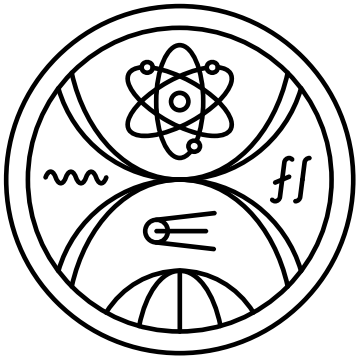
Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Augmentation of Synthetic Human 3D Scans for Deep Learning

Master's Thesis

2024
Bc. Martin Halaj

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics



Augmentation of Synthetic Human 3D Scans for Deep Learning

Master's Thesis

Study Programme:	Applied informatics
Field of Study:	Computer Science
Department:	Department of Applied Informatics
Supervisor:	RNDr. Martin Madaras, PhD.
Consultant:	Mgr. Dana Škorvánková

Bratislava, 2024
Bc. Martin Halaj

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Martin Halaj
Study programme:	Applied Computer Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Augmentation of Synthetic Human 3D Scans for Deep Learning

Annotation: Machine learning-based approaches require a large amount of training data to perform well on previously unseen test data. If there is no available real annotated training dataset a synthetic one can be rendered and used. Regarding all human body-related tasks, it is important to make the virtual data qualitatively as similar as possible to the real human body scans, so that the statistical models are able to generalize well in real-world scenarios. Therefore, we would like to enhance SMPL-based models with cloth, hair, or other details missing in the SMPL-X models and render the synthetic data as similar to real scans as possible.

Aim: The main goal of the thesis is to augment synthetically generated data to adapt the synthetic domain to the domain of real human body data. Add more details to the generated human body data, such as clothes, hair, and realistic background scene. Moreover, the aim is to evaluate the data by using it as a supervision signal to train and test a machine learning model.
- Study relevant papers concerning synthetic data generation, especially human body data (SMPL, SMPL-X)
- Use the existing framework for synthetic data generation using SMPL and enhance the framework for the generation of clothes, hair, backgrounds, etc.
- Generate a set of training and testing data and perform a quantitative evaluation using neural networks for pose estimation and body measurements estimation
- Write a thesis with focus given to experiments and quantitative evaluation of the generated data

Literature: Loper et al. 2015, SMPL: A Skinned Multi-Person Linear Model, ACM Trans. Graphics Proc. SIGGRAPH Asia, p. 1-16, volume 34, 2015

Pavlakos et al. 2019, Expressive Body Capture: 3D Hands, Face, and Body from a Single Image, Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, p. 10975-10985, CVPR 2019

Keywords: synthetic data, human body, SMPL, deep learning

Supervisor:	RNDr. Martin Madaras, PhD.
Consultant:	Mgr. Dana Škorvánková
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	doc. RNDr. Tatiana Jajcayová, PhD.

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

Assigned: 04.10.2022

Approved: 07.10.2022 prof. RNDr. Roman Ďurikovič, PhD.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Martin Halaj
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Augmentation of Synthetic Human 3D Scans for Deep Learning
Vylepšovanie syntetických 3D skenov ľudí pre hlboké učenie

Anotácia: Prístupy využívajúce strojové učenie vyžadujú veľké množstvo trénovacích dát na správne fungovanie na neznámych testovacích dátach. Ak nie je k dispozícii žiadny skutočný anotovaný súbor trénovacích dát, je možné vykresliť a použiť syntetický 3D sken. V kontexte všetkých úloh súvisiacich s ľudským telom je potrebné, aby sa virtuálne dáta kvalitatívne čo najviac podobali reálnym skenom ľudí, aby štatistické modely správne generalizovali v reálnom nasadení. Preto by sme chceli rozšíriť modely založené na SMPL s oblečením, vlasmi a inými chýbajúcimi detailami v SMPL-X modeloch a renderovať syntetické dáta čo najpodobnejšie reálnym skenom.

