applied sciences

MDPI

Article

Producing Synthetic Dataset for Human Fall Detection in AR/VR Environments

Denis Zherdev [1,2], Larisa Zherdeva [2,*], Sergey Agapov [3], Anton Sapozhnikov [3], Artem Nikonorov [1,2] and Sergej Chaplygin [3]

[1] Image Processing Systems Institute of RAS—Branch of the FSRC "Crystallography and Photonics" RAS, 443001 Samara, Russia; denis.zherdev.91@mail.ru (D.Z.); artniko@gmail.com (A.N.)
[2] Samara National Research University, 443086 Samara, Russia
[3] Samara State Medical University, 443099 Samara, Russia; s.n.agapov@samsmu.ru (S.A.); a.a.sapozhnikov@samsmu.ru (A.S.); info-iir@samsmu.ru (S.C.)
* Correspondence: lara.zherdeva.taskina@gmail.com

Abstract: Human poses and the behaviour estimation for different activities in (virtual reality/augmented reality) VR/AR could have numerous beneficial applications. Human fall monitoring is especially important for elderly people and for non-typical activities with VR/AR applications. There are a lot of different approaches to improving the fidelity of fall monitoring systems through the use of novel sensors and deep learning architectures; however, there is still a lack of detail and diverse datasets for training deep learning fall detectors using monocular images. The issues with synthetic data generation based on digital human simulation were implemented and examined using the Unreal Engine. The proposed pipeline provides automatic "playback" of various scenarios for digital human behaviour simulation, and the result of a proposed modular pipeline for synthetic data generation of digital human interaction with the 3D environments is demonstrated in this paper. We used the generated synthetic data to train the Mask R-CNN-based segmentation of the falling person interaction area. It is shown that, by training the model with simulation data, it is possible to recognize a falling person with an accuracy of 97.6% and classify the type of person's interaction impact. The proposed approach also allows for covering a variety of scenarios that can have a positive effect at a deep learning training stage in other human action estimation tasks in an VR/AR environment.

Keywords: modelling and simulation; depth maps; segmentation; human fall; CNN; machine learning

Citation: Zherdev, D.; Zherdeva, L.; Agapov, S.; Sapozhnikov, A.; Nikonorov, A.; Chaplygin, S. Producing Synthetic Dataset for Human Fall Detection in AR/VR Environments. *Appl. Sci.* **2021**, *11*, 11938. https://doi.org/10.3390/app112411938

Academic Editor: João M. F. Rodrigues

Received: 4 November 2021
Accepted: 13 December 2021
Published: 15 December 2021

1. Introduction

With the rapid progress of deep learning models gathering the necessary amount of training, data is a challenging task [1]. Typically, synthetic data is used in the neural networks training process to reduce the costs of collecting big diversity of the dataset and solving domain-adaptation problems in visual tasks [2]. In this case, developing and improving three-dimensional modelling and rendering software aims to achieve synthetic data modelling for solving non-standard problems in network training.

The existing synthetic dataset samples cover an impressive amount of applications in recognition tasks, such as autonomous robots navigation [3–5] and unmanned aerial vehicles [6–9]. There can be large-scale datasets of interiors with any number of furniture sets, or separate datasets with objects of the environment. On the other hand, there can be various samples of residential or non-residential environments [10,11] simulated under different lighting conditions [12,13]. There are also large-scale urban datasets, including modelled natural areas and landscapes [14,15], and it is shown that such datasets have a good effect on convolutional neural network (CNN) training. The dataset modelling task has many variants of realization that combine photogrammetry methods and computer visualization engines [6,16].

Synthetic training datasets can be created using human body modelling with a certain approximation that has been confirmed more than once to be useful in solving standard problems in recognizing the form and actions of a person [17–19]. For general cases of human posture assessment, datasets [20,21] collect an impressive amount of annotated data. However, in particular problems of recognizing human actions, existing sets may not cover all scenarios [22], and their variability may not be full enough [23]. The person action annotation task with the environment is one of the resource-intensive tasks [24], while this sort of data is being increasingly applied in health care and tracking tasks, e.g., automatic fall detection systems [25,26], with subsequent analysis of their causes to improve the quality and safety for people.

Computer visualization and rendering provides the ability to generate a large amount of data with perfect labels [27], and the creation of a pipeline for generating a synthetic dataset based on human actions and interactions with the environment is an important and urgent task. The main emphasis in this research focuses on the usage of a simulation environment for applied tasks of human action predictions with the environment, including for further use in rehabilitation based on virtual reality. In the study, there is a high interest in applying the approach in the healthcare industry, particularly in regard to analysing the causes of falls. Because of the study, the life quality of people with disabilities [27] can be improved.

The synthetic data generation pipeline [28] describing the consequences of a fall has serious limitations: evaluating what part of the body had been hit; usage of the dataset in the task of detecting the consequences of a fall; intentionality of the actions of persons in the reconstruction process of the fall. The dataset obtained using our approach has several differences from the existing synthetic options [20,28–30]. For the more realistically based simulation, the digital human behaviour algorithm and inverse kinematics methodology were applied in the research. In [20,22,29], authors performed the human fall data generation without considering the inverse kinematics human model aspects, which is in contrast to our approach. In the existing works [28,29], modelling scenarios are not sufficiently considered, and in the dataset [29] there is the dynamics of a person falling from an initial static state. We give a notice to produce the hit masks that indicate the places which have been hit, and it can be used in many applications, such as to analyse the causes of falls in the healthcare industry, in VR controllers design, etc. In [30], authors use a hit mask in some sense with the purpose of pressure assessment of bodies at rest on a mattress. Moreover, our dataset includes the diversity of clothes for simulated digital humans. Our research has put forward the following contributions:

(1) We implemented a physical model of digital humans to simulate the fall of a person and their interaction with the environment with inverse kinematics for producing various fall dynamics;
(2) We took into account the coordinates of impact with the object of interaction in the process of registering a digital human fall to improve the segmentation ability of the hit mask;
(3) We implemented the integration of digital human behaviour simulation for automatic "playing" of various scenarios of interactions and falls in a 3D scene automatically;
(4) We apply the deep learning approach to examine the ability for training with synthetic data to recognize the real datasets.

For the validation of the proposed synthetic dataset in recognition results, two classes were taken into account (fall and not fall). However, at the training stage, we have three categories that were labelled automatically using our generation pipeline: wall, floor and hit mask, obtained as an area of floor and body touch.

The manuscript is structured as follows. In Section 2, we consider modern approaches to the definition of falling people. In Section 3, we present a modular approach to the problem of generating and collecting synthetic data from digital humans to study a person's fall and the consequences of a fall using the example of interaction with the environment. In Section 4, we present the implementation of the pipeline, the main results and the char-

acteristics of the generated dataset, discussing the applicability of the results and further development of the project for the application of virtual reality. Finally, the conclusions and general ability for the pipeline applications are given in Section 5.

