sensors

MDPI

Article

3DMesh-GAR: 3D Human Body Mesh-Based Method for Group Activity Recognition

Muhammad Saqlain [1], Donguk Kim [1], Junuk Cha [1], Changhwa Lee [2], Seongyeong Lee [2] and Seungryul Baek [1,*]

[1] AI Graduate School, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea; saqlain@unist.ac.kr (M.S.); dukim@unist.ac.kr (D.K.); jucha@unist.ac.kr (J.C.)
[2] Department of Computer Science and Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea; changhwalee@unist.ac.kr (C.L.); skwithu@unist.ac.kr (S.L.)
* Correspondence: srbaek@unist.ac.kr; Tel.: +82-52-217-2205

Abstract: Group activity recognition is a prime research topic in video understanding and has many practical applications, such as crowd behavior monitoring, video surveillance, etc. To understand the multi-person/group action, the model should not only identify the individual person's action in the context but also describe their collective activity. A lot of previous works adopt skeleton-based approaches with graph convolutional networks for group activity recognition. However, these approaches are subject to limitation in scalability, robustness, and interoperability. In this paper, we propose 3DMesh-GAR, a novel approach to 3D human body Mesh-based Group Activity Recognition, which relies on a body center heatmap, camera map, and mesh parameter map instead of the complex and noisy 3D skeleton of each person of the input frames. We adopt a 3D mesh creation method, which is conceptually simple, single-stage, and bounding box free, and is able to handle highly occluded and multi-person scenes without any additional computational cost. We implement 3DMesh-GAR on a standard group activity dataset: the Collective Activity Dataset, and achieve state-of-the-art performance for group activity recognition.

Keywords: 3D human activity recognition; human body mesh estimation; feature extraction; deep learning; video understanding

check for updates

Citation: Saqlain, M.; Kim, D.; Cha, J.; Lee, C.; Lee, S.; Baek, S. 3DMesh-GAR: 3D Human Body Mesh-Based Method for Group Activity Recognition. *Sensors* **2022**, *22*, 1464. https://doi.org/10.3390/s22041464

Academic Editor: Muhammad Arsalan

Received: 27 December 2021
Accepted: 10 February 2022
Published: 14 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

The purpose of human activity recognition is to identify what a human is doing in a scene using images and video frames or inertial, environmental, and physiological sensors data [1]. It is one of the most active areas of research and an immensely significant component of computer vision and computer graphics fields. Group activity recognition (GAR) is a subset of the human activity recognition problem that focuses on a group of people's collective behavior resulting from their individual actions and interactions [2]. The GAR is a critical task in many domains for automatic analysis of human behavior, including intelligent surveillance, crowd monitoring, human-computer interaction, social behavior comprehension, robotics, sports video analysis, virtual reality, etc. [3]. Additionally, with the increasing population of elderly people, GAR is becoming a powerful tool to monitor functional, cognitive, and physical health at their homes or in hospitals [4]. Furthermore, it is critical to determine individual activities and interactions of people when recognizing group activity because these actions and interactions often constitute the group activity. Recent studies for this task have explored different methods for feature representations, such as optical flow [5], RGB (i.e., red, green, blue) image sequences [6,7], human skeletons [8], depth image sequences [9], and audio waves [10].

The majority of group activity recognition models either explicitly or implicitly examine human actions. Some studies identify individual activities and group activities in a combined framework using probabilistic graphical models [11] or neural networks that implement the capability of graphical models [12]. Other approaches simulate the

relationship between individual persons' activities and group activities by applying various pooling operations such as max-pooling [13] or attention pooling [14] on individual person representations. The temporal evolution of individual actions and group activities is another significant component in recognizing group activities. Recurrent Neural Networks (RNNs) have been used in several studies to recognize individual and group actions over time [14,15]. This strategy provides a concrete way to model group activities in video datasets. Moreover, some existing works attempt to inject temporal data via Convolutional Neural Networks (CNNs) on optical flow fields computed between two consecutive frames as an additional input to each time step of RNNs [15,16]. Recently, multi-stream convolutional networks beat RNNs in action identification tests [17], where CNN model temporal information on optical flow fields. However, this method can be expensive due to the computation of optical flow for numerous individuals and several forward runs of CNNs on the input optical flow field for all people in the video.

Researchers have proposed new action detection methods that can run on simpler hardware with few restrictive constraints in recent years. These methods do not need high-resolution cameras, special bodysuits, or in-studio recording but only require some scaled cameras for capturing the persons who are doing their everyday activities [18–22]. Moreover, various recent studies have used Deep Neural Networks (DNNs) to create 2D or 3D skeleton data with pose and shape information from a single RGB image, which are further used for action recognition tasks [3,8,23]. Unfortunately, most of the previous methods need more computation power and fail to get the required results for multi-person action recognition due to high occlusions, complicated backgrounds, a large variety of scenes and appearances, and depth uncertainties [24]. Additionally, current 3D skeleton methods show a deficiency in precise modeling of body-bone length distribution, which may predict impractical body structure such as abnormal limbs proportions and right-left asymmetry [25].

More recently, 3D human body mesh reconstruction aims to create full-body 3D meshes of humans in an image or video [26], and further, these meshes are being widely used for motion re-targeting [27], action recognition [28], virtual try-on [29], etc. A single-stage 3D mesh reconstruction method called ROMP (i.e., Regress all meshes in a One-stage fashion for Multiple 3D People) is presented for multi-person scenarios [21]. Therefore, this paper proposed a 3D human body mesh-based method for multi-person action recognition called 3DMesh-GAR. Our method is composed of three stages. Stage I contains a 3D mesh reconstruction network, which takes RGB image frames as the input, applies a mesh reconstruction network, and regresses the 3D body meshes of all people in the frame. Our mesh reconstruction network is computation cost-efficient, bounding box-free, and can learn pixel-level features in an end-to-end fashion. Stage II consists of the concatenation method and the feature extraction network. During the concatenation, all body meshes created in Stage I are merged into a single 3D body mesh by averaging them. Next, a feature extraction network is applied to the concatenated body mesh, converting the complex 3D mesh parameter into simple 2D trainable features. Finally, these features are treated as the input of a fully connected Deep Artificial Neural Network (D-ANN) in Stage III, which decides the action class of the input frame. We evaluate the performance of our method on a benchmark group activity recognition dataset called the Collective Activity Dataset [30]. The experimental results demonstrate that the proposed 3D mesh-based action recognition method obtained superior performance to the current state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 contains related studies. Section 3 explains research materials and methods followed by this study. Dataset, experiments, and result analyses are provided in Section 4. The whole study is concluded in Section 5.

2. Related Work

We briefly review the recent literature on various deep learning-based and 3D skeleton-based group activity recognition approaches (Section 2.1). Then, we analyze recent one-

stage/multi-stage methods designed for a single person and multi-person 3D human body mesh reconstruction (Section 2.2).

2.1. Group Activity Recognition Approaches

For the last decade, the computer vision research community has widely studied group activity recognition from video datasets. Most of these studies are based on visual features extracted using some 2D Convolutional Neural Networks (2D-CNNs) for each individual in an input frame and then building probability graphical networks on the prior knowledge for recognizing group activities [31–34]. Recently, CNNs models have achieved extensive success in activity recognition problems [14,15,35–37]. Additionally, ResNet [38], derivatives [39], and Inception [40] consolidate explicit knowledge flowing from initial to later feature extraction layers in the network via skip and summation connections. This strategy allows the training of deeper and more powerful models. With the advancements in deep learning techniques, feature extraction from input objects has been jointly optimized with the latest relational modeling methods such as Deep CNNs [41], Graph Neural Networks (GNNs) [42–44], Recurrent Neural Networks (RNNs) [38,45–47], and Transformers [48–50].