Cieľ: Hlavným cieľom diplomovej práce je rozšíriť synteticky generované údaje tak, aby sa syntetická doména prispôsobila oblasti skutočných údajov o ľudskom tele. Pridajte ďalšie podrobnosti k vygenerovaným údajom o ľudskom tele, ako sú oblečenie, vlasy a realistická scéna na pozadí. Okrem toho je cieľom vyhodnotiť údaje ich použitím ako kontrolného signálu na trénovanie a testovanie modelu strojového učenia.
- Preštudujte si relevantné články týkajúce sa generovania syntetických dát, najmä údajov o ľudskom tele (SMPL, SMPL-X)
- Použite existujúci systém na generovanie syntetických údajov pomocou SMPL a vylepšite systém na generovanie oblečenia, vlasov, pozadia atď.
- Vytvorte množinu trénovacích a testovacích dáta, vykonajte kvantitatívne vyhodnotenie pomocou neurónových sietí na odhad pózu a odhad telesných rozmerov
- Napíšte diplomovú prácu so zameraním na experimenty a kvantitatívne vyhodnotenie získaných dát

Literatúra: Loper et al. 2015, SMPL: A Skinned Multi-Person Linear Model, ACM Trans. Graphics Proc. SIGGRAPH Asia, p. 1-16, volume 34, 2015

Pavlakos et al. 2019, Expressive Body Capture: 3D Hands, Face, and Body from a Single Image, Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, p. 10975-10985, CVPR 2019

Kľúčové slová: syntetické dáta, telo človeka, SMPL, hlboké učenie

Vedúci: RNDr. Martin Madaras, PhD.

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Konzultant: Mgr. Dana Škorvánková
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 04.10.2022

Dátum schválenia: 07.10.2022 prof. RNDr. Roman Ďurikovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

I hereby declare that I have written this thesis by myself, only with help of referenced literature, under the careful supervision of my thesis advisor.

.....

Bratislava, 2024 Bc. Martin Halaj

Acknowledgments:

Abstract

Keywords: synthetic data, human body, SMPL, deep learning

Abstrakt

Kľúčové slová: syntetické dáta, telo človeka, SMPL, hlboké učenie

Contents

List of Figures

List of Tables

Introduction

Chapter 1

Motivation

Chapter 2

Theoretical Background

2.1 Machine learning

2.1.1 Regression

2.1.2 Neural Networks

Basics

Convolutional Neural Network

Generative Neural Network

Segmentation

intersection over union (IOU), pixel accuracy measures for evaluating the segmentation

2.2 Mesh Processing

2.2.1 3D Avatars

2.2.2 Evaluation Metrics

Chapter 3

Related Work

In this chapter, we will describe researches related to our work. Since we use 3D avatars throughout our work, we will first describe popular 3D models and inspect carefully how they work. Next, we will take a closer look at body measurements field.

3.1 3D Avatars

Depicting of human body is crucial in various fields like the game or film industry. In addition, it is also used in diverse types of simulations, animations, or for research purposes. Because the shape of the human body is very complex and likewise asymmetrical, it is complicated to realistically model a human body. The reason why is this hard is that the human body consists of many different parts, we have about 600 muscles, approximately 200 bones, many joints, and so forth. The result is that many experts around the world are interested in this issue. Thanks to this, a number of different models depicting the human body were created. Now we take a closer look at some of the best and the most popular models of the human body at this time.

SMPL model

The former model of the SMPL family is the base SMPL model, which stands for a Skinned Multi-Person Linear model. We can say without any hesitation that models from this family are the most popular body models among scientists and researchers right now. The model was introduced in 2015 [16].