2. Related Works

2.1. Human Fall Detection

The problem of fall detection has been widely studied by the authors in [27,31] using neural networks. Significant experiments have been carried out to study human falls in simulated laboratory conditions [32–34]. However, not all datasets are publicly available (for privacy reasons). The fall assessment methods presented in these studies have varied depending on the estimation-based data.

In [35], the authors used human motion segmentation with background subtraction. The main measure was a significant change in visual information between subsequent image frames. However, differences in the background and the presence of occlusion objects, different camera viewpoints, and variability of the person's appearance led to poor generalization of methods applied to the real world [27]. The authors of work [36] made a great contribution to the study of 3D information about a person falling from several cameras and sensors, and the analysis was carried out using depth maps. As noted by the authors in the work [37], systems with multiple cameras generated more accurate fall detection results. However, these methods have additional limitations, such as data synchronization and depth detection for 3D sensors, and it is not useful for cases in which cheapness, quick responses and simplified usage are required. Fall detection with a monocular camera has an advantage against other methods due to the absence of the above restrictions.

One of the wider and open-source datasets that can be taken into account is [38], which concerns a human falling. Solving the detection task can be complex due to variable lighting and exposure conditions using such datasets. In most cases, these datasets are not fully able to provide fall detection from one camera because of the high occlusion coefficient when data from different cameras is analysed [39]. Comparable research to our approach was presented in [28], where the authors demonstrated fall recognition derived from synthetic data samples based on the alignment of MoCap poses and human models. They generated the values of the skeletal joints, the segmentation mask and the "fall-no fall" label in the created dataset. As a result, the authors [28] also introduced a deep learning framework for fall detection in complex non-obvious real-world conditions. However, the detection result does not provide the interaction masks with the environment during a fall.

The presented approaches and datasets have two common disadvantages: (1) this is the intentional actions of persons in the process of reconstructing a fall (although the researchers note that they studied data on an unexpected and unintentional fall); (2) the presented variations of falls are often limited and difficult to use in the task of detecting the consequences of one. For example, it is difficult to evaluate what part of the body had been hit and the object of hit. In such tasks, human detections inevitably make mistakes in predictions and pixel segmentations [28].

2.2. Existed Training Sets

As mentioned above in Section 2.1, recognition of human actions often faces problems associated with changing the camera's point of view and external lighting, as well as the shape of the person's body and clothes [40]. At the same time, the camera orientation change makes a significant contribution to the recognition, and the same action can lead to different results. In addition, annotations of body joints that are heavily occluded can have many errors [28]. The high cost of annotating large-scale data has prompted researchers to look for efficient ways to synthesize large data sets for the reliable recognition of actions [22,41]. The main advantage of synthetic data is complete control over the virtual environment and the ability to generate datasets with high variance [20,22].

The general idea of the work in [29], in which authors developed and presented a framework synthesizing training data for synthetic 3D people models, inspired us to pursue this research. The dataset developed in [29] shows Human Pose Models that represent RGB and depth images of human poses independent of external parameters such as clothing, lighting and camera viewpoints. The authors in [29] take the synthetic data approach, as it promises a wide range of actions due to its diversity and scale of variation. Therefore, in a virtual environment, a researcher can fix as many variations of the same action as required to solve a specific problem, while such an implementation in the real world would require large costs [14,42]. It was experimentally demonstrated in [22,29] that the method based on synthetic data is superior to existing modern methods of recognizing actions in conventional RGB and RGB-D videos.

The authors in [20] also used the synthetic data generation method for the problem of body segmentation and depth estimation. Another dataset, Human3.6M [43], presents a realistic rendering of people in mixed reality. In [20,22], it was also shown that CNN trained using the synthetic dataset allows one to accurately estimate 3D depth and segment the human part in real-life images. However, results in [29] noted that the realism of synthetic data directly affects generalization in regard to real data. As a solution in [29], the authors proposed several methods of adapting the subject area. As noted in [22], there are two main approaches for creating a dataset, namely rendering only a synthetic dataset and combining synthetic and real training data. During the analysis, a significant advantage of mixed datasets based on synthetic and real data was revealed due to the control of model retraining using just synthetic data features. However, in a real experiment it is difficult to obtain truly satisfying data regarding hit maps of human falls because there is so much outlay here. For example, a pavilion with tactile surfaces with feedback was needed to register impacts with high time costs and the chance of the subject being injured. The authors in [29] noticed that the developed dataset of synthetic people images was created using 3D models and lighting variations. At the same time, 2D backgrounds were used as the environment, on which the lighting was not transferred, and the transition between the model and the background was sharp and unrealistic, leading to the additional usage of methods that "improve" the synthetic data.

Thus, the use of synthetic data based on the combination of physical modelling and digital humans can improve the quality and variability of the dataset, and the approach itself allows for the generation of really large-scale data.

3. Proposed Digital Human Falling Dataset Generation Pipeline

It is necessary to have an environment that combines both a physics and animation engine, as well as a realistic rendering in real-time, for the successful implementation of the pipeline. Moreover, such an environment has to provide modularity and access to any component within that pipeline. There are at present many modern engines for modelling and creating your own synthetic datasets, and a detailed comparison can be found in [44]. Therefore, we chose a modern engine, Unreal Engine 4 (ue4), implemented on c++ [45] as a platform for simulation. The ue4 combines all the architectures and tools we need, and it is also capable of providing high performance 3D simulations.

The proposed pipeline in the manuscript of synthetic data generation consists of three main modules:

- Three-dimensional scene generation for various textures, meshes, skeleton physics, etc. (Sections 3.1–3.3);
- Various behaviour simulation of a digital human (Section 3.4);
- Masks of digital human hit registration with a 3D scene environment (Section 3.5).

3.1. Digital Human Construction

We used 3D scanned models of real people from the RenderPeople [46] dataset as a base three-dimensional human model. Thus, 4 models (2 male and 2 female) with 17–23 K triangles were included in the sample. To control the colour diversity of the digital human

and his clothes, materials for rendering models were implemented. We used skin tone variety palettes and texture patterns for clothing. In Figure 1a, the examples of digital humans are shown.