There have been several attempts to address the problem of group activity detection using probabilistic graphical models. Sun et al. [51] used a latent graph model for multi-target tracking, activity group localization, and group recognition. Hand-crafted features are used as input to the model in the initial probabilistic techniques. With the recent success of deep neural networks in various computer vision applications, these networks are now being used as feature extractors and inference engines in probabilistic group activity detection models. Deng et al. used CNNs as an initial classifier to come up with unary potentials [52]. They used a deep neural network to create the graphical model and performed messages traveling through the network to refine initial predictions. The authors presented multi-stream convolutional frameworks for group activity recognition in [2,53], in which new input modalities are simply included in the model by adding new convolutional streams. Li et al. proposed a real-time inference method for multi-person tracking and collective activity detection at individual, interaction, and group activity levels [54]. Simonyan and Zisserman [5] employed a two-stream CNNs architecture that can independently attain representation on optical flow assembled frames and RGB images. Azar et al. represented a CNNs based spatial relational method for group activity detection [41]. Wang et al. proposed an effective CNNs framework for action recognition from videos by dividing the input video into many chunks and applying a multi-stream method to combine each chunk with their corresponding part in a learnable way [17].

Some of the recent studies depend on spatio-temporal information extracted using RNNs. Ramanathan et al. proposed a multi-stage RNNs model to recognize only similar events in input videos by extinguishing irrelevant information [14]. Their model learns to recognize activities in videos while spontaneously attending to the main objects responsible for an activity. Ibrahim et al. [13] presented a deep architecture for modeling group activities in a principled structured temporal framework. The first part of their two-stage technique modeled individual-level activities, then merged all individual-level information to reflect collective activities. The Long Short-Term Memory (LSTM) network was used to represent the temporal representation of the model. Ibrahim et al. [37] introduced a hierarchical relational network that computed relational representations of persons based on graph structures that describe their potential relationships. Individual human representations and hypothetical relationship networks were provided to each related layer. Based on their connections with corresponding graphs, relational representations of each person were constructed. This method can be used to acquire relational feature representations that can successfully distinguish between different types of single-person and group activities. Another study looked into the use of RNNs for message passing [12]. They also proposed gating functions for learning the graph's structure. A few studies also look at using structured RNNs to predict the scene context [15,55] or generate captions [16]. For group

activity recognition, Li et al. suggested a two-stage semantic-based approach [16]. In the first stage, they implement the LSTM model to create captions for all video frames, while in the second stage, another LSTM model is implemented to identify the final activity class based on the created captions in the first stage.

Although attention mechanisms were originally designed for Natural Language Processing (NLP) problems, recently researchers have also proposed them for group activity detection by consolidating attention via pooling methods [56], graphs [57], or LSTM models [58]. Tang et al. joined attention mechanisms to get compact representations by allocating varying pooling weights to the various individual or group interactions [59]. Lu et al. proposed a spatio-temporal attention mechanisms-based method to utilize spatial configuration and temporal dynamic in a collective scene [60]. Moreover, attention-based models can also be applied for various modalities, such as motion [58] and pose [61]. Fernando et al. represented a temporal pooling function-based method for learning features of an input video through ranking machines. Later on, these features were used to recognize the input video with some classifier such as support vectors machines (SVMs) [62].

Graph Convolutional Networks (GCNs), a semi-supervised learning method, has recently become an emerging research topic in deep learning [63]. Some researchers have applied GCNs to recognize single-human activity [8,57] and group activity [42]. Wu et al. [42] proposed a flexible and efficient Actor Relation Graph (ARG) to simultaneously capture the appearance and position relation between actors. The connections in ARG were automatically learned using the GCNs from group activity videos in an end-to-end manner, and the inference on ARG were efficiently performed with standard matrix operations. Vaswani et al. proposed a transformer network that can learn long-term dependencies in a better way as compared to RNN due to its self-attention mechanism [45]. Girdhar et al. introduced a transformer network combined with 3D CNN representation for video action localization and action recognition [64]. Gavrilyuk et al. proposed an actor-transformer method for group activity recognition, which uses optical flow for temporal dynamics representation while pose information has been applied for interpreting spatial information of multiple people [65]. However, using optical flow information as input requires a high computational cost. Moreover, most of the current studies for group activity recognition consist of complex multi-stage architecture and required bounding boxes and well-designed hand-crafted features for each individual in the frame. Our method is simple, easy to use, bounding box free, and uses 3D mesh features for group activity recognition.

2.2. 3D Human Body Mesh Reconstruction Methods

Previous studies estimate monocular 3D-pose estimation for a single person in the form of a non-parametric 3D shape [66,67] or body skeleton [68–72]. The SMPL (i.e., Skinned Multi-Person Linear model) parametric model [73] has also been widely used for human body mesh recovery. The SMPL is adapted to convert a highly complicated 3D body mesh into a simple vector with very low dimensions, which can be regressed by an image [74]. Bogo et al. proposed the first optimization-based method called SMPLify, which can continuously train SMPL with the learned 2D joints [75].

Some recent studies applied deep neural networks in a multi-stage manner for direct SMPL parameters regression. These studies first approximate intermediary representations such as silhouettes and keypoints from the input images and then regress SMPL parameters by mapping them [76–78]. Some other studies regressed SMPL parameters directly from images, either by leveraging temporal learning [79,80] or complex model training methods [81,82]. Moreover, some researchers employed CNNs-based learning methods to get satisfactory results for 3D pose estimation [83–85]. Transfer learning methods have also been implemented to enhance 3D pose estimation using features obtained from 2D pose datasets [86–88]. Recently, Cha et al. proposed a self-supervised learning based method for 3D human body pose and shape estimation using only 2D images [89]. Their method does not require other type of supervision signals such as video-level, multi-view priors, or 2D/3D skeletons. However, all these methods achieved higher accuracy only

for single-person problems, and it remains ambiguous how to use these methods for more common multi-person problems.

For multi-person 3D regression, most of the existing methods are composed of a multi-stage approach that provides the single-person model with a 2D person identifier to solve multi-person problems [90,91]. Recently, many researchers have been focusing on multi-person 3D pose estimation, for which they use a top-down paradigm [92,93]. In top-down methods, firstly, all individual person instances are detected, and then features are extracted using the bounding-box method, and finally, body joins locations are regressed using those features for each person [76,80,81,94]. Different multi-stage methods have also been proposed for this purpose using Faster Region-based Convolutional Neural Networks (R-CNN) [95], such as 3D Multi-Person Pose Estimation (3DMPPE) [96] and Localization Classification Regression Network (LCR-Net++) [93]. Moon et al. applied a prior person identification step and transferred all resized and cropped images of identified persons to the 3D pose estimation network [96]. However, top-down methods depend considerably on human detection for localization of each individual prior to body joint estimation within the identified bounding boxes [97,98]. These methods have no realization of persons who are out of the bounding boxes and their possible interaction. Moreover, human detection becomes unreliable for highly occluded scenarios, resulting in misleading the pose estimation of the targeted person with the nearby persons.

Recent bottom-up approaches for 3D mesh estimation do not require individual person detection and thus can achieve higher accuracy in the presence of multi-person scenarios [99,100]. These methods take all persons in a frame simultaneously and distinguish their joints in a better way. Fabbri et al. proposed a pose estimation method called Learning on Compressed Output (LoCO) for mapping the images into the volumetric heatmaps and using these heatmaps to estimate multi-person 3D poses through an encoder-decoder framework [101]. More recently, Zhang et al. represented a single-stage method that shows instances of multi-person in the space of spatial depth where all points are associated with their corresponding body meshes [26]. Their method can directly identify human body meshes through concurrently localizing human instance points and predicting corresponding 3D meshes. Sun et al. also proposed a single-stage method called ROMP for 3D mesh regression of multi-person scenarios [21]. ROMP can simultaneously identify mesh parameter map and body center heatmap, which are jointly used to predict a 3D person map on the pixel level. Their method achieved state-of-the-art results on the highly occluded and crowded dataset. So, we also used a single-stage 3D mesh reconstruction method as the backbone of our pipeline for 3D mesh regression from the group activity datasets.