Their goal was to create a model that would be as realistic as possible, including the natural deformation of tissue or simulating the believable motion of soft tissue. They also want to create a model that can be easily rendered without the need for manual intervention and is fully compatible with existing rendering engines or graphics tools. This base model does not support finger movements or facial expressions. The resulting model is easy to animate or control. For our purposes, we focus on high-level model

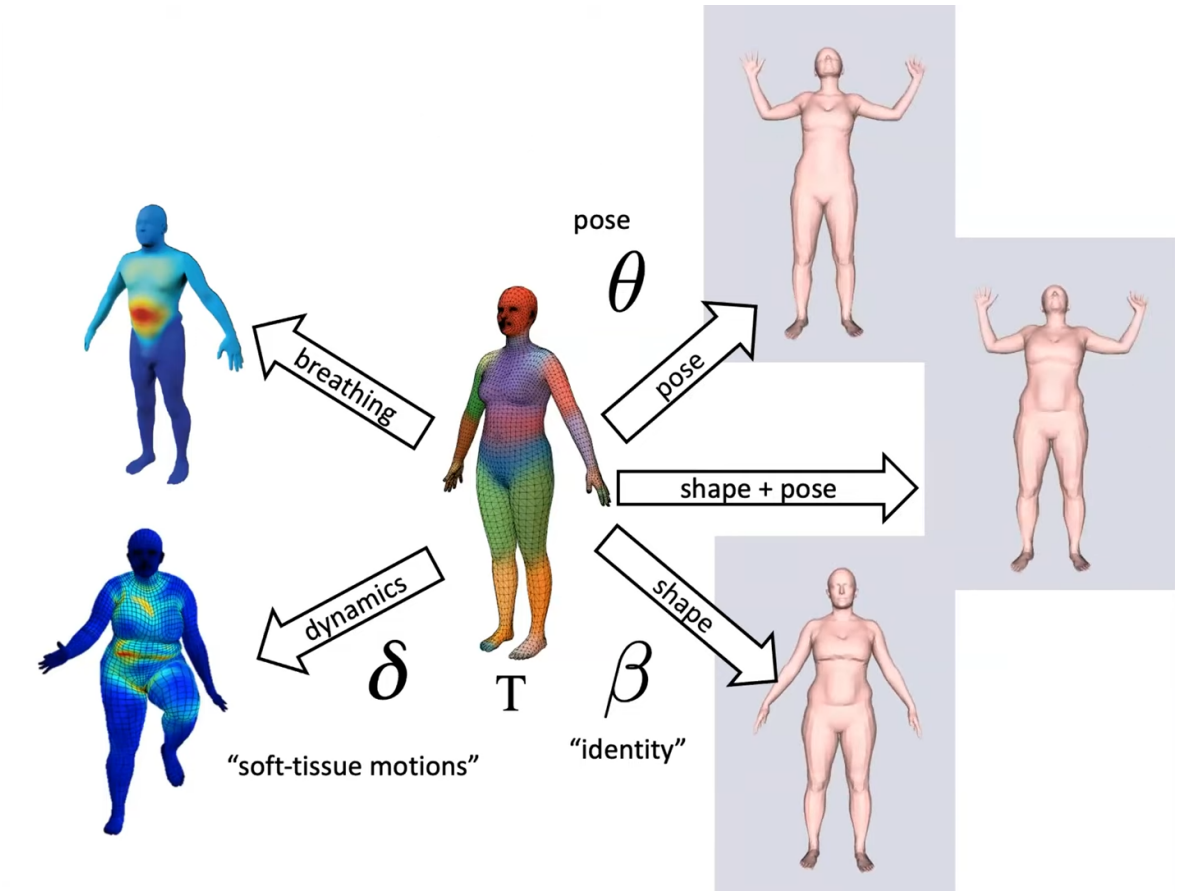


Figure 3.1: Representation of SMPL model [3].

definition with an emphasis on practical use. You can find omitted details in the original article.