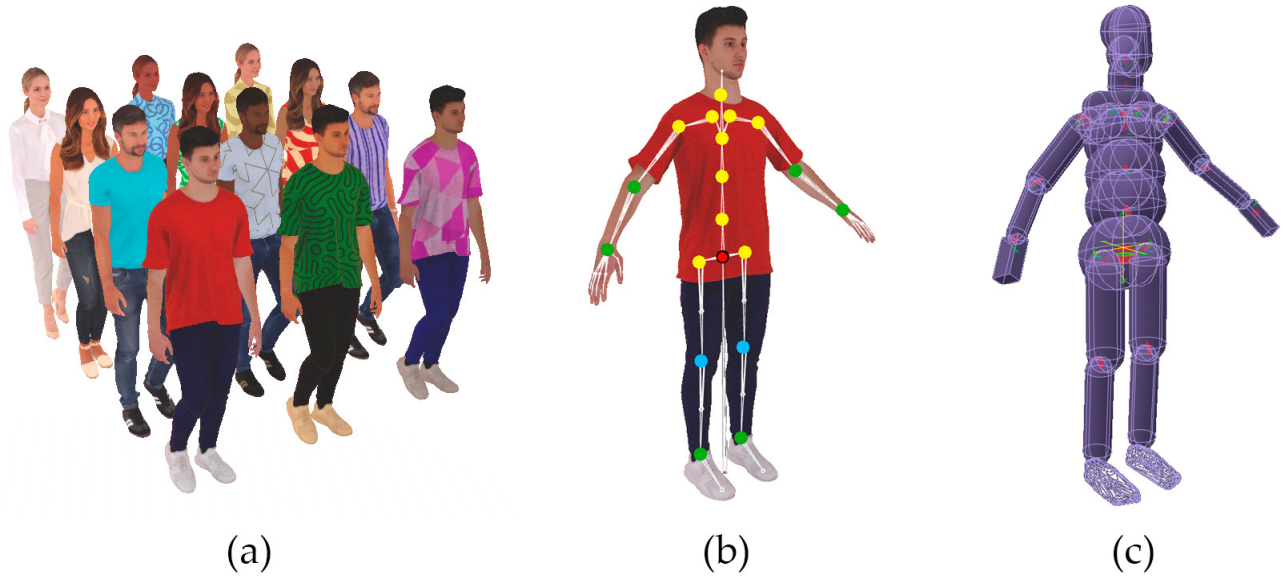


Figure 1. Digital human examples used in experiments (**a**); digital human skeleton hierarchy (**b**); physical model of digital human (**c**).

We executed the cloth texture generation pipeline. It contains the following steps: (1) select 5 typical patterns for the main background of clothing (including one-color), (2) transform and rotate patterns, (3) project the pattern over the entire model according to UV, (4) change the colour of patterns across the entire spectrum of the standard rgb palette.

3.2. Physical Modeling

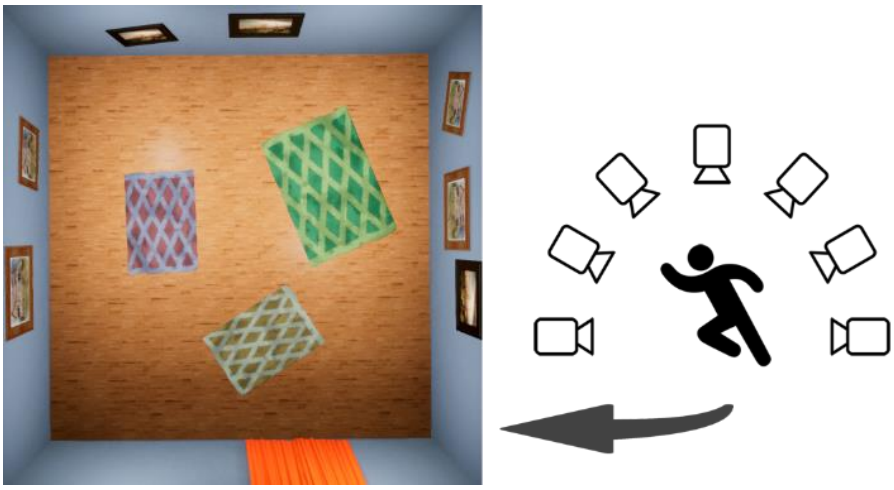
We used a skeleton containing the 22 main bones of the body and legs and 30 bones of the fingers, which has a standard tree structure. Figure 1b,c shows the anatomy of 3D scanned people models corresponding to the hierarchy of skeletal bones. It is outlined that the human skeleton has the freedom degrees (DOF) for each joint—red 6 DOF, yellow 3 DOF, green 2 DOF, blue 1 DOF. The physical model of a digital human is based on a hinge system of primitives with corresponding constraints for each joint.

As shown in Figure 1b, certain bones have their own DOF. Below in the experiments, we will discuss the process of a person's fall as closely as possible. Such imitations are possible using a system based on physical capsules corresponding to the skeleton hierarchy [47]. Figure 1c shows a schematic diagram of the physical model. It is a person modelled as an articulated rigid body system consisting of primitives such as capsules. Each bone is assigned a capsule with the appropriate weight and simulation parameter value.

The physical model for simulating the rigid body dynamics of a human model based on capsules was used. The idea is based on the movement simulation of the rigid body, which takes into account the applied forces and moments. However, we used local constraint, where the obtainable angular transformation is limited at each stage by the upper and lower limits [48]. This minimizes and avoids unrealistic joint displacements during physical simulation.

3.3. Background and Rendering

A physical model of a person is placed in a 3D environment in which the human model interacts with a 3D interior to obtain simulation data (Figure 2a). It is necessary to notice that the term "interaction" of a person implies the activation of a digital human behaviour scenario, whose algorithm will be described in Section 3.4.



(a) **(b)**

Figure 2. Visual representation of simulation room (**a**) and camera system for digital human (**b**).

In the experiment, we used a 3D room of fixed sizes. The weight and length were 8 m and the height is 3 m. On the walls were randomly placed objects of interior paintings, and on the floor were rugs. The rendering material type for each type of environment model (floor, walls, etc.) was assigned. Moreover, the following variability of parameters within the material were implemented: texture scale, texture blending colour, normal coefficient and roughness.

The simulation environment has the significant ability to adjust the light while recording visual changes by moving the camera to any angle and any desired point of view. For experiments, we arranged and configured three types of light sources (basic skylight, directional source and 9 light windows that simulate office lighting sources).

In Figure 2b, it is shown that the layout of cameras for filming during the experiments was a contained system of 16 units for digital human registration. This system issued images with fall registration at different angles simulated by different cameras in a 3D environment.

We executed the next scheme to obtain simulation data for images generated by the described system. We implemented a separate object that follows a digital human and consists of 16 virtual cameras, located in the hemisphere at the same distance and where the centre of interest is a person model. The field of view of each camera was specified and equalled 90 degrees. Recording and registration from virtual cameras was carried out synchronously.

In general, the method allows one to effectively include large variations of any number of virtual cameras and many factors that affect the final result when generating data.

3.4. Digital Human Behavior Simulation

Initially, our simulation approach was to autonomously observe a virtual person with actions and "play" various scenarios of his interaction with the 3D environment. Thus, a behaviour system was implemented for our digital human.

The implementation of human behaviour in the experiment was carried out using the behaviour tree (BT) [49]. Such algorithm representation sets a certain digital human action command as a leaf of BT. A node in the BT either encapsulates the action to be performed or acts as a component of the control flow that directs the traversal of the BT.