3. Materials and Methods

The details of our proposed method are given in this section. Figure 1 illustrates all the important elements of the proposed learning framework for group action recognition from 3D human body meshes. We first show how the 3D human body meshes are created in multi-person scenarios from simple RGB images (Section 3.1). Then, we introduce a concatenation method to aggregate all the meshes into a single 3D mesh and a feature extraction network to convert 3D mesh parameters into trainable features (Section 3.2). Finally, a fully connected deep neural network is presented to train and classify the group actions from the trainable features (Section 3.3).

3.1. 3D Body Mesh Reconstruction: Stage I

Stage I contains a 3D body mesh reconstruction network for body mesh regression from the multi-person dataset. As illustrated in our framework, a ResNet-50 [38] network is applied as the default backbone to the input RGB image of size 512 × 512 and extracts a backbone feature vector $yf \in \mathbb{R}^{34\times H_b\times W_b}$ where H_b and W_b are the height and width of the backbone feature, respectively, their values are set to 128. From the backbone feature, three different head networks are built to find three types of maps, such as body center heatmap (C_m), camera map (A_m), and SMPL map (S_m). These maps comprehensively describe the

estimated 3D body mesh information. The body center heatmap predicts the probability of all positions being people body centers. Using these position parameters, the camera map and SMPL maps are applied to get camera parameters and SMPL parameters, respectively, which are further gathered to define the 3D mesh parameter map (P_m). The size of all maps is given by $n \times H \times W$, where n represents the total number of channels, and H and W are the height and width of the maps, respectively. The value of both height and width of all the maps is set to 64. Each map is further elaborated in the following subsections.

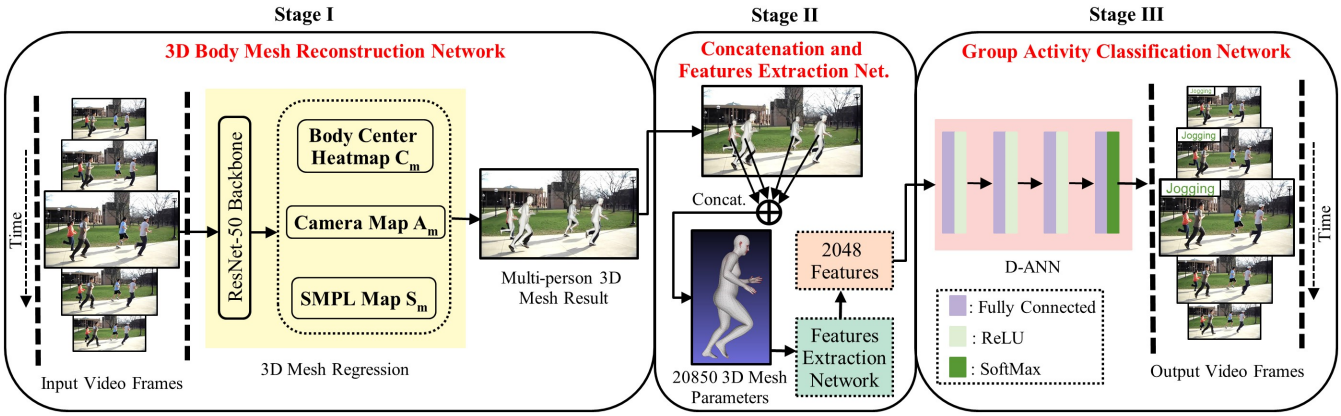

Figure 1. An overview of our 3DMesh-GAR framework for 3D body mesh-based group activity recognition. Stage I infers 3D body mesh reconstruction for each input RGB frame in a video using a 3D mesh regression model. Stage II provides the concatenation and features extraction networks used to concatenate all 3D body meshes of the frame into a single averaged 3D mesh and extract learnable features from the concatenated 3D body mesh, respectively. Finally, Stage III provides a fully-connected deep artificial neural network (D-ANN) trained with learnable mesh features to classify the group activities.

3.1.1. Body Center Heatmap: $C_m \in \mathbb{R}^{1\times H\times W}$

The C_m map is the heatmap showing the 2D person body's central point in the input RGB image. All body centers are presented as a Gaussian distribution in the (C_m), calculated by Gaussian kernel size k of all person centers in terms of their 2D bodies. The value of k is derived by the following equation:

$$k = k_l + (\frac{d_{b-box}}{\sqrt{2}W})^2 k_r \quad (1)$$

where k_l, d_{b-box}, W, and k_r are minimum kernel size, diagonal length of the human bounding box, the width of the body center heatmap, and variation range of k, respectively. The values of k_l and k_r were set to 2 and 5, respectively, by default. The body center heatmaps of all persons from different actions of the collective active dataset [30] are shown in Figure 2b.

3.1.2. Camera Map: $A_m \in \mathbb{R}^{3\times H\times W}$

The camera map consists of three camera parameters, one 2D scale s parameter, and two translation $t = (t_x, t_y)$ parameters for each person in the image. The scale s represents the size and depth of the person's body. Whereas the translation t_x and t_y, whose values range in $(-1,1)$, represent a normalized translation of the person's body relative to the center of the image on the x-axis and y-axis, respectively.

3.1.3. SMPL Map: $S_m \in \mathbb{R}^{142\times H\times W}$

The S_m map consists of 142-dimensional SMPL parameters, obtained by employing the SMPL parametric model for person body representation [73]. It allows the use of pose parameters θ and shape parameters β to describe the full 3D human body mesh. The pose

parameters are defined as $\theta \in \mathbb{R}^{6\times22}$, containing a 6D representation of 3D rotational information for 22 human body joints. The shape parameters are defined as $\beta \in \mathbb{R}^{10}$, which are parameterized by the top-10 principal components of the 3D shape space. An efficient mapping is established by applying an SMPL differentiable function that takes the θ and β parameters as input and results in a triangular body mesh $M \in \mathbb{R}^{6890\times3}$ with 6890 vertices. The 3D joints are reconstructed by a PM process, where P is defined as $P \in \mathbb{R}^{K\times6890}$, which is an infrequent weight matrix that expresses the linear mapping through 6890 vertices of human body mesh M to the body joint K.