As we can see in the Figure 3.1, the model is defined by four basic parameters T, β, θ and δ [18]. The first parameter T represents a template mesh that is used as the foundation mesh of the resulting SMPL avatar. The mesh was created by the artists and is composed of 6890 vertices and 23 joints. The topology of the mesh is the same for both sexes and the neutral gender model. The second parameter is β. It influences the shape of the avatar. We can image it as a vector of numbers, usually in the range $\langle -3; 3 \rangle$. The number of these parameters can vary up to 300. In general, using 10 parameters is sufficient enough to obtain a lifelike body model. Utilizing more than 10 parameters is used for representing finer details on the human body like dimples on the cheeks and things like that. We can also think about β as a layer that can be added on top of the template mesh we mentioned before in order to create a human with a different body shape. These parameters directly influence physical proportions of resulting avatar like height, waist circumference, chest circumference and many other metrics. The third parameter that is part of the SMPL model definition is θ. This parameter is responsible for the pose in which the resulting avatar is rendered. As in

the previous case, the pose is also defined by a vector of numbers. In this vector, we have defined the rotation for every joint of our model. Each joint is defined by three numbers that represent the rotation of the joint in the direction of the x, y, and z axes. We mentioned above that the model consists of 23 joints. It is therefore obvious that the total length of the pose vector is 69 in the case of the base SMPL model. The rotation of all joints, except the root joint, is relative to the parent joint according to the kinematic tree. Also, it is important to mention that all angles are in radians. The last parameter that we gave above was δ. In contrast with other mentioned parameters, δ is usually not a frequently utilized parameter. This parameter is responsible for the movement of soft tissue.

The default value for any parameter is a vector that consists of zeros. If we define all parameters as zero vectors in the required shape, we will define a model $M(\beta, \theta, \delta)$ that will produce an avatar standing in the T-pose. Avatars rendered in this way also have a skeleton, where one bone connects two different joints together.

SMPL-H model

After releasing the base SMPL model, it started gaining popularity among scientists and other research communities. However, as we mentioned before, it has some drawbacks and researchers tried to address them in various works. The first of the most significant works that tried to enhance the base SMPL model was the MANO model [25]. The base SMPL model has hands that cannot change hand pose. Because the human body and hands cannot be separated, the authors of this work decided to confront this issue. They focused on creating a model of a hand that could use individual fingers and move with them exactly how people can. This new model was learned from approximately one thousand 3D scans of human hands. Scans were captured by scanning the hands in various poses of 31 participants. The resulting hand model is composed of bones and joints similar to the base SMPL model. The SMPL-H model represents the conjunction of the base SMPL model and previously described the MANO model. It is also created based on four parameters T, β, θ, and δ. The only difference between SMPL and SMPL-H models is the shape of the vectors that represent the individual parameters because of the increased number of joints. In the base SMPL model, each hand uses only one joint, but in SMPL-H each hand consists of 15 joints. The resulting avatar based on the SMPL-H model is composed of 51 joints instead of only 23. The β vector representing the shape of the body is the same for both models. The same applies to a vector that defines the δ parameter. On the other hand, the vector defining θ that is responsible for pose does not match with the SMPL one. Because of the increased number of joints, it also has to be longer, thus this vector contains 153 values for each joint and all axes.

