With multiple models running separately, the training and inference of the OpenPose full-body recognition model are both complex and expensive. Our end-to-end trainable single network eliminates these drawbacks. SN [34] extended PAF [35] for whole-body pose estimation. Similar to PAF [35], they predicted heat maps for each keypoint and grouped them using partial affinity maps. The SN model is trained on a dataset that is sampled from different datasets. However, SN falls short of handling scale variations between whole-body parts and does not take advantage of the semantic relationships between keypoints.

The release of the COCO WholeBody Dataset [33] builds a benchmark for whole-body pose estimation. Jin et al. extended the original COCO keypoints [36] dataset by further annotating face, hand, and foot keypoints. They also proposed a robust, two-stage, top-down model to perform whole-body pose estimation on the COCO WholeBody Dataset. Similar to top-down human pose estimation methods, Jin et al. [33] first obtained candidate person boxes in an image using FasterRCNN [37]. Then, using a single network called ZoomNet, the detection of whole-body keypoints was performed on the person boxes. ZoomNet consists of four sub-CNNs, which use HRNet-W32 [38] for BodyNet and HRNetV2p-W18 [39] for the FaceHead and HandHead networks. Besides, HPRNet [40] presents a new bottom-up one-stage method for whole-body pose estimation. To handle the scale variance among different body parts, they built a hierarchical point representation of body parts and jointly regressed them.

These methods have multiple independent modules that can be fine-tuned separately for hands and faces, which ignore the semantic association of pose relations for each part

of the whole body. We analyzed the whole-body pose estimation task from this new perspective based on the semantic relations of the parts of the torso, hands, and face. We noticed that the semantics of the various parts of the human body are related and can form a complete skeleton graph. Based on this, we extended the graph convolutional networks to perform whole-body pose estimation by introducing the heat-map-based feature embedding of keypoints. We show that GCN-based whole-body semantic analysis can significantly improve the accuracy of pose estimation.

3. Heat-Map-Based Skeletal-Structural Graph Convolutional Network

As shown in Figure 2, we propose a novel graph convolutional network scheme to learn whole-body semantic representations. Our approach consists of two main steps. First, we propose a novel heat-map-based keypoint embedding, which encodes the position information and feature information of the keypoints of the human body. Second, the proposed semantic-structural graph convolutional network consists of a structure-based graph layer to capture skeleton structure information and a data-dependent non-local layer to analyze long-range grouped joint features.

Figure 2. Overview of the proposed SSGCN scheme. It comprises three fundamental components: feature extraction backbone, heat-map-based keypoint embedding, and the skeletal-structural graph convolutional network.

3.1. One-Dimensional Heat-Map-Based Keypoint Embedding

The proposed heat-map-based keypoint embedding encodes the position information and feature information of the keypoints of the human body, which are shown in Figure 3a,b, respectively.

3.1.1. Heat-Map-Based Keypoint Position Embedding

The heat-map-based keypoint embedding estimates x -axis and y -axis 1D heat maps of all human joints $P^H = \{P^{H,x}, P^{H,y}\}$. The backbone network extracts image feature $F_p \in \mathbb{R}^{c \times h \times w}$ from the input image. As shown in Figure 3a, we define the x -axis and y -axis 1D keypoint heat map as follows:

$$P^{H,x} = \text{avg}^y(f_p^{kp}(F_p)) \quad (1)$$

$$P^{H,y} = \text{avg}^x(f_p^{kp}(F_p)) \quad (2)$$

where f_p^{kp} denotes the keypoint head, outputting the 2D keypoint heat maps. avg^i denotes i -axis marginalization by averaging. $P^{H,y}$ and $P^{H,x}$ are concatenated as joint representations F_P , which are later fed to the GCN model.

Figure 3. One-dimensional heat map for keypoint representation including (a) location embedding and (b) feature embedding.

3.1.2. Heat-Map-Based Keypoint Feature Embedding

We computed the K th keypoints heat map from an intermediate representation output from the backbone. During training, we trained the keypoint heat map generator under the supervision of the ground truth heat map using the MSE loss.

We aggregated the representations of all the pixels weighted by their degrees related to the k th keypoint, forming the k th keypoint/finger representation:

$$\mathbf{f}_k = \sum_{i \in \mathcal{I}} \tilde{m}_{ki} \mathbf{x}_i. \quad (3)$$

where $\mathbf{x}_i$ is the representation of pixel p_i . $\tilde{m}_{ki}$ is the normalized degree for pixel p_i related to the k th keypoint/finger representation. We used spatial softmax to normalize each keypoint heat map M_k to obtain $\tilde{m}_{ki}$.