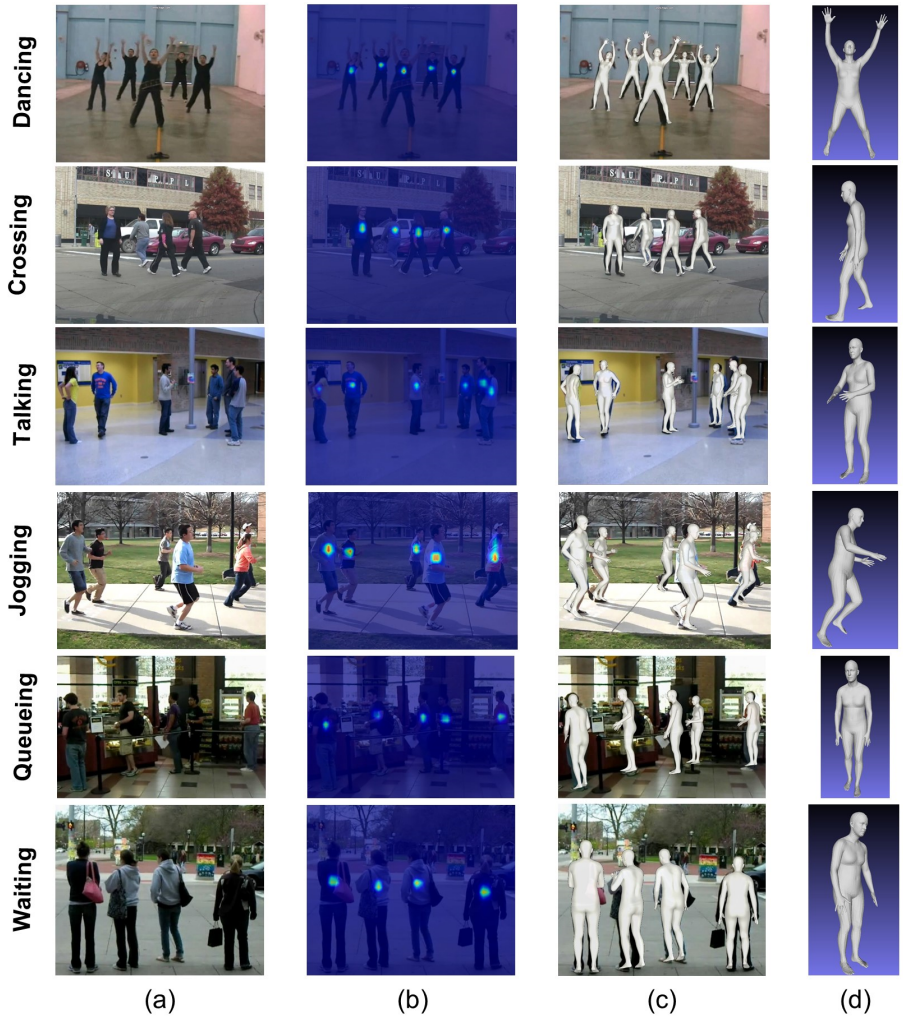

Figure 2. 3D body mesh reconstruction examples: We applied a 3D mesh reconstruction network to create 3D body meshes for each individual person from the input RGB frames. Each row presents examples from the Collective Activity Dataset, and each column corresponds to (**a**) input RGB frames, (**b**) body center heatmaps of each person in the frame, (**c**) 3D body meshes of each person in the frame, and (**d**) concatenated 3D body mesh, respectively.

3.1.4. Mesh Parameter Map: $P_m \in \mathbb{R}^{145\times H\times W}$

The mesh parameter map is a combination of the SMPL map and camera map. The 3D human body mesh parameters are estimated by considering the positions of SMPL and camera maps to the centers of the human bodies. A weakly-perspective camera model is implemented to estimate 3D human body joints $J = (x_k, y_k, z_k)$, where $k = 1 \ldots K$. These 3D body joints are used to get 2D projection joints $\widehat{J} = (\widehat{x}_k, \widehat{y}_k)$, where $\widehat{x}_k$ and $\widehat{y}_k$ are derived by camera parameters as $\widehat{x}_k = sx_k + t_x$, $\widehat{y}_k = sy_k + t_y$, respectively. It helps the mesh reconstruction model to train with 2D pose datasets, which increases the generalization and strength.

3.1.5. Collision-Aware Representation (CAR)

Conventionally, human body centers are defined using the center of bounding boxes. However, this method failed to identify the body centers in highly occluded scenes. Thus, Sun et al. [21] proposed a novel method called collision-aware representation (CAR) to define the human body centers in densely overlapping people cases. Using this method, a repulsion field is created, in which each body center is considered positively charged. These same charged body centers repel each other, and their radius of repulsion is the same as the size of the Gaussian kernel defined by Equation (1). So, the CAR plays a vital role in pushing apart the closer body centers in multi-person occluded cases and using these differential body centers to create a body center heatmap. The impact of CAR can be seen in the *Tracking* and *Jogging* rows of Figure 2, where heatmaps of occluded persons are successfully identified. Moreover, by sampling these body center heatmaps with mesh parameter maps, 3D mesh features are extracted for each individual person, as shown in Figure 2c.

3.2. Concatenation and Features Extraction: Stage II

After Stage I, we have a per-frame 3D body mesh for each individual. The Stage II of our framework contains two simple networks such as (1) concatenation network and (2) feature extraction network. The prior network gets the output of Stage I with multi-person 3D mesh results as its input and implements a concatenation function to find a single 3D body mesh by averaging all body meshes. The value of concatenated 3D body mesh S^{3D} is derived as:

$$S^{3D} = \left(\frac{S_1^{3D} + S_2^{3D} + \cdots + S_n^{3D}}{n}\right), n = 1 \ldots N \tag{2}$$

where N is the total number of persons in the input frame. As each input frame belongs to one of the group activities, all persons in that specific frame represent the corresponding activity. Thus, we concatenated all frames with different numbers of persons into a single 3D body mesh corresponding to a specific group activity class, see Figure 2d. The final 3D body mesh contains three types of parameters, such as 3D coordinates of body mesh, SMPL pose, and 2D coordinates of 2D keypoints of shapes $(6890, 3)$, $(72, 1)$, and $(54, 2)$, respectively. So, the total number of raw parameters of a single 3D body mesh is $20,850$. The second network of Stage II takes the raw mesh parameters from the concatenation network as the input. It applies a simple non-trainable linear network to convert the input 3D mesh parameters into trainable parameters of size 2048.

3.3. Activity Classification Network: Stage III

After Stage II, we have trainable features for each input RGB image. Stage III contains a deep artificial neural network (D-ANN) that takes the trainable features as the input and passes them to the multiple hidden layers for training using activation functions. It gives the output with one of the group activity classes. A D-ANN is a multi-layer fully-connected neural network and comprises an input layer, several hidden layers, and an output layer. All nodes of one layer are connected to all other nodes in the next layer. Nowadays, D-ANN models are being used in several real-world applications because of their outstanding performances [102]. The success of deep learning models from the last decade is due to a combination of both theoretical progression such as improved learning rate methods, optimization techniques, availability of numerous big datasets, etc., and easy access to improved and cheap hardware resources such as multi-processor graphics cards or graphics processing units (GPUs) [103]. The D-ANN methods are now routinely implemented with impressive results in areas such as pattern recognition, image analysis, object detection, fault diagnosis, self-driving cars, speech recognition, natural language processing, and robotics, to name a few areas.

We proposed a deep classification network by introducing a total of three hidden layers, each with Rectified Linear Units (*ReLU*) activation function. We preferred the

ReLU activation function because it performs comparatively better than other sigmoidal activation functions in deep learning models and helps to get the best results on numerous benchmark problems in multiple domains [104]. Our fully connected deep neural network is illustrated in Figure 3, which contains an input layer with 2084 features from 3D human body mesh, 3 hidden layers with a different number of nodes such as 480, 120, and 84 nodes for the first, second, and third hidden layers, respectively, and an output layer with various possible activity classes. A simple learning function between nodes of different layers is calculated by an activation function, which can be defined as:

$$x = \Sigma_i^n w_i \cdot x_i + b \tag{3}$$

where $x_i = (x_1, x_2 \ldots . x_n)$ and $w_i = (w_1, w_2 \ldots w_n)$ are the values of the previous layer with a total of n nodes and their corresponding weights, and b is the bias value. The result of this learning function is passed through the non-linear *ReLU* activation function, which converts all negative values to zero while the remaining values are kept unchanged, as given in the following equation.

$$f(x) = max(0, x) \tag{4}$$

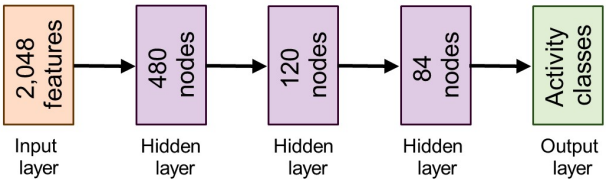

Figure 3. The architecture of a fully connected deep artificial neural network that forms Stage III of our pipeline.

The final action recognition is done at the output layer, whose nodes are equal to the number of input activity classes. The output layer uses the $SoftMax$ activation function instead of *ReLU*, which gives the maximum probability value to the most precise activity class and vice versa.

4. Experiments and Result Analysis

In this section, we evaluate the effectiveness of the proposed method using a public benchmark group activity dataset, comparing our approach with current state-of-the-art methods for the same benchmark. A detailed explanation of the dataset is given in Section 4.1. The implementation details and hardware environment are explained in Section 4.2. The comparison of the proposed method with the state-of-the-art is described in Section 4.3. Result analysis and discussion are provided in Section 4.4. Limitations of this study and expected future works are given in Section 4.5.

4.1. Dataset

We conducted experiments on a widely-adopted public group activity dataset, namely the Collective Activity Dataset (CAD) [30]. It contains 44 short video sequences with 5 different group activities such as *Crossing*, *Waiting*, *Queueing*, *Walking*, and *Talking*. The group activity label for a specific frame is determined by an activity in which the majority of people are participating. Additionally, we used an augmented collective activity dataset with two additional activity classes such as *Dancing* and *Jogging* for further evaluation of our model. As this dataset is collected from outdoor activities with consumer hand-held digital cameras, the quality of some frames is not high. Thus, we pre-processed the dataset by keeping only those frames for which at least two body meshes are created and removing the other frames. We divide the whole dataset into 80% and 20% for the model's training and validation.