We examined the followed implementation of the behaviour in the experiment. Figure 3 presents the scheme of decision making by a digital human with the description of BT component types.

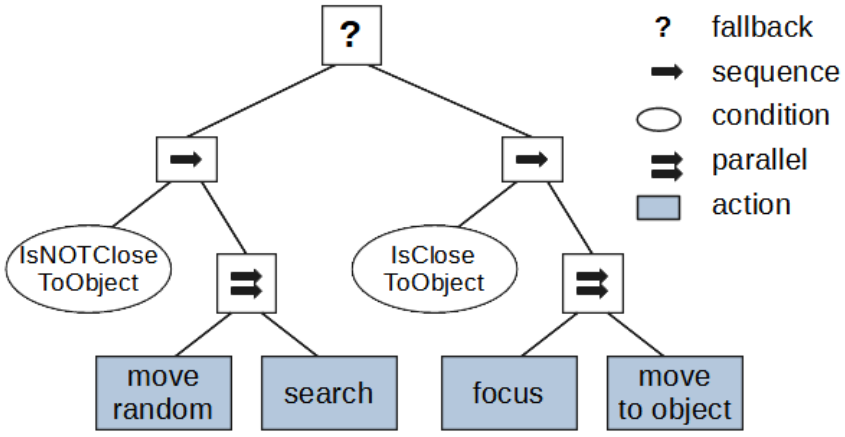


Figure 3. Behaviour tree for the digital human.

The BT consists of one root and two sequences, two parallel, two condition and four activity nodes. Firstly, at the BT start point, the first branch on the left is the "explore the room" sequence. Using such an algorithm, the digital human can choose any point of free area, and the action "move" occurs with the playback of the corresponding animation. Secondly, if a digital human in the process of moving comes close enough to the object it can interact with, then the second branch in the BT is activated. In that case, the digital human focuses the attention on the object and moves closer. Upon completion of the action, the BT returns success. Finally, the digital human continues the loop by walking around the room and the object of interest during the searching process.

The slides of the registration process for synthetic data simulation are shown in Figure 4. In the process of digital human random motion at some arbitrary moment, the action is triggered at t_start time, which leads to a "fainting" after a t_fall time. From the moment the trigger t_start is activated, the automatic system begins a comprehensive data collection. At t_fall, the physical model of a person is activated, which corresponds to the simulation of a sudden loss of consciousness in a person. The animation contribution to the movement of the joints tends to go to zero, and only the simulation of the capsule model remains—the person continues to fall in the direction of inertia of the last skeletal pose until the moment t_end, when the physical model is fully balanced. In addition, we can capture the digital human during his motion in the behaviour process, which provides non-fall images.

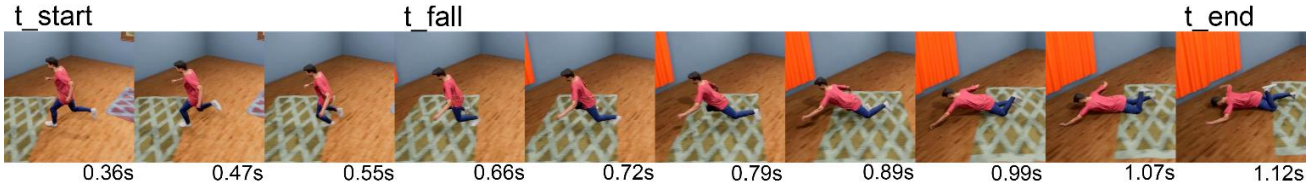


Figure 4. Single-frame registration digital human from the moment the fall simulation trigger is activated.

3.5. Hit Masks and Visualisation

For further comprehensive research of deep learning, the data was presented in several versions. A dataset is synchronously captured from each camera in each frame with standard 60 Hz update frequency. Therefore, this includes the main rendering maps: rgb, normal, depth and object segmentation. Figure 5 shows a sample obtained in a fixed simulation frame.

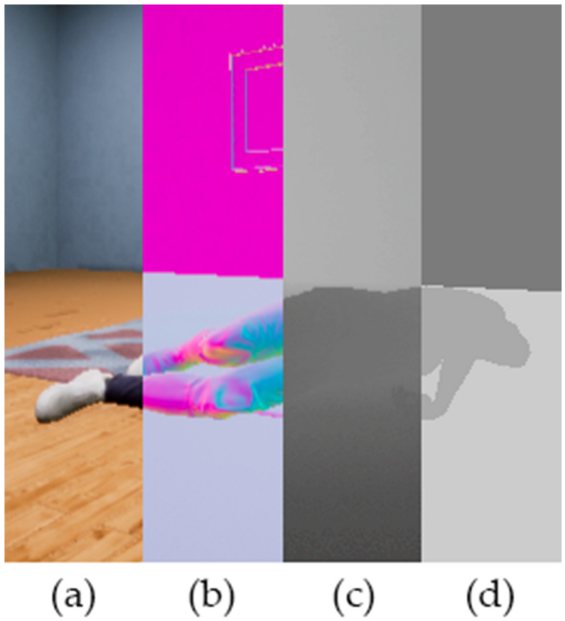


Figure 5. Simulation results: rgb (**a**), normal (**b**), depth (**c**), segmentation (**d**) rendering maps.

Simulation of digital human interaction in a virtual scene has significant advantages. We can obtain annotated data using such an approach with minimal costs, different types, accuracy and gradation. Thus, in Figure 5d, it is shown that we compute accurate segmentation maps of objects: floor, wall and the human body. A fact of using additional controlled parameters such as normal and depth maps (Figure 5b,c) makes it possible to comprehensively assess the orientation of digital human body parts in the task of recognizing the consequences of a fall. Moreover, we presented an advanced dataset from human interaction maps with the environment at the time of fall impact. We were interested in the possibility of generating and training a collision recognition model. Therefore, in the research, we register human collisions with the floor and use of a simplified registration system for which several stage outputs are presented below in Figure 6.

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

Figure 6. 2D render target map representing the scene depth buffer relative to the floor plane (**a**) and map after filtering by render depth (**b**).

This is the main scene-capturing component, which is located in the plane of the floor and is directed perpendicular to the ceiling of the room. The component registers render target texture in a 2D frame by frame in an orthographic projection, and displays the scene depth buffer relative to the floor plane. Figure 6a shows the results of depth buffer rendering with a maximum scan value of 0.17 m. The resulting frame with the depth mask

has the threshold value of 0.037 m and is normalized relative to the entire mask (Figure 6b). Thus, we obtain a "fingerprint" of the digital human body on the surface of a floor, which is further projected according to the view of the activated camera.