3.2. Skeletal-Structural Graph Convolutional Network

The proposed skeletal-structural graph convolutional network analyzes the semantics of the whole-body keypoint to output the pose estimation of the whole body. It consists of cascaded structure-based graph layers and a data-dependent non-local layer. The structure-based graph layers are applied on the graph to capture the skeleton-structure information and to extract the high-level features. The data-dependent non-local layers are then employed to analyze long-range grouped joint features.

3.2.1. Structure-Based Graph Layer

The graph convolution operation on vertex v_i is represented as [41]:

$$f_{out}(v_i) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{in} \cdot w(l_i(v_j)) \quad (4)$$

where f denotes the feature map and v denotes the vertex of the graph. B_i denotes the sampling area of the convolution for v_i , which is defined as the one-distance neighbor vertexes v_j of the target vertex v_i . w is the weighting function, similar to the original convolution operation, which provides a weight vector based on the given input. Equation (4) can be transformed as follows:

$$f_{out} = \sum_k^{k_v} W_k(f_{in} A_k) \odot M_k \quad (5)$$

$$A_k = D_k^{-\frac{1}{2}} (\tilde{A}_k + I) D_k^{-\frac{1}{2}}, D_{ii} = \sum_k^{k_v} (\tilde{A}_k^{ij} + I_{ij}) \quad (6)$$

where k_v denotes the kernel size on the spatial dimension, $\tilde{A}_k$ is the adjacency matrix of the undirected graph representing intrabody connections, I is the identity matrix, and W_k is a trainable weight matrix.

3.2.2. Data-Dependent Non-Local Layer

Since the whole-body pose estimation task needs to deal with a large number of human joint points, the basic GCN cannot handle long-range whole-body relationships. Therefore, the data-dependent non-local layer is proposed to capture global and long-range relationships among nodes in the whole-body graph. We follow the non-local [42,43] concept and define the operation as:

$$x_i^{l+1} = x_i^l + \frac{W_x}{K} \sum_{j=1}^k f(x_i^l, x_j^l) \cdot g(x_j^l) \quad (7)$$

where W_x is initialized as zero, f denotes a pairwise function to compute the affinity between node i and all other j , and g is a function computing the representation of the node j . In practice, we implement Equation (7) in a similar way to [43].

3.2.3. Keypoint Group Representations

In order to deal with the complex structure of the whole-body network, the keypoints are grouped according to the semantics of the body structure. As shown in Figure 4, the graph pooling mechanism was applied to group the keypoints of the face, hand, and foot.

We grouped hand keypoints to represent the higher level of semantic relations. Specifically, we averaged the joint heat maps in groups. As shown in Figure 4, we grouped the keypoints of the face, hands, and feet. For a group of keypoints, set $\mathcal{G}$, the keypoint group representation is:

$$f_g = \frac{1}{N} \sum_{i \in \mathcal{G}} f_i \quad (8)$$

where f_i is the i th keypoint in the keypoint group $\mathcal{G}$. After the keypoints are grouped, the body graph is reconstructed, and then, the subsequent graph layer and non-local layer operations are performed.

In the implementation, we kept the joint points of body parts and grouped the feet, face, and hands separately. Corresponding to the key point definition of COCO wholebody dataset, the specific keypoint grouping sequence number is shown in Figure 4.

Figure 4. Visualization of keypoint groups in the SSGCN model.

3.3. Keypoint-Based Pose Estimation

Figure 5 shows the structure of the proposed SSGCN. We adopted the residual block [44] built of two GCN layers, followed by one local layer [43]. This block is repeated three times. In the second block, we grouped the keypoints of the face, hand, and foot for a higher level of the semantic feature.

Figure 5. Structure of the proposed SSGCN. We adopted the residual block [44] built by two GCN layers, followed by one local layer [43]. This block is repeated three times. In the second block, we grouped the keypoints of the face, hand, and foot for a higher level of the semantic feature.