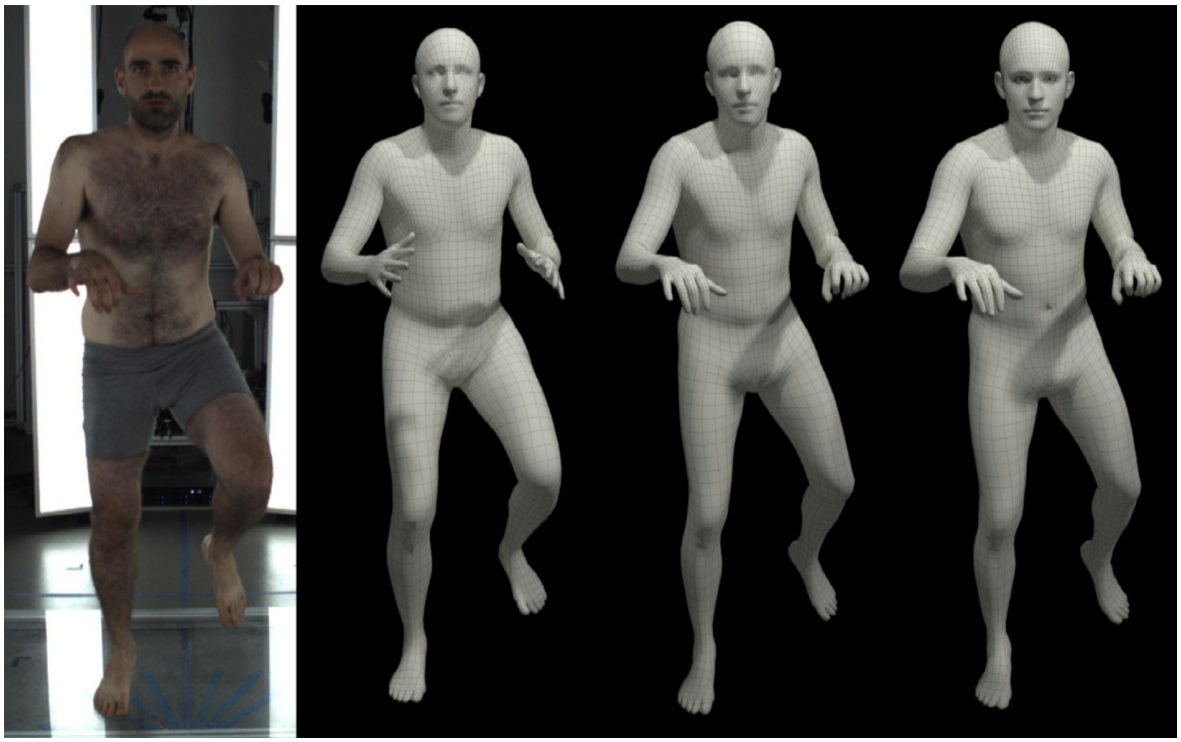


Figure 3.2: Comparison of SMPL, SMPL-H and SMPL-X model [20].

SMPL-X model

The second crucial thing the basic SMPL model is missing is the ability to express facial expressions. Scientists and researchers at the Max Planck Institute for Intelligent Systems in California have attempted to address this last major shortcoming. In 2017, they introduced a new FLAME model [15]. The FLAME is an abbreviation of Faces Learned with Articulated Model. Similar to the basic SMPL model, one of their basic goals was to create a model that can be easily used with existing graphics software. The model uses linear shape space. During the research, they collected approximately 3800 scans of human heads. Later, they used these data to train linear shape space.

The resulting model outperformed both popular models at that time, Basel Face Model [21] from 2009 and FaceWarehouse [5] from 2014. The first mentioned work result in the Basel Face Model. The model was trained using 200 scans of a human face that consisted predominantly of young Europeans with neutral expression. The first half consists of male scans and the second female scans. Their main goal was to create a generative model that can capture 3D space of human face. The model should be able to capture it from standard 2D picture but also from 3D scan. In the comparison with the first model, the second work includes variance in ethnicity and age of scanned subjects. They scanned 150 people together and for every one they capture 20 scans with different facial expression. For scanning they used in that time highly available and also affordable Microsoft Kinect 360 or Microsoft Kinect V1

as sometimes called. This device was introduced in 2010 and using structured light principle for creating 3D scans or more precisely point clouds. The final product of this work was FaceWarehouse dataset with template mesh for every scan.

Similar to the SMPL-H model, the SMPL-X was created by combining two models, in this case SMPL-H with the described FLAME model. The resulting SMPL-X model has a mesh made of 10 475 vertices and the model has 54 joints. In comparison to SMPL-H model, there are 3 extra joints. Each eye has its own joint. It is designed to turn the eye in a specific direction. The last joint is used to move the jaw, so a person can keep their mouth open or closed thanks to it. As with other models, SMPL-X model is defined by four main parameters T, β, θ and δ. The only difference is the size of the vectors that contain the values for the various parameters. These increased equivalently to the increase in the SMPL-H model.