To further analyze the performance of the proposed method for single-person as well as multi-person cases, we used NTU-60 RGB+D Dataset (NTU60) [105]. The original

NTU60 dataset is very big and contains 56,880 videos samples with 60 different human action classes. We therefore used a subset of the original dataset with five random classes such as $Back-Pain$, Bow, $Brush-Teeth$, $Cheer-Up$, and $Hands-Shaking$. The first four classes contain single person actions and the last class contain multi-person action. We divided the dataset into 80% and 20% for the model's training and validation.

4.2. Implementation Details

We used the ResNet-50 network as the backbone of the 3D mesh reconstruction network. All input RGB images are resized into 512×512 with zero padding and a similar aspect ratio. The size of backbone feature vectors resulting from ResNet-50 is $34 \times 128 \times 128$. These feature vectors are used to develop three head networks to find body center map, camera map, and SMPL map, each with output sizes of $1 \times 64 \times 64$, $3 \times 64 \times 64$, and $142 \times 64 \times 64$, respectively. The maximum number of persons in one frame whose 3D meshes can be created is manually set to 64. The values of the body center heatmap threshold and repulsion coefficient of CAR are set to 0.2 for each. Each person's final 3D body mesh contains $20,850$ mesh parameters that are converted to 2048 trainable features.

For the training of the D-ANN at Stage III, we use the $ADAM$ optimizer [106] with fixed hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-10}$. We adopt a mini-batch size of 1 sample with a learning rate ranging from 0.001 to 0.000001. We train and validate the network for CAD and NTU60 datasets in 80 and 50 epochs, respectively. Our method is implemented on PyTorch based deep learning framework. The inference time for a single frame is approximately 1.6 ms on a single $TITAN-RTX2080Ti$ GPU. The implementation code of our proposed method will be available online upon publication.

4.3. Comparison with the State-of-the-Art

We evaluate the proposed method on CAD dataset. The results are provided in Table 1, along with a comparison of different previous methods. Our 3D body mesh-based action recognition method outperforms all previous methods and shows the state-of-the-art results with 93.6% group activity recognition accuracy. All provided values have validation accuracy as pre-determined by the dataset authors. Meanwhile, our method slightly outperforms the recently published methods by Wu et al. [42] and Gavrilyuk et al. [65] with the validation accuracy of 90.0% and 92.8%, respectively. These outstanding results represent the effectiveness of the proposed method for group activity recognition using 3D body mesh reconstruction of multiple people scenarios.

Table 1. Collective activity dataset comparison with state-of-the-art methods for group activity recognition. **Bold** denotes the best results.

Methods	Backbone	Group Activity
Lan et al. [33]	None	79.7%
Choi and Savarese [11]	None	80.4%
Deng et al. [12]	AlexNet	81.2%
Ibrahim et al. [13]	AlexNet	81.5%
Hajimirsadeghi et al. [32]	None	83.4%
Azar et al. [41]	I3D	85.8%
Li and Chuah [16]	Inception-v3	86.1%
Shu et al. [35]	VGG16	87.2%
Qi et al. [55]	VGG16	89.1%
Ehsanpour et al. [107]	I3D	89.4%
Wu et al. [42]	Inception-v3	91.0%
Gavrilyuk et al. [65]	I3D	92.8%
Ours (3DMesh-GAR)	ResNet-50	**93.6%**

To understand the effectiveness of our method, we compare the accuracy of each individual class of the CAD dataset with the previous state-of-the-art method [65] in Table 2.

It can be seen that the proposed method outperformed the previous method to recognize the *Crossing* and *Waiting* classes with the accuracy of 88.8% and 96.8%, respectively. While, the accuracy of other classes such as *Queueing* (95.2%), *Talking* (99.7%), and *Walking* (86.7%) are also very high, their differences with [65] are negligible.

Table 2. Comparison of each activity class recognition of the CAD dataset with the previous state-of-the-art method. **Bold** denotes the best results (%).

Methods	Crossing	Queueing	Talking	Waiting	Walking
Gavrilyuk et al. [65]	83.3	**100**	**100**	96.1	**88.1**
Ours	**88.8**	95.2	99.7	**96.8**	86.7

To further analyze the efficiency of the proposed method for single-person as well as multi-person cases, we compare the accuracy of a subset of the NTU60 dataset [105] with the previous state-of-the-art method proposed by Huang et al. [108] in Table 3. It is clear that our method outperformed the previous method with an overall validation accuracy of 92.2%. Our method also achieved higher accuracy for all individual classes such as $Back-Pain$ (90.2%), *Bow* (93.7%), $Brush-Teeth$ (88.1%), and $Cheer-Up$ (95.0%), except one class $Shaking-Hands$ (91.9%). It shows that the proposed 3D body mesh-based action recognition method is the most suitable approach for general-purpose Guman Activity Recognition (GAR) in vast physical environments.

Table 3. Comparison of five different activity classes recognition of the NTU60 dataset with the previous state-of-the-art method. **Bold** denotes the best results (%).

Methods	Back-Pain	Bow	Brush-Teeth	Cheer-Up	Shaking-Hands	Overall Accuracy
Huang et al. [108]	88.7	90.9	85.9	87.7	**93.8**	89.4
Ours	**90.2**	**93.7**	**88.1**	**95.0**	91.9	**92.2**

4.4. Results Analysis

To analyze the performance of the proposed method, we present confusion matrices on the Collective Activity Dataset for multi-person activity recognition, see Figure 4. Similar to [65], our method also struggles to distinguish between the *Crossing* and *Walking* classes. The confusing ratio of the *Crossing* class with *Walking* is 9.5%, whereas, the *Walking* class with *Crossing* is 13.2%, as shown in Figure 4a. Therefore, following [15,50], we merged the *Crossing* and *Walking* classes into a single *Moving* class because there is no clear difference in physical appearance and pose of all persons of these classes and their 3D body mesh features are also similar. Figure 4b presents a confusion matrix of our method for group activity recognition with the merged *Moving* class. It is clear that our method achieves accuracy over 93%, with the least accuracy for the *Waiting* class (93.7%). The accuracy for the *Moving* class, which is a combination of the *Crossing* and *Walking* classes, is also improved up to 98.4%. The most confusion occurs for recognizing the *Queueuing* class with *Moving* (3.0%) and the *Waiting* class with *Moving* (4.6%).

For further analysis of our method, we used an augmented dataset of the Collective Activity Dataset with two additional outdoor activity classes such as *Dancing* and *Jogging*, see Figure 5a. It can be seen that the *Crossing* class achieves the lowest accuracy of 81.3% and is highly confused with the *Jogging* and *Walking* classes with a ratio of 5.2% and 6.3%, respectively. This is because of the similar physical appearance of human bodies in these classes. Similarly, the *Jogging* class is confused with the *Crossing* and *Walking* classes with a ratio of 5.5% and 3.5%, respectively. In addition, the *Walking* class is confused with the *Crossing* and *Jogging* classes with a ratio of 8.8% and 4.9%, respectively. We again merged the *Crossing* and *Walking* classes into the *Moving* class and implemented our method on the merged dataset, see Figure 5b. It is clear that the proposed method attains improved

accuracies for all classes of the merged dataset, except for *Queueing* class, which slightly deteriorated from 92.4% to 91.1%.