4. Experiments

4.1. Dataset

In this article, we performed the following digital human simulation and hit mask recognition research. Several experiments were carried out and the simulation values were initialized in each experiment: firstly, a digital human was randomly selected and forwarded in one specially generated room; secondly, many digital humans were randomly selected and placed in four unique generated rooms. Finally, the synthetic dataset contains of 577 simulations (first experiment) and 271 simulations (second experiment) accordingly. Each experiment amount of data includes three images from 16 cameras (48 images at one simulation): rgb frame, ground truth hit mask for "fall" and "not fall" cases, ground truth segmentation masks for floor and wall. In accordance with the digital human behaviour simulation algorithm, half of the simulations were carried out with fall results, while the other one captured non-fall results. For the second experiment, we specifically reduced the number of simulations, but on the other hand, the criteria number for generating the appearance of the digital human and the environment that affected the variance of the data was increased.

For the registration of an arbitrary moment without falling (walk or stand) and the moment of falling from the different views, 16 cameras were performed. The duration of each simulation was 1–3 s. Simulations were run on the PC with an 8th core and 16th threaded processor with 4.3 GHz and GeForce RTX 2080 video card. If a Person's location took place in the proximity of the walls and corners of the room, then some amount of cameras were expected to be outside the room. Therefore, data from such cameras were automatically excluded from the sample. As a result of two experiments, 27,476 and 12,698 images of 512 × 512 size were modelled and collected accordingly.

4.2. CNN Training

The CNN approaches can be applied in many areas. In fact, other CNN are unable to obtain a mask in order to segment the coordinates of impact with the object of interaction in the process of registering a digital human fall, therefore the Mask-R-CNN network [50] effectiveness was examined. We used our generated synthetic data to train the Mask R-CNN network for the prediction of pixel masks produced after a person interacts with any environment. In our case, we checked hitting the floor after a digital human's fall. We divided the dataset into training and test samples in a ratio 4:1, which were randomly selected from all amounts of simulation samples. The library Tensorflow 2.3 was used in the framework. We used SGD for gradient optimization with learning rate 0.001, momentum 0.9, weight decay 0.0001 and patch size 100. Initial weights were taken based on the ResNet-101 model weights pre-trained on the COCO dataset [51]. Segmentation masks involved three classes—floor, fall and does not fall, and the model itself was trained for 250 epochs.

The values of the loss function, which combines classification, localization, and segmentation mask losses, equal 0.483 and 0.394, respectively, as a result of training with synthetic data. Figure 7 shows the plot of the loss function curves.

The loss function was a combination of many resulting values for the basic loss parameters of the training model that generated a sum. There is the cross-entropy of an anchor classifier loss, bounding box difference (smooth L_1 norm) between target and recognized object and the difference (smooth L_1 norm) between target and predicted object mask. The difference or L_1 norm was calculated using the following expression:

$$L_{1,smooth} = \begin{cases} 0.5(y-x)^2, & if|y-x| < 1 \\ |y-x| - 0.5, & otherwise \end{cases}$$

where x is a training sample and y is a predicted result. Additionally, for cross-entropy, the function was used in the form:

$$L_{cross-entropy} = -\sum_{i=1}^{n} t_i \log(p_i)$$

where t_i is the truth label and p_i is the Softmax probability for the i-th class.

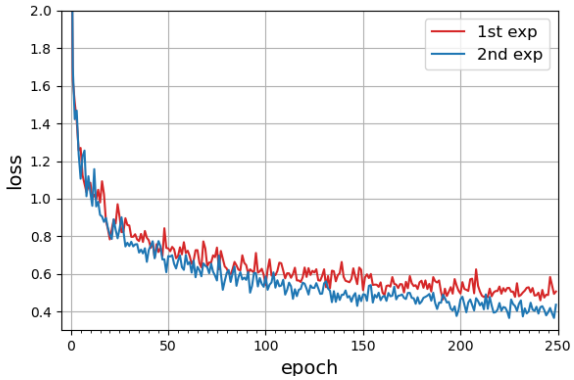
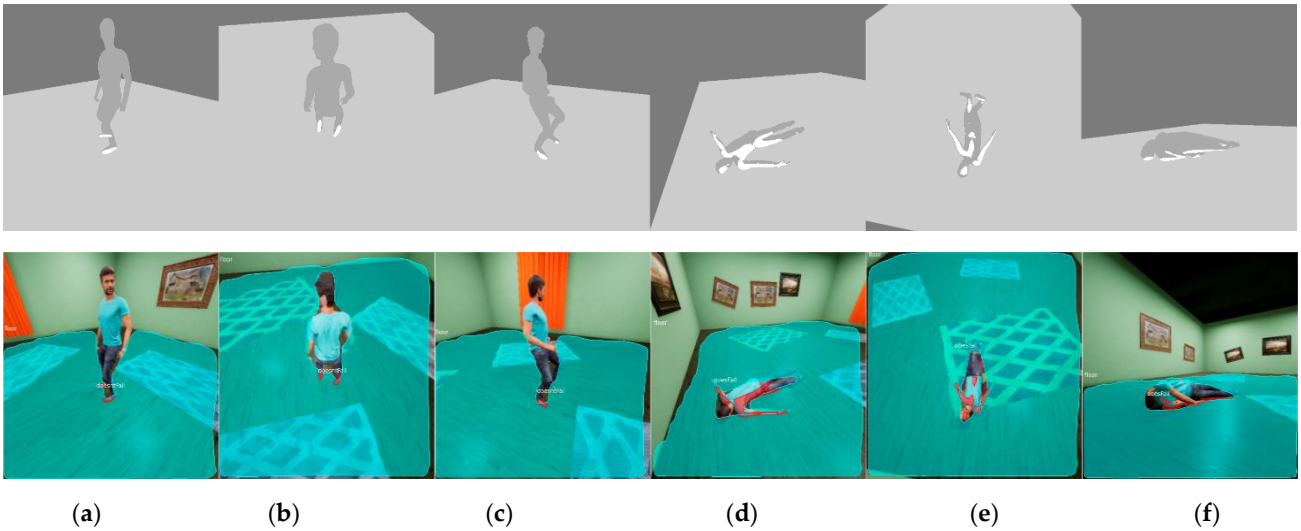


Figure 7. Loss functions for Mask R-CNN model trained on digital human fall data for the first and second simulation experiment.