Finally, you can see the final comparison of all three described model in Figure 3.2. On the very left, you can see a real photo of man that was used as reference model for quality comparison among different SMPL models. We will describe models from left to right. The first image represents the basic SMPL model. As you can see, this model cannot imitate hands properly nor the facial expression. The model is able to adapt to the position of the hand, but not to rotate it. It also cannot place individual fingers in the right shape. In the second image is depicted SMPL-H model. We can observe significant improvement in positioning of hand and individual fingers. The last image in the very right represents SMPL-X model. You can see that it has the same positioning of hands and fingers as SMPL-H model. In addition, we can see quite believable depiction of facial expression. It is also observable that SMPL-X model has finer body in comparison with SMPL and SMPL-H model. This is caused by finer mesh thanks to higher vertices number compared to others. This finer mesh is the best noticeable on the forehead.

Unity Synthetic Humans

There is evidence that using synthetic data during training the neural network can enhance final performance of the whole model [23]. One of the most popular game engines right now is Unity [1]. Unity also recognizes the fact that synthetic data can improve the performance of machine learning models, so they decided to develop the Unity Computer Vision [2] package. This package is aimed to scientists and researchers. The package contains several tools for different tasks. For now, we will only talk about Synthetic Humans package in a little bit more detail. The package is now available for academic use only.

[1] `https://unity.com/`

[2] `https://unity.com/products/computer-vision`

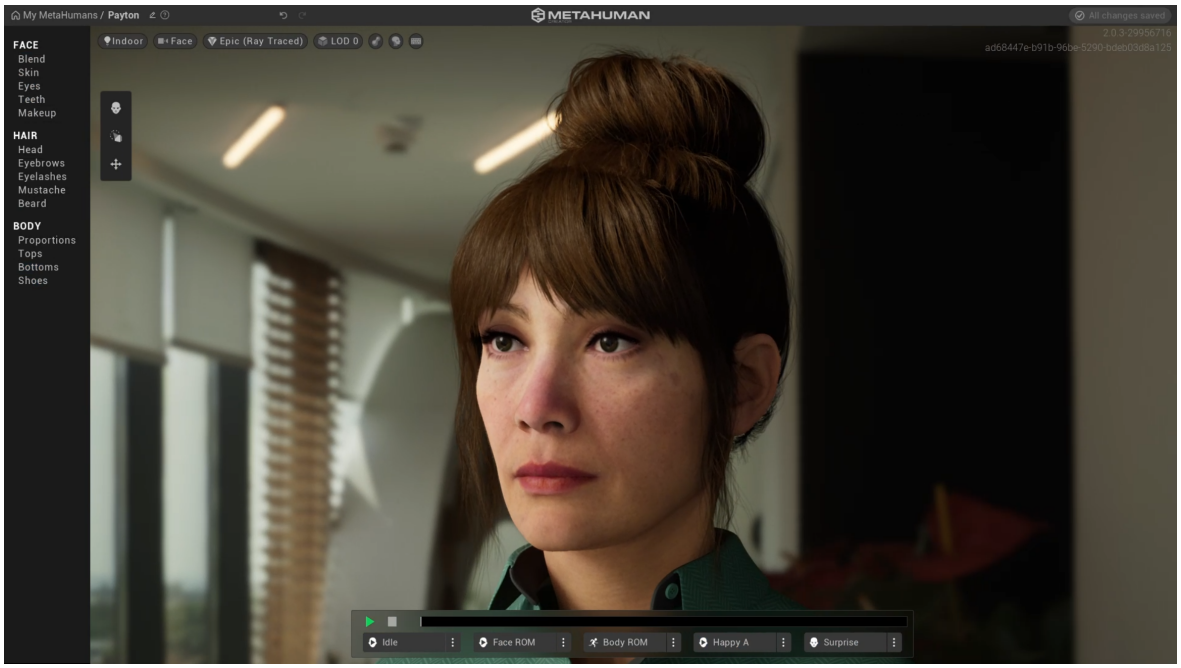


Figure 3.3: Creating avatar in MetaHuman Creator.