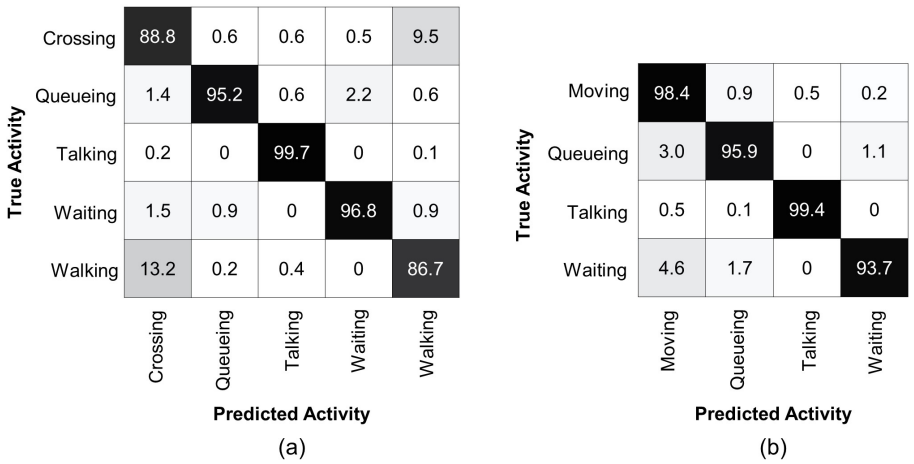

Figure 4. Confusion matrix for group activity recognition using Collective Activity Dataset. (**a**) denotes the confusion matrix for the original five classes. Most confusion occurs when distinguishing between the *Crossing* and *Walking* classes. (**b**) denotes the confusion matrix after merging the *Crossing* and *Walking* classes into a single *Moving* class. Our method achieves over 93% accuracy for each group activity class.

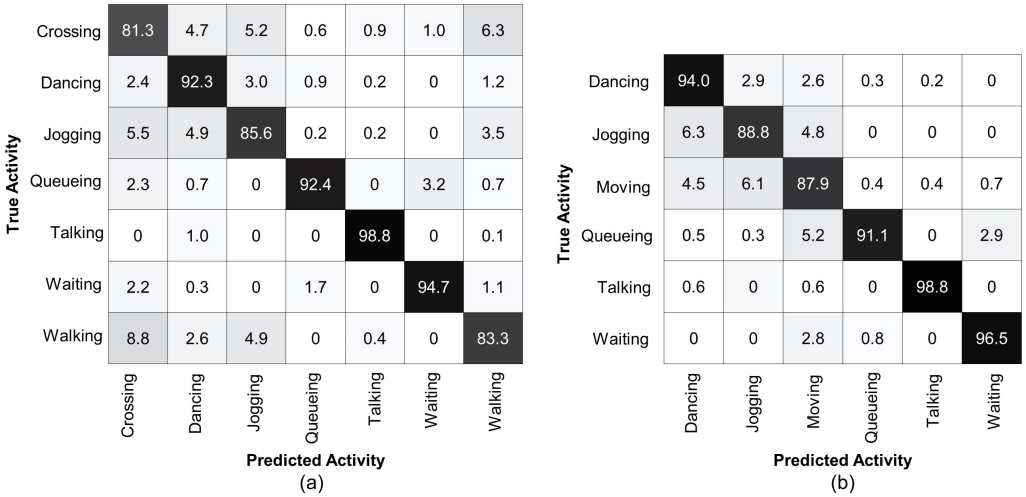

Figure 5. Confusion matrix for group activity recognition using augmented Collective Activity Dataset. (**a**) denotes the confusion matrix for the original seven classes. Most confusion occurs in distinguishing between the *Crossing* and *Walking* classes. (**b**) denotes confusion matrix after merging the *Crossing* and *Walking* classes into a single *Moving* class. Our method achieves over 87% accuracy for each group activity class.

Figure 6 illustrates the accuracy convergence of our method concerning epochs using different dataset settings of the Collective Activity Dataset. Figure 6a shows results for five classes dataset such as *Crossing*, *Queueing*, *Taking*, *Waiting*, and *Walking*. It shows the highest validation accuracy of 93.6% at the 60th epoch. Training accuracy starts at 70.0% and goes up to 98.0%. Then, training accuracy convergence becomes stable. Validation accuracy starts at 77.2% and goes up to 93.6%. It slightly went down at various epochs during the first 40 epochs and then became almost stable. Meanwhile, training loss for this dataset starts from 0.73 and decreases up to 0.06 with increasing epochs. Figure 6b shows results for the six classes dataset such as *Dancing*, *Jogging*, *Moving*, *Queueing*, *Taking*, and *Waiting*. It depicts the highest validation accuracy of 92.5% at the 72nd epoch. Training accuracy starts at 67.5% and goes up to 96.6%. Validation accuracy starts at 74.6% and goes

up to 92.4%. It slightly went down at various epochs during the first 50 epochs and then became nearly stable. The value of training loss for this dataset reduces from 0.85 to 0.10 as the epochs increase.

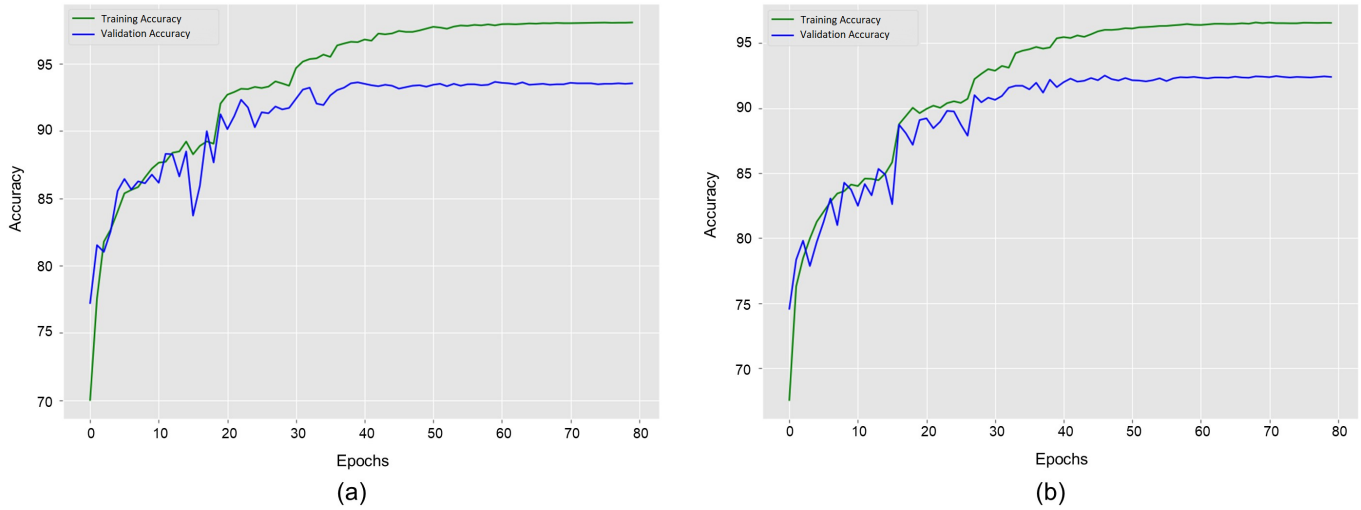

Figure 6. Model accuracy curves of group activity recognition using the training and validation datasets on the Collective Activity Dataset. (**a**) denotes the learning curves for the original dataset with the five classes. (**b**) denotes the learning curves for the augmented dataset with the six merged classes.

4.5. Limitations and Future Work

The proposed 3DMesh-GAR method is the first to implement real-time multi-person activity recognition using 3D body mesh. Nonetheless, it has some limitations, which will be addressed in this subsection.

As our 3D body mesh reconstruction approach can process multi-persons simultaneously without relying on bounding boxes for level human detection, it is naturally sensitive to person scale variations, which limits its applicability on the wild images. Moreover, our mesh reconstruction is based on 2D human representation, such as 2D Body Center Heatmap, which makes it difficult to learn the mapping function. Our method successfully reconstructs the 3D human body meshes even in the presence of person-to-person occlusion. However, it still shows the limitation of creating meshes in extremely occluded scenarios and in the presence of long-distance between persons and camera, as shown in Figure 7. All persons with missing 3D body meshes are highlighted with red circles.

Figure 7. Qualitative analysis of 3D Mesh reconstruction in highly occluded and long distance scenarios.