The 2D detector takes the feature maps F and outputs 2D keypoint heat map $\mathcal{H}_j$, where $J = 133$ is the number of whole-body keypoints. Compared to regression methods, the heat map is more robust for the 2D keypoint detector. The 2D detector contains a two-layer CNN, and the pixels in 2D heat map $\mathcal{H}_j$ show the confidence of that pixel belonging to joint j . Supervised by ground truth 2D annotations, the 2D detector head is used for heat map extraction in the mid-term detection and the final keypoint output in the 2D task. The proposed SSGCN model outputs 1D heat maps of mesh vertices $M^H = \{M^{H,x}, M^{H,y}\}$. We converted the discretized heat maps of M^H to continuous coordinates $M^C = \{M^{C,x}, M^{C,y}\} \in \mathbb{R}^{V \times 2}$.

3.4. Loss Functions

The networks are supervised by an intermediate 2D heat map loss $\mathcal{L}_{2D}$ and final pose loss $\mathcal{L}_{1D}$. In our implementation, we trained the model with the combined loss $\mathcal{L}_{overall}$.

3.4.1. 2D Heat Map Loss

The 2D heat map loss is given as:

$$\mathcal{L}_{2D} = \sum_{j=1}^J \|\mathcal{H}_j - \hat{\mathcal{H}}_j\|_2^2 \quad (9)$$

where $\mathcal{H}_j$ and $\hat{\mathcal{H}}_j$ are the ground truth and estimated heat maps, respectively. We set the heat map resolution as 48×64 px. The ground truth heat map is defined as a 2D Gaussian with a standard deviation of 4 px centered on the ground truth 2D joint location.

3.4.2. 1D Heat Map Loss

To train the proposed model, we used the L1 loss function, defined as follows:

$$\mathcal{L}_{1D} = \|P^C - P^{C*}\|_1 \quad (10)$$

where P^C and P^{C*} indicate the output 1D heat maps and the ground truth 1D heat maps, respectively.

3.4.3. Overall Loss

We obtained the overall loss by summing the losses from all branches as follows:

$$\mathcal{L}_{overall} = \lambda \mathcal{L}_{2D} + \mathcal{L}_{1D} \quad (11)$$

where $\mathcal{L}_{2D}$ and $\mathcal{L}_{1D}$ are the intermediate 2D loss and the final pose 1D loss, respectively. The loss weight λ was set to 0.1.

4. Experiments and Results

4.1. Datasets and Metrics

Datasets. We followed the paper [33] and conducted experiments on the dataset COCO WholeBody. COCO WholeBody Dataset is a large-scale dataset with keypoint and bounding box annotations. As shown in Figure 6, this dataset extended the existing COCO keypoint dataset by further annotating face, hand, and foot keypoints. For each person in a valid box, there is a total of 133 keypoints (17 for body, 6 for feet, 68 for face, and 42 for hands). About 130 K face and left/right hand boxes are labeled, resulting in more than 800 K hand keypoints and 4 M face keypoints in total.

Metrics: All of our experiments were conducted on the COCO WholeBody Dataset, and the results are presented as the keypoint AP (APkp) and keypoint recall AR (ARkp) metrics without any test time augmentation. All results were obtained on the COCO WholeBody validation set.

4.2. Implementation Details

We adopted HRNet-w32 as our backbone network, and it takes images at a resolution of $\mathcal{I} \in \mathbb{R}^{192 \times 256}$ and outputs heat maps of $\mathcal{H} \in \mathbb{R}^{48 \times 64}$. We implemented our method with the PyTorch framework, and all experiments were conducted on NVIDIA GTX1080Ti GPUs. The networks were trained using the Adam optimizer with mini-batches of size 32. The learning rate was set as 10^{-3} to 10^{-4} , and the loss weight λ was set to 0.1.

Figure 6. Whole-body keypoints as defined in the COCO WholeBody Dataset.

4.3. Experimental Results

Learned weighting matrices:

The proposed network consists of four residual blocks, where each block contains two SemGConv layers. Figure 7 visualizes the learned weighting matrices of the SSGCN in the network, including the whole-body weight in (a), the torso weight in (b), and two hands' weight in (c) and (d). We made two important observations.

Figure 7. Visualization of learned weighting matrices M of the SSGCN in the network.

First, the weight in the lower right is larger than the upper left, which means that the central node has a higher impact than the end node. In other words, the keypoints of the torso posture guide the hand and foot posture estimation through the GCN. This verifies the calibration effect of the GCN module on the task of whole-body pose estimation.