Unity Synthetic Humans serve as generator to procedurally generate one or many synthetic humans [29]. It is worth to note that this package uses Unity's own avatars and not the ones from the SMPL family. Users can influence various parameters of avatar generation. They can choose age, height, type of body, clothes and etcetera of generated avatar. Furthermore, it is also possible to change position or rotation of avatar. But the potential of Synthetic Humans will start to increasing after the connection with another Synthetic Homes package. This package is capable of generating synthetic backgrounds for avatars. In addition, it generates depth map, normals, segmentation and so on for the generated scene [28]. Together, these two packages can produce a training dataset for a neural network. This data can subsequently improve the performance of the final model.

On the other hand, the disadvantage is that avatars generated this way cannot be used outside Unity. It means user cannot export generated avatar and use it in another graphic software.

MetaHuman

As we mentioned at the beginning of this chapter, one of the areas where 3D avatars play an important part is the video game industry. Humans are an integral part of many modern video games. Therefore, companies that develop game engines or games are forced to participate on research of 3D avatars but also bring innovations in order to gain a competitive advantage.

Unreal Engine [3], similarly to Unity, rank among the most popular game engines right now. In 2021, Epic Games, the developer of Unreal Engine, introduced a new type of avatars called MetaHuman [4] that are one of the state-of-the-art. For work with avatars, they also presented MetaHuman Creator, MetaHuman Animator and Mesh to MetaHuman tools. In the Figure 3.3, we can see a screenshot of MetaHuman Creator interface. You can likewise notice how realistic these avatars are. It is a browser-based graphic user interface that allows users to create and adjust their own MetaHuman that they can subsequently transfer to Unreal Engine and use it in video games or animations. Epic Games decided to use the browser as a way to run MetaHuman Creator because it is very hardware intensive and would be challenging for users to use it on their own machines [8]. Compared with Creator part, MetaHuman Animator works directly in Unreal Engine. This part is used to capture facial performance of actors and transfer it to MetaHuman using only iPhone and computer, but this process is quite demanding on hardware performance [7]. Last tool Mesh to MetaHuman is used to convert user's mesh to MetaHuman.

Unfortunately, the huge disadvantage is that it is not possible to use MetaHuman avatar outside Unreal Engine. On the other hand, using MetaHuman inside Unreal Engine is possible without any extra payment.

3.2 Virtual Body Measurements

Creation of an exact virtual copy of a person is not a simple task. However, there is increasing demand for this type of service. To have your own virtual character that looks just like you has a good deal of different utilization. One of the usages is that you can put yourself into video game or animation, which could even more enhance user's gaming experience. A more practical application are virtual dressing rooms. In addition, these could also be useful for medical or ergonomic purposes. There is a vision that you will have your virtual avatar, then you can browse at e-shop with clothes and try the clothes from e-shop to your avatar to see how it would look like on you.

How we mentioned in the beginning of the section, creating exact copy of a person is complex task. Nevertheless, creating a copy of face is more complicated than creating a copy of body. This is due to the fact that it is relatively easy to parameterize the body through various metrics such as height, weight, etc. There are two ways how this can be performed. One way is extraction of these parameters from photo or SMPL model. The second way is generating a model from input parameters. We will now describe some tools and frameworks that focus on these tasks.

[3] `https://www.unrealengine.com`

[4] `https://www.unrealengine.com/en-US/metahuman`

3.2.1 Non-commercial Solutions

There are a number of different studies dealing with this issue. These researches gradually advance the possibilities in the given area, but at the same time they leave room for further research and development. Now we will take a closer look at several researches that are connected to our work.

SMPL Anthropometry

Anthropometry is a scientific field that deals with the study of the measurement of the human body. Thanks to anthropometry, we can parametrize human bodies, which is very useful for our purposes.

In 2023, David Boja released his