Because our deep network for action recognition in Stage III is trained with mesh features, our major focus in future studies will be to improve the mesh reconstruction network in Stage I. To do so, we can replace the 2D body center heatmap with 3D skeletons for multi-person detection in the frame. Thus, the efficiency of the mesh reconstruction network can be improved by collaborating the 3D joints and 3D body meshes. In addition, we will extend our work for short abnormal actions detection in a video sequence of long normal actions.

5. Conclusions

Autonomous group activity recognition has become a growing research field in the past few years and has achieved impressive progress. Activity recognition methods based on deep learning approaches are playing a vital role in GAR. Thus, this paper proposed 3DMesh-GAR, an efficient and flexible 3D human body mesh-based method for group activity recognition. It takes RGB image frames as input and product 3D body meshes using their body center heatmaps and mesh parameter maps. The 3D mesh features are very simple and easy to train with a linear neural network. A deep artificial neural network architecture has been developed and optimized to learn and recognize group activities from the proposed description in an end-to-end manner. We evaluate the proposed method on a benchmark group activity recognition dataset and establish a new state-of-the-art performance. Various experiments and results analysis show how an intuitive and simple 3D mesh-based method regarding information representation can be successful for high-level feature extraction from a complex dataset.

Author Contributions: Conceptualization, M.S. and S.B.; formal analysis, J.C., C.L. and S.L.; funding acquisition, S.B.; methodology, M.S. and S.B.; project administration, S.B.; software, M.S. and D.K.; supervision, S.B.; validation, M.S. and S.B.; visualization, D.K., J.C., C.L. and S.L.; writing—original draft, M.S.; writing—review and editing, M.S. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. 2021R1F1A1047920) and Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01778 Development of human image synthesis and discrimination technology below the perceptual threshold, No. 2020-0-01336 Artificial Intelligence Graduate School Program(UNIST)). This work was also supported by the Settlement Research Fund (1.200099.01) of UNIST (Ulsan National Institute of Science and Technology).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

GAR	Group Activity Recognition
3DMesh-GAR	3D Mesh-based Group Activity Recognition
GCNs	Graph Convolutional Networks
RNNs	Recurrent Neural Networks
CNNs	Convolutional Neural Networks
DNNs	Deep Neural Networks
D-ANN	Deep Artificial Neural Network
GNNs	Graph Neural Networks
LSTM	Long Short-Term Memory

NLP	Natural Language Processing
SVMs	Support Vectors Machines
ARG	Actor Relation Graph
SMPL	Skinned Multi-Person Linear model
R-CNN	Region-based Convolutional Neural Networks
3DMPPE	3D Multi-Person Pose Estimation
LCR-Net++	Localization Classification Regression Network
LoCO	Learning on Compressed Output
GPUs	Graphics Processing Units
ReLU	Rectified Linear Units

References

1. Demrozi, F.; Pravadelli, G.; Bihorac, A.; Rashidi, P. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE Access* **2020**, *8*, 210816–210836. [CrossRef] [PubMed]
2. Azar, S.M.; Atigh, M.G.; Nickabadi, A. A multi-stream convolutional neural network framework for group activity recognition. *arXiv* **2018**, arXiv:1812.10328.
3. Ren, B.; Liu, M.; Ding, R.; Liu, H. A survey on 3d skeleton-based action recognition using learning method. *arXiv* **2020**, arXiv:2002.05907.
4. Wang, Y.; Cang, S.; Yu, H. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Syst. Appl.* **2019**, *137*, 167–190. [CrossRef]
5. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
6. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018, pp. 6450–6459.
7. Lin, J.; Gan, C.; Han, S. Temporal shift module for efficient video understanding. In *Proceedings of the CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 7082–7092.
8. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv* **2018**, arXiv:1801.07455
9. Xu, C.; Govindarajan, L.N.; Zhang, Y.; Cheng, L. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *Int. J. Comput. Vis.* **2017**, *123*, 454–478. [CrossRef]
10. Xiao, F.; Lee, Y.J.; Grauman, K.; Malik, J.; Feichtenhofer, C. Audiovisual slowfast networks for video recognition. *arXiv* **2020**, arXiv:2001.08740.
11. Choi, W.; Savarese, S. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the European Conference on Computer Vision*; Springer: Heidelberg, Germany, 2012; pp. 215–230.
12. Deng, Z.; Vahdat, A.; Hu, H.; Mori, G. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
13. Ibrahim, M.S.; Muralidharan, S.; Deng, Z.; Vahdat, A.; Mori, G. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
14. Ramanathan, V.; Huang, J.; Abu-El-Haija, S.; Gorban, A.; Murphy, K.; Fei-Fei, L. Detecting events and key actors in multi-person videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
15. Wang, M.; Ni, B.; Yang, X. Recurrent modeling of interaction context for collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
16. Li, X.; Choo Chuah, M. Sbgar: Semantics based group activity recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
17. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*; Springer: Heidelberg, Germany, 2016; pp. 20–36.
18. Huang, Y.; Bogo, F.; Lassner, C.; Kanazawa, A.; Gehler, P.V.; Romero, J.; Akhter, I.; Black, M.J. Towards accurate marker-less human shape and pose estimation over time. In *Proceedings of the 2017 International Conference on 3D Vision (3DV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 421–430.
19. Fang, H.S.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.C. Learning pose grammar to encode human body configuration for 3d pose estimation. *arXiv* **2017**, arXiv:1710.06513
20. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
21. Sun, Y.; Bao, Q.; Liu, W.; Fu, Y.; Black, M.J.; Mei, T. Monocular, One-stage, Regression of Multiple 3D People. *arXiv* **2020**, arXiv:2008.12272.

22. Xiang, D.; Joo, H.; Sheikh, Y. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
23. Duan, H.; Zhao, Y.; Chen, K.; Shao, D.; Lin, D.; Dai, B. Revisiting Skeleton-based Action Recognition. *arXiv* **2021**, arXiv:2104.13586.
24. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Elgharib, M.; Fua, P.; Seidel, H.P.; Rhodin, H.; Pons-Moll, G.; Theobalt, C. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Trans. Graph. (TOG)* **2020**, *39*. [CrossRef]
25. Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; Lu, C. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 3383–3393.
26. Zhang, J.; Yu, D.; Liew, J.H.; Nie, X.; Feng, J. Body meshes as points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
27. Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; Gao, S. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
28. Varol, G.; Laptev, I.; Schmid, C.; Zisserman, A. Synthetic humans for action recognition from unseen viewpoints. *Int. J. Comput. Vis.* **2021**, *129*, 2264–2287. [CrossRef]
29. Mir, A.; Alldieck, T.; Pons-Moll, G. Learning to transfer texture from clothing images to 3d humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
30. Choi, W.; Shahid, K.; Savarese, S. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*; IEEE: Piscataway, NJ, USA, 2009; pp. 1282–1289.
31. Choi, W.; Savarese, S. Understanding collective activitiesof people from videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1242–1257. [CrossRef]
32. Hajimirsadeghi, H.; Yan, W.; Vahdat, A.; Mori, G. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
33. Lan, T.; Wang, Y.; Yang, W.; Robinovitch, S.N.; Mori, G. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1549–1562. [CrossRef]
34. Amer, M.R.; Lei, P.; Todorovic, S. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 572–585.
35. Shu, T.; Todorovic, S.; Zhu, S.C. Cern: confidence-energy recurrent network for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
36. Bagautdinov, T.; Alahi, A.; Fleuret, F.; Fua, P.; Savarese, S. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4315–4324.
37. Ibrahim, M.S.; Mori, G. Hierarchical relational networks for group activity recognition and retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 721–736.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385
39. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
40. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv* **2016**, arXiv:1602.07261
41. Azar, S.M.; Atigh, M.G.; Nickabadi, A.; Alahi, A. Convolutional relational machine for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
42. Wu, J.; Wang, L.; Wang, L.; Guo, J.; Wu, G. Learning actor relation graphs for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
43. Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; Wu, F. Multi-modality cross attention network for image and sentence matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
44. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2016; pp. 3464–3468.
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2017.
46. Sendo, K.; Ukita, N. Heatmapping of people involved in group activities. In *Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
47. Choi, W.; Shahid, K.; Savarese, S. Learning context for collective activity recognition. In *Proceedings of the CVPR 2011*; IEEE: Piscataway, NJ, USA, 2011; pp. 3273–3280.
48. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In *Proceedings of the 2017 IEEE international Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2017; pp. 3645–3649.
49. Musgrave, K.; Belongie, S.; Lim, S.N. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*; Springer: Heidelberg, Germany, 2020; pp. 681–699.