4.3. Recognition Results

Figure 8 presents the simulation results of the digital human in two different rooms, where top—ground truth segmentation maps (hit map = 1.0; floor = 0.79; human = 0.66; walls = 0.48), and bottom—hit map predictions by model with rendered rgb images. In Figure 8a–c, the example images before falling from different cameras and labelled "did not fall" were shown, and Figure 8d–f, images are labelled "fall". The recorded samples of fall scenarios represent different interactions with the environment during a fall, which can be divided into three main cases: the subject falls forward (Figure 8d), falls back (Figure 8e) and falls on the side (Figure 8f), the description of which will be detailed below.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

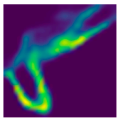
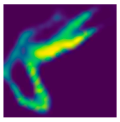
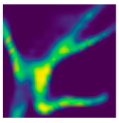
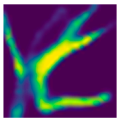
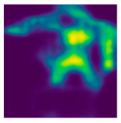
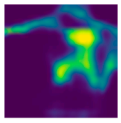
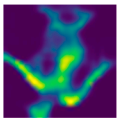
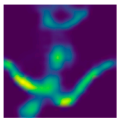
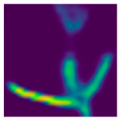
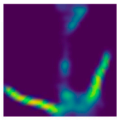
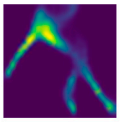
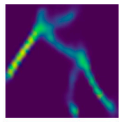
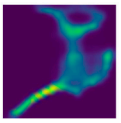
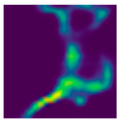
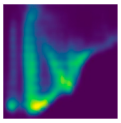
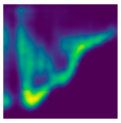
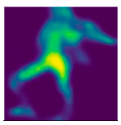
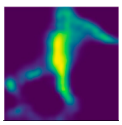
Figure 8. Ground truth (top) synthetic data generated in fall simulations; hit map predictions for digital human stand and fall positions (bottom). (**a**–**f**) Minimum detection confidence = 0.9.

It is possible to evaluate the unique scenarios of the fall, and the recorded data allows one to judge the nature of the impact as a result of conducted simulations with digital humans. Finally, all three simulation examples in Figure 8d–f have a head hit registration, as seen from the hit map data from the bottom of the figure.

As expected, increasing the variance of data generation in the second experiment (2nd exp) improved the model efficiency at the testing stage, while the amount of data used for

training was more than halved. The accuracy of the testing set was calculated considering "fall" and "not fall" classes detection. The value was 91.4% and 75.8% (1st exp) compared with 97.6% and 92.1% (2nd exp). Additionally, the 2nd exp trained model data appears to be more effective in the hit map prediction. Table 1 shows examples of comparison of specific predicted hit maps training data on the example of three types of falls.

Table 1. Ground truth (GT) and specific predicted hit maps for Mask R-CNN model trained synthetic data.

Fall Type	GT	Predicted Map, 1st Exp	Predicted Map, 2nd Exp
Forward fall (Camera #2)			
Forward fall (Camera #2)			
Forward fall (Camera #11)			
Back fall (Camera #9)			
Back fall (Camera #1)			
Back fall (Camera #4)			
Side fall (Camera #1)			
Side fall (Camera #6)			
Side fall (Camera #1)			

Hit map recognition on the generated synthetic data shows promising results. It was shown in Table 1 that predicted hit maps are able to convey information about the impact of individual parts of the body, so in all examples a head hit is traced. For comparative analysis, we specify the minimum detection confidence for the predicted mask with a value of 0.9. As a result, for a forward human fall example, we had following DICE metrics of 75.0% (2nd exp) with a comparison of 71.9% (1st exp) and IoU metrics of 59.7% (2nd exp) with a comparison of 55.5% (1st exp). The results of the segmentation metrics comparison were presented in Table 2.

Table 2. The values of the DICE and IoU segmentation metrics (detection_min_confidence = 0.9) corresponding to the registered fall types.

Data Generation	**1st Exp (577 Simulations) 1 Random Character, 1 Random Room**			**2nd Exp (271 Simulations) 2 Random Characters, 4 Random Rooms**		
Fall type	fw	back	side	fw	back	side
Total simulations	191	196	190	89	90	92
Total images		27,476			12,698	
DICE (%)	71.9	62.1	59.7	75.0	69.3	66.7
IoU (%)	55.5	48.0	44.7	59.7	52.3	49.4

In order to ensure that the proposed synthetic data generation pipeline is effective in real-world datasets, we conducted the experiment using the UR Fall dataset (URDFD) [52]. In Table 3, the accuracy for results for fall and not fall detection was given. This experiment shows the reliable efficiency of the proposed approach, and the results are comparable with the recognition accuracy obtained during training on real data presented in [53], where an accuracy of 95% was achieved. Figure 9 provides the recognition results on real data using Mask R-CNN trained by the synthetic data.

Table 3. The detection accuracy for human fall on URDF dataset.

Camera Type	**Camera1 (Filming Top-Down)**	**Camera0 (Filming Side)**
Accuracy	97.9%	87.5%

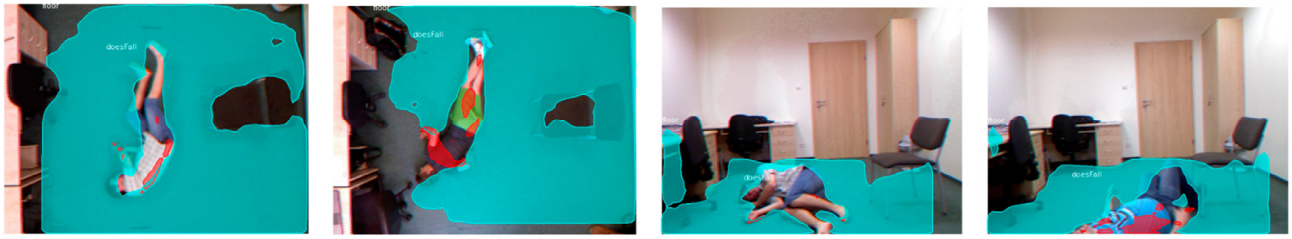


Figure 9. Examples of real URDF dataset human fall recognition results.

4.4. Discussion

There can be many other possibilities of such synthetic data generation pipeline usage. The described approach of digital human synthetic data generation can be effectively used in many applications. In addition, based on the received positive results, we can shed light on developing possibilities for a type of VR controller that can produce a new level of virtual environment interaction and deepen user VR experiences. The concept of such a controller is shown in Figure 10. The main idea is to provide the VR operator with a matrix of several sensors (pictured as blue and red on the suit) which use mechanic or electrical signals. As demonstrated in Figure 10, it is the connection between the virtual reality results of a digital human body interaction with the environment and with sensors on a suit that is schematically lighted as red for the correspondent hit map and blue—sensors that are not activated.

We also agree with the idea mentioned in [54] by M.A. Fallon and colleagues, who argue that virtual reality, in addition to collecting data from a neural network, allows one to determine the emotional state of a person as well as his psychophysiological load. At the same time, a properly constructed research design will allow us not only to solve the problems of assessing a person's fall in a three-dimensional environment, but also to additionally train him in the balance pose and other physiological exercises [55].