Second, the specific structure of the body also has an important impact on the weights, for example the eyes and ears of the face (b) and the end joints of the fingers (c). These keypoints have a relatively fixed structural relationship, and the weights obtained by training were higher. Figure 7 proves that the GCN correctly parses the structure of the keypoints of the human skeleton and guides the estimation of the whole-body posture.

Ablation study: As shown in the ablation study Table 1, the results are presented as the keypoint AP (APkp) and keypoint recall AR (ARkp) metrics without any test time augmentation on the COCO WholeBody Dataset. Specifically, we evaluated (1) our model design compared to the baseline HRNet-w32, (2) the basic GCN (3), intermediate supervision, (4) and the keypoint group representations. As a baseline, we compared to [33] as the implementation used the HRNet-w32 module and input size 192×256 . The proposed SSGCN module significantly improved the final results.

Table 1. Ablation study on the COCO WholeBody Dataset. All models were trained with the HRNet-w32 backbone. “GCN”, “IS”, and “Group” represent the graph convolutional network, intermediate supervision, and keypoint group representation, respectively.

Method	Body		Foot		Face		Hand		Whole-Body		All Mean	
	AP	AR	AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Baseline (HRNet-w32)	65.9	70.9	31.4	42.4	52.0	58.2	30.0	36.3	43.2	52.0	44.6	51.9
GCN	68.5	73.4	43.5	55.3	54.6	63.5	39.1	44.8	48.0	57.5	50.7	58.9
GCN + IS	74.2	79.6	79.1	84.9	63.2	73.7	47.0	58.8	54.5	66.6	63.6	72.7
GCN + IS + Group (Full)	73.8	79.3	80.7	87.1	75.9	87.9	52.5	65.2	55.0	67.2	67.5	77.3

Comparison with the state-of-the-art: As shown in Table 2, the best results are boldfaced, and the SSGCN performed the best among all models. For bottom-up methods, although they have the advantage of speed in the case of multi-person pose estimation, it is still difficult to achieve the accuracy of top-down methods. The proposed SSGCN method belongs to the top-down structure and surpassed all the bottom-up methods. In the final whole-body AP and AR metrics, our method outperformed the latest bottom-up method HPRNet [40] by about 16~20.

For the top-down methods, our method still achieved optimality in most of the metrics. Although it was slightly lower than ZoomNet in body torso recognition, the SSGCN was significantly ahead of other methods in the hand and face metrics. Our method significantly outperformed the latest methods such as ZoomNet [33] in the estimation of hands, feet, and faces due to the consideration of the relationship between the semantics of the end limbs and the semantics of the whole torso. Overall, the SSGCN was ahead of the other methods in whole-body metrics and faster than traditional top-down methods such as ZoomNet in inference.

Model complexity: Compared with the traditional whole-body human pose estimation method, the proposed SSGCN has fewer parameters (34.52 M) and less computational complexity (4.17 GMac). Since most of the existing methods for whole-body pose estimation only analyze the inference speed without publishing the number of parameters and the amount of computation (FLOPs), in Table 3, we focus on the comparative analysis of the running speed. The proposed model was able to achieve a running time of 125 ms, which is significantly faster than traditional methods such as SN and ZoomNet, and the speed was close to the bottom-up methods such as PAF. ZoomNet [33] also achieved a relatively high accuracy, but they used multiple independent models to detect each part of the body separately, which resulted in a slow runtime. Although OpenPose [32] ran the fastest, it was far below our method in terms of accuracy. The experimental results showed that

our proposed SSGCN model can significantly improve the accuracy of whole-body pose recognition with only a small increase in the model complexity.

Table 2. Comparison with the state-of-the-art on the COCO WholeBody validation set.

Method	Body		Foot		Face		Hand		Whole-Body		All Mean	
	AP	AR	AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Bottom-up methods:												
PAF [35]	26.6	32.8	10.0	25.7	30.9	36.2	13.3	32.1	14.1	18.5	19.0	29.1
SN [34]	28.0	33.6	12.1	27.7	38.2	44.0	13.8	33.6	16.1	20.9	21.6	32.0
AE [45]	40.5	46.4	7.7	16.0	47.7	58.0	34.1	43.5	27.4	35.0	31.5	39.8
HPRNet [40]	59.4	68.3	53.0	65.4	75.4	86.8	50.4	64.2	34.8	49.2	54.6	66.8
Top-down methods:												
OpenPose [32]	56.3	61.2	53.2	64.5	48.2	62.6	19.8	34.2	33.8	44.9	42.3	53.5
HRNet [38]	65.9	70.9	31.4	42.4	52.3	58.2	30.0	36.3	43.2	52.0	44.6	51.9
ZoomNet [33]	74.3	80.2	79.8	86.9	62.3	70.1	40.1	49.8	54.1	65.8	62.1	70.6
Ours	73.8	79.3	80.7	87.1	75.9	87.9	52.5	65.2	55.0	67.2	67.5	77.3

Table 3. Comparison in terms of model accuracy and processing speed.