50. Hu, G.; Cui, B.; He, Y.; Yu, S. Progressive relation learning for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
51. Sun, L.; Ai, H.; Lao, S. Localizing activity groups in videos. *Comput. Vis. Image Underst.* **2016**, *144*, 144–154. [CrossRef]
52. Deng, Z.; Zhai, M.; Chen, L.; Liu, Y.; Muralidharan, S.; Roshtkhari, M.J.; Mori, G. Deep structured models for group activity recognition. *arXiv* **2015**, arXiv:1506.04191.
53. Wang, L.; Zang, J.; Zhang, Q.; Niu, Z.; Hua, G.; Zheng, N. Action recognition by an attention-aware temporal weighted convolutional neural network. *Sensors* **2018**, *18*, 1979. [CrossRef]
54. Li, W.; Chang, M.C.; Lyu, S. Who did what at where and when: simultaneous multi-person tracking and activity recognition. *arXiv* **2018**, arXiv:1807.01253.
55. Qi, M.; Qin, J.; Li, A.; Wang, Y.; Luo, J.; Van Gool, L. stagnet: An attentive semantic rnn for group activity recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
56. Long, X.; Gan, C.; De Melo, G.; Wu, J.; Liu, X.; Wen, S. Attention clusters: Purely attention based local feature integration for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
57. Wang, X.; Gupta, A. Videos as space-time region graphs. *arXiv* **2018**, arXiv:1806.01810
58. Li, Z.; Gavrilyuk, K.; Gavves, E.; Jain, M.; Snoek, C.G. Videolstm convolves, attends and flows for action recognition. *Comput. Vis. Image Underst.* **2018**, *166*, 41–50. [CrossRef]
59. Tang, Y.; Zhang, P.; Hu, J.F.; Zheng, W.S. Latent embeddings for collective activity recognition. In *Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
60. Lu, L.; Di, H.; Lu, Y.; Zhang, L.; Wang, S. Spatio-temporal attention mechanisms based model for collective activity recognition. *Signal Process. Image Commun.* **2019**, *74*, 162–174. [CrossRef]
61. Baradel, F.; Wolf, C.; Mille, J. Human activity recognition with pose-driven attention to rgb. In Proceedings of the BMVC 2018-29th British Machine Vision Conference, Snowmass, CO, USA, 1–5 March 2020.
62. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank pooling for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 773–787. [CrossRef] [PubMed]
63. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
64. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
65. Gavrilyuk, K.; Sanford, R.; Javan, M.; Snoek, C.G. Actor-transformers for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
66. Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; Schmid, C. Bodynet: Volumetric inference of 3d human body shapes. *arXiv* **2018**, arXiv:1804.04875
67. Gabeur, V.; Franco, J.S.; Martin, X.; Schmid, C.; Rogez, G. Moulding humans: Non-parametric 3d human shape estimation from single images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
68. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
69. Tome, D.; Russell, C.; Agapito, L. Lifting from the deep: Convolutional 3d pose estimation from a single image. In Proceedings of the EEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
70. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal depth supervision for 3d human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
71. Zhang, J.; Nie, X.; Feng, J. Inference stage optimization for cross-scenario 3d human pose estimation. *arXiv* **2020**, arXiv:2007.02054.
72. Gong, K.; Zhang, J.; Feng, J. PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
73. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (TOG)* **2015**, *34*, 1–16. [CrossRef]
74. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent Advances in Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *arXiv* **2021**, arXiv:2104.11536.
75. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 561–578.
76. Kolotouros, N.; Pavlakos, G.; Daniilidis, K. Convolutional mesh regression for single-image human shape reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
77. Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; Schiele, B. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *Proceedings of the 2018 International Conference on 3D Vision (3DV)*; IEEE: Piscataway, NJ, USA, 2018; pp. 484–494.
78. Tung, H.Y.F.; Tung, H.W.; Yumer, E.; Fragkiadaki, K. Self-supervised learning of motion capture. *arXiv* **2017**, arXiv:1712.01337.
79. Arnab, A.; Doersch, C.; Zisserman, A. Exploiting temporal context for 3D human pose estimation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.

80. Kocabas, M.; Athanasiou, N.; Black, M.J. Vibe: Video inference for human body pose and shape estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
81. Guler, R.A.; Kokkinos, I. Holopose: Holistic 3d human reconstruction in-the-wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
82. 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, 18–23 June 2018.
83. Tekin, B.; Katircioglu, I.; Salzmann, M.; Lepetit, V.; Fua, P. Structured prediction of 3d human pose with deep neural networks. *arXiv* **2016**, arXiv:1605.05180.
84. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
85. Sun, X.; Shang, J.; Liang, S.; Wei, Y. Compositional human pose regression. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
86. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
87. Tekin, B.; Márquez-Neila, P.; Salzmann, M.; Fua, P. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
88. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proceedings of the 2017 International Conference on 3D Vision (3DV)*; IEEE: Piscataway, NJ, USA, 2017.
89. Cha, J.; Saqlain, M.; Lee, C.; Lee, S.; Lee, S.; Kim. D.; Park, W.-H.; Baek, S. Towards single 2D image-level self-supervision for 3D human pose and shape estimation. *Appl. Sci.* **2021**, *11*, 9724. [CrossRef]
90. Zanfir, A.; Marinoiu, E.; Sminchisescu, C. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
91. Jiang, W.; Kolotouros, N.; Pavlakos, G.; Zhou, X.; Daniilidis, K. Coherent reconstruction of multiple humans from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
92. Dabral, R.; Mundhada, A.; Kusupati, U.; Afaque, S.; Sharma, A.; Jain, A. Learning 3d human pose from structure and motion. *arXiv* **2017**, arXiv:1711.09250
93. 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, Honolulu, HI, USA, 21–26 July 2017.
94. Kanazawa, A.; Zhang, J.Y.; Felsen, P.; Malik, J. Learning 3d human dynamics from video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
95. 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**, *28*, 91–99. [CrossRef] [PubMed]
96. Moon, G.; Chang, J.Y.; Lee, K.M. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
97. Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; Tan, R.T. Occlusion-aware networks for 3d human pose estimation in video. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
98. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 7753–7762.
99. Li, J.; Wang, C.; Liu, W.; Qian, C.; Lu, C. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. *arXiv* **2020**, arXiv:2008.00206.
100. Lin, J.; Lee, G.H. Hdnet: Human depth estimation for multi-person camera-space localization. In *Proceedings of the European Conference on Computer Vision*; Springer: Heidelberg, Germany, 2020; pp. 633–648.
101. Fabbri, M.; Lanzi, F.; Calderara, S.; Alletto, S.; Cucchiara, R. Compressed volumetric heatmaps for multi-person 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
102. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
103. Imtiaz, S.I.; ur Rehman, S.; Javed, A.R.; Jalil, Z.; Liu, X.; Alnumay, W.S. DeepAMD: Detection and identification of Android malware using high-efficient Deep Artificial Neural Network. *Future Gener. Comput. Syst.* **2021**, *115*, 844–856. [CrossRef]
104. Dahl, G.E.; Sainath, T.N.; Hinton, G.E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
105. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
106. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

107. Ehsanpour, M.; Abedin, A.; Saleh, F.; Shi, J.; Reid, I.; Rezatofighi, H. Joint learning of social groups, individuals action and sub-group activities in videos. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IX 16; Springe: Heidelberg, Germany, 2020; pp. 177–195.
108. Huang, L.; Huang, Y.; Ouyang, W.; Wang, L. Part-level graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press: Palo Alto, CA, USA, 2020.