Based on analyses of the data obtained for the DICE and IoU metrics, it is necessary to notice the high recognition of falling to the side (Table 2) despite the small number of training examples. It is demonstrated by Table 2 that both experiments are characterized by a slight difference in the back and side fall types of DICE and IoU metrics. Despite the

fact that we set the minimum detection confidence for the predicted mask to a 0.9 value, the method for generating synthetic data for the hit map prediction as a whole allows us to achieve encouraging characteristics. Hereby, the Mask R-CNN model trained only via synthetics can be useful in a task regarding a prototype for the novel VR controller shown in Figure 10.

Figure 10. Concept of a human suit for VR environment interaction generation.

5. Conclusions

This paper presented a modular pipeline for generating synthetic data for the tasks of human interaction recognition with a 3D environment. The research included the following contributions: a synthetic dataset based on the procedural generation of realistic movements and falls, which take into account the physics models of digital humans; registering basic rgb and segmentation rendering maps while simulating a digital human fall; in segmentation maps, we presented unique hitting coordinate masks with the interaction of the human model and 3D scenes.

The pipeline modules included generating a human's and 3D environment's appearance, and also fall simulation based on a physical model of a digital human. We integrated the behaviour of digital humans in automatic scenarios. All modules of the pipeline were implemented in the open-source game engine, which allowed for high reconstruction availability for simulation. Our generated data included rgb maps, segmentation maps of 3D scene objects and hit maps.

It was also noted that one of the main challenges at the preparation stage in training neural networks is the collection of large-scale annotated data sets with minimal time and resource costs, especially the detection and classification of human interactions with a high occlusion coefficient, as a falling person can become a difficult task. Moreover, there is a high probability of erroneous and inaccurate manual annotation of such spatial kinds of data. It was shown that, by training the Mask R-CNN model via our generated synthetic data, it is possible to recognize a fallen human with an accuracy of 97.6%.

Author Contributions: Conceptualization, D.Z., L.Z., A.N. and S.C.; methodology, software D.Z. and L.Z.; resources, data curation, L.Z., S.A. and S.C.; validation, formal analysis, A.N., S.C. and D.Z.; writing—original draft preparation, D.Z. and L.Z.; writing—review and editing, A.N. and D.Z.; visualization L.Z. and A.S.; supervision, project administration A.N. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of communications of the Russian Federation, Grant 0000000007119P190002.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Ministry of Digital Development, Communications and Mass Media of the Russian Federation. The results were obtained as a part of the implementation of the program of activities of the Leading Research Centre (Grant Agreement No. 003/20 03-17-2020, Grant ID—0000000007119P190002).

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

References

1. Le, T.; Baydin, A.; Zinkov, R.; Wood, F. Using synthetic data to train neural networks is model-based reasoning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 3514–3521.
2. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
3. Wang, S.; Yue, J.; Dong, Y.; He, S.; Wang, H.; Ning, S. A synthetic dataset for Visual SLAM evaluation. *Robot. Auton. Syst.* **2020**, *124*, 1–13. [CrossRef]
4. Sharma, S.; Ball, J.E.; Tang, B.; Carruth, D.W.; Doude, M.; Islam, M.A. Semantic Segmentation with Transfer Learning for Off-Road Autonomous Driving. *Sensors* **2019**, *19*, 2577. [CrossRef]
5. Iqbal, J.; Xu, R.; Sun, S.; Li, C. Simulation of an Autonomous Mobile Robot for LiDAR-Based In-Field Phenotyping and Navigation. *Robotics* **2020**, *9*, 46. [CrossRef]
6. Wang, W.; Zhu, D.; Wang, X.; Hu, Y.; Qiu, Y.; Wang, C.; Hu, Y.; Kapoor, A.; Scherer, S. TartanAir: A Dataset to Push the Limits of Visual SLAM. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 1–8.
7. Antonini, A.; Guerra, W.; Murali, V.; Sayre-McCord, T.; Karaman, S. The Blackbird Dataset: A Large-Scale Dataset for UAV Perception in Aggressive Flight. In *Proceedings of the 2018 International Symposium on Experimental Robotics*; Xiao, J., Kröger, T., Khatib, O., Eds.; Springer Proceedings in Advanced Robotics; Springer: Cham, Switzerland, 2020; pp. 130–139.
8. Anwar, A.; Raychowdhury, A. Autonomous Navigation via Deep Reinforcement Learning for Resource Constraint Edge Nodes Using Transfer Learning. *IEEE Access* **2020**, *8*, 26549–26560. [CrossRef]
9. Muñoz, G.; Barrado, C.; Çetin, E.; Salami, E. Deep Reinforcement Learning for Drone Delivery. *Drones* **2019**, *3*, 72. [CrossRef]
10. Li, W.; Saeedi, S.; McCormac, J.; Clark, R.; Tzoumanikas, D.; Ye, Q.; Leutenegger, S. InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset. *arXiv* **2018**, arXiv:1809.00716.
11. Danielczuk, M.; Matl, M.; Gupta, S.; Li, A.; Lee, A.; Mahler, J.; Goldberg, K. Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7283–7290.
12. McCormac, J.; Handa, A.; Leutenegger, S.; Davison, A. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2678–2687.
13. Hu, J.; Choe, G.; Nadir, Z.; Nabil, O.; Lee, S.-J.; Sheikh, H.; Yoo, Y.; Polley, M. Sensor-realistic Synthetic Data Engine for Multi-frame High Dynamic Range Photography. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Virtual, 14–19 June 2020; pp. 516–517.
14. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243.
15. Richter, S.; Hayder, Z.; Koltun, V. Playing for Benchmarks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2213–2222.
16. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. Towards Semantic Segmentation of Urban-Scale 3D Point Clouds: A Dataset, Benchmarks and Challenges. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 4977–4987.
17. Dwivedi, S.; Athanasiou, N.; Kocabas, M.; Black, M.J. Learning to Regress Bodies from Images using Differentiable Semantic Rendering. In Proceedings of the 2021 IEEE International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 11250–11259.
18. Quirós-Ramírez, M.A.; Streuber, S.; Black, M.J. Red shape, blue shape: Political ideology influences the social perception of body shape. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 148. [CrossRef]
19. Zhang, S.; Zhang, Y.; Ma, Q.; Black, M.J.; Tang, S. PLACE: Proximity Learning of Articulation and Contact in 3D Environments. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 642–651.

20. Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.J.; Laptev, I.; Schmid, C. Learning from Synthetic Humans. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 109–117.
21. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
22. Hoffmann, D.; Tzionas, D.; Black, M.J.; Tang, S. Learning to Train with Synthetic Humans. In Proceedings of the 2019 German Conference on Pattern Recognition (GCPR), Dortmund, Germany, 11–13 September 2019; pp. 609–623.
23. Stewart Williams, J.; Kowal, P.; Hestekin, H.; O'Driscoll, T.; Peltzer, K.; Wu, F.; Arokiasamy, P.; Chatterji, S. Prevalence, risk factors and disability associated with fall-related injury in older adults in low- and middle-incomecountries: Results from the WHO Study on global AGEing and adult health (SAGE). *BMC Med.* **2015**, *13*, 147. [CrossRef]
24. Taheri, O.; Ghorbani, N.; Black, M.J.; Tzionas, D. GRAB: A Dataset of Whole-Body Human Grasping of Objects. *arXiv* **2020**, arXiv:2008.11200.
25. Harrou, F.; Zerrouki, N.; Sun, Y.; Houacine, A. An Integrated Vision-Based Approach for Efficient Human Fall Detection in a Home Environment. *IEEE Access* **2019**, *7*, 114966–114974. [CrossRef]
26. Wang, Z.; Ramamoorthy, V.; Gal, U.; Guez, A. Possible Life Saver: A Review on Human Fall Detection Technology. *Robotics* **2020**, *9*, 55. [CrossRef]
27. Kong, Y.; Huang, J.; Huang, S.; Wei, Z.; Wang, S. Learning Spatiotemporal Representations for Human Fall Detection in Surveillance Video. *J. Vis. Commun. Image Represent.* **2019**, *59*, 215–230. [CrossRef]
28. Asif, U.; Mashford, B.; Cavallar, S.; Yohanandan, S.; Roy, S.; Tang, J.; Harrer, S. Privacy Preserving Human Fall Detection using Video Data. In Proceedings of the 2020 Machine Learning for Health NeurIPS Workshop, PMLR, Virtual Conference, 11 December 2020; pp. 39–51.
29. Liu, J.; Rahmani, H.; Akhtar, N.; Mian, A. Learning Human Pose Models from Synthesized Data for Robust RGB-D Action Recognition. *Int. J. Comput. Vis.* **2019**, *127*, 1545–1564. [CrossRef]
30. Clever, H.; Erickson, Z.; Kapusta, A.; Turk, G.; Liu, K.; Kemp, C. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 6215–6224.
31. Min, W.; Zou, S.; Li, J. Human fall detection using normalized shape aspect ratio. *Multimed. Tools Appl.* **2019**, *78*, 14331–14353. [CrossRef]
32. Martínez-Villaseñor, L.; Ponce, H.; Brieva, J.; Moya-Albor, E.; Núñez-Martínez, J.; Peñafort-Asturiano, C. UP-Fall Detection Dataset: A Multimodal Approach. *Sensors* **2019**, *19*, 1988. [CrossRef]
33. Fan, K.; Wang, P.; Zhuang, S. Human fall detection using slow feature analysis. *Multimed. Tools Appl.* **2019**, *78*, 9101–9128. [CrossRef]
34. Fan, Y.; Levine, M.D.; Wen, G.; Qiu, S. A deep neural network for real-time detection of falling humans in naturally occurring scenes. *Neurocomputing* **2017**, *260*, 43–58. [CrossRef]
35. Mirmahboub, B.; Samavi, S.; Karimi, N.; Shirani, S. Automatic monocular system for human fall detection based on variations in silhouette area. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 427–436. [CrossRef]
36. Gasparrini, S.; Cippitelli, E.; Spinsante, S.; Gambi, E. A depth-based fall detection system using a kinect® sensor. *Sensors* **2014**, *14*, 2756–2775. [CrossRef]
37. Auvinet, E.; Multon, F.; Saint-Arnaud, A.; Rousseau, J.; Meunier, J. Fall Detection With Multiple Cameras: An Occlusion-Resistant Method Based on 3-D Silhouette Vertical Distribution. *IEEE Trans. Inf. Technol. Biomed.* **2011**, *15*, 290–300. [CrossRef] [PubMed]
38. Charfi, I.; Miteran, J.; Dubois, J.; Atri, M.; Tourki, R. Optimized spatio-temporal descriptors for real-time fall detection: Comparison of support vector machine and adaboost-based classification. *J. Electron. Imaging* **2013**, *22*, 041106. [CrossRef]
39. Asif, U.; Cavallar, S.; Tang, J.; Harrer, S. SSHFD: Single Shot Human Fall Detection with Occluded Joints Resilience. In Proceedings of the 2020 European Conference on Artificial Intelligence (ECAI), Santiago de Compostela, Spain, 29 August–8 September 2020; pp. 1–8.
40. Rahmani, H.; Mahmood, A.; Huynh, D.; Mian, A. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2430–2443. [CrossRef]
41. Bak, S.; Carr, P.; Lalonde, J.F. Domain adaptation through synthesis for unsupervised person re-identification. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2014; pp. 1–17.
42. Shahroudy, A.; Liu, J.; Ng, T.; Wang, G. NTU RGB+D: A large scale dataset for 3d human activity analysis. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
43. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339. [CrossRef]
44. Kiourt, C.; Koutsoudis, A.; Pavlidis, G. DynaMus: A fully dynamic 3D virtual museum framework. *J. Cult. Herit.* **2016**, *22*, 984–991. [CrossRef]
45. Unreal Engine 4. Available online: https://www.unrealengine.com/en-US/what-is-unreal-engine-4 (accessed on 16 October 2021).
46. RenderPeople—A Photorealistic Human 3D Models. Available online: https://renderpeople.com/ (accessed on 16 October 2021).

47. Millington, I. *Game Physics Engine Development: How to Build a Robust Commercial-Grade Physics Engine for Your Game*; CRC Press: Boca Raton, FL, USA, 2010.
48. Aristidou, A.; Lasenby, J.; Chrysanthou, Y.; Shamir, A. Inverse Kinematics Techniques in Computer Graphics: A Survey. *Comput. Graph. Forum* **2017**, *37*, 35–58. [CrossRef]
49. Iovino, M.; Scukins, E.; Styrud, J.; Ögren, P.; Smith, C. A Survey of Behavior Trees in Robotics and AI. *arXiv* **2020**, arXiv:2005.05842.
50. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
51. 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 2014 European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
52. Kwolek, B.; Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput. Methods Programs Biomed.* **2014**, *117*, 489–501. [CrossRef] [PubMed]
53. Núñez-Marcos, A.; Azkune, G.; Arganda-Carreras, I. Vision-based fall detection with convolutional neural networks. *Wirel. Commun. Mob. Comput.* **2017**, *2017*, 9474806. [CrossRef]
54. Fallon, M.; Riem, M.; Kunst, L.; Kop, W.; Kupper, N. Multi-modal responses to the Virtual Reality Trier Social Stress Test: Acomparison with standard interpersonal and control conditions. *Int. J. Psychophysiol.* **2021**, *161*, 27–34. [CrossRef]
55. Garduño, H.; Martínez, M.; Castro, M. Impact of Virtual Reality on Student Motivation in a High School Science Course. *Appl. Sci.* **2021**, *11*, 9516. [CrossRef]