Method	All Mean (AP)	All Mean (AR)	R. Time (ms)
PAF [35]	19.0	29.1	100
SN [34]	21.6	32.0	216
HPRNet [40]	54.6	66.8	101
OpenPose [32]	42.3	53.5	45
ZoomNet [33]	62.1	70.6	175
Ours	67.5	77.3	125

Qualitative results: In Figure 8, we show the visual results of our method on the COCO WholeBody Dataset. As seen, our method was able to accurately predict the whole-body pose in various complex actions and scenes, including complex situations where limbs overlap and are obscured. This indicates that SSGCN can effectively encode the relationships among whole-body joints.

Figure 8. Sample whole-body keypoint detection results of the SSGCN.

4.4. Analysis

Intermediate supervision: The supervision for the heat map ensures that the keypoint heat map can accurately identify the characteristics of each joint point. The ablation study showed that the quality of the keypoint heat map had a great impact on the final results. Intermediate supervision ensures the accuracy of the heat map, which would have a very big impact on the subsequent GCN network.

Semantic–structural graph convolutional network: The proposed semantic–structural graph convolutional network leverages the whole-body graph structure to analyze the semantics of each part of the body and improve the accuracy of pose estimation. By constructing a whole-body skeleton map to connect the semantic features of individual parts of the whole body together, the accuracy of whole-body pose estimation is greatly improved while increasing by a small amount of the model’s computation.

Keypoint groups: The group mechanism enables the model to handle higher-level semantic information, and the nonlocal mechanism helps to handle the remote relationship of joint nodes. As shown in Table 1, the semantic–structural graph convolutional module significantly improves the whole-body pose estimation accuracy. Although the ablation experimental results showed that the grouped graph convolutional machine had some disturbing effects on the estimation of the body, there were significant improvements in the foot, face, and hand metrics.

5. Conclusions

In this paper, we performed the semantic fusion of whole-body poses based on the whole-body skeleton and leveraged the heat-map-based graph convolutional network to calibrate human whole-body human pose estimation. We proposed a novel semantic–structural graph convolutional network scheme for hand pose estimation. The proposed scheme leverages the whole-body graph structure to analyze the semantics of each part of the body through the graph convolutional network and improves the accuracy of pose estimation. Our method achieved competitive performance on whole-body pose estimation benchmarks, and at the same time, there was only a small increase in the computational effort and model complexity. The experimental results showed that integrating whole-body semantics based on the GCN had a great improvement on whole-body recognition, especially on pose estimation such as the hands and face. We believe that a whole-body pose semantic fusion approach through skeletal–structure-based pose estimation is an important direction to optimize the whole-body pose estimation task.

Author Contributions: Conceptualization, W.L.; funding acquisition, S.C.; methodology, W.L.; Writing—original draft, W.L.; writing—review and editing, R.D. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDC02070600.

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

References

1. Cimen, G.; Maurhofer, C.; Sumner, B.; Guay, M. Ar poser: Automatically augmenting mobile pictures with digital avatars imitating poses. In Proceedings of the 12th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing, Madrid, Spain, 18–20 July 2018.
2. Elhayek, A.; Kovalenko, O.; Murthy, P.; Malik, J.; Stricker, D. Fully automatic multi-person human motion capture for vr applications. In Proceedings of the International Conference on Virtual Reality and Augmented Reality, London, UK, 22–23 October 2018.
3. Xu, W.; Chatterjee, A.; Zollhoefer, M.; Rhodin, H.; Fua, P.; Seidel, H.P.; Theobalt, C. Mo2cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 2093–2101. [[CrossRef](#)] [[PubMed](#)]
4. Choi, H.; Moon, G.; Lee, K.M. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In Proceedings of the European Conference on Computer Vision, online, 23–28 August 2020.

5. Kundu, J.N.; Rakesh, M.; Jampani, V.; Venkatesh, R.M.; Babu, R.V. Appearance Consensus Driven Self-supervised Human Mesh Recovery. In Proceedings of the European Conference on Computer Vision, online, 23–28 August 2020.
6. Iqbal, U.; Xie, K.; Guo, Y.; Kautz, J.; Molchanov, P. KAMA: 3D Keypoint Aware Body Mesh Articulation. *arXiv* **2021**, arXiv:2104.13502.
7. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
8. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
9. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
10. Yan, A.; Wang, Y.; Li, Z.; Qiao, Y. PA3D: Pose-action 3D machine for video recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
11. Huang, L.; Huang, Y.; Ouyang, W.; Wang, L. Part-aligned pose-guided recurrent network for action recognition. *Pattern Recognit.* **2019**, *92*, 165–176. [[CrossRef](#)]
12. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3d pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
13. Moeslund, T.B.; Granum, E. A Survey of Computer Vision-Based Human Motion Capture. *Comput. Vis. Image Underst.* **2001**, *81*, 231–268. [[CrossRef](#)]
14. 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.
15. Tompson, J.; Jain, A.; LeCun, Y.; Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *Adv. Neural Inf. Process. Syst.* **2014**, *1*, 1799–1807.
16. Ramakrishna, V.; Munoz, D.; Hebert, M.; Bagnell, J.A.; Sheikh, Y. Pose machines: Articulated pose estimation via inference machines. In Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2014.
17. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
18. Luo, Y.; Ren, J.; Wang, Z.; Sun, W.; Pan, J.; Liu, J.; Pang, J.; Lin, L. Lstm pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
19. Artacho, B.; Savakis, A. UniPose: Unified Human Pose Estimation in Single Images and Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
20. Athitsos, V.; Sclaroff, S. Estimating 3D hand pose from a cluttered image. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; Volume 2, p. II–432. [[CrossRef](#)]
21. de La Gorce, M.; Fleet, D.J.; Paragios, N. Model-Based 3D Hand Pose Estimation from Monocular Video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1793–1805. [[CrossRef](#)] [[PubMed](#)]
22. Cao, X.; Wei, Y.; Wen, F.; Sun, J. Face alignment by explicit shape regression. *Int. J. Comput. Vis.* **2014**, *107*, 177–190. [[CrossRef](#)]
23. Tzimiropoulos, G. Project-out cascaded regression with an application to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
24. Trigeorgis, G.; Snape, P.; Nicolaou, M.A.; Antonakos, E.; Zafeiriou, S. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
25. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 918–930. [[CrossRef](#)] [[PubMed](#)]
26. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
27. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [[CrossRef](#)]
28. Oberweger, M.; Wohlhart, P.; Lepetit, V. Hands deep in deep learning for hand pose estimation. In Proceedings of the 20th Computer Vision Winter Workshop, Seggau, Austria, 9–11 February 2015.
29. Oberweger, M.; Lepetit, V. DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017.
30. Sharp, T.; Keskin, C.; Robertson, D.; Taylor, J.; Shotton, J.; Kim, D.; Rhemann, C.; Leichter, I.; Vinnikov, A.; Wei, Y.; et al. Accurate, robust, and flexible real-time hand tracking. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015.
31. Sridhar, S.; Mueller, F.; Oulasvirta, A.; Theobalt, C. Fast and robust hand tracking using detection-guided optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
32. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [[CrossRef](#)] [[PubMed](#)]

33. 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, online, 23–28 August 2020.
34. Hidalgo, G.; Raaj, Y.; Idrees, H.; Xiang, D.; Joo, H.; Simon, T.; Sheikh, Y. Single-network whole-body pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
35. 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, Honolulu, HI, USA, 21–26 July 2017.
36. 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 Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2014.
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *1*, 91–99. [[CrossRef](#)] [[PubMed](#)]
38. 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, Long Beach, CA, USA, 15–20 June 2019.
39. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
40. Samet, N.; Akbas, E. HPRNet: Hierarchical Point Regression for Whole-Body Human Pose Estimation. *arXiv* **2021**, arXiv:2106.04269.
41. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *arXiv* **2018**, arXiv:1801.07455.
42. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; Volume 2, pp. 60–65. [[CrossRef](#)]
43. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803. [[CrossRef](#)]
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Newell, A.; Deng, J.; Huang, Z. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. *Adv. Neural Inf. Process. Syst.* **2017**, *1*, 2274–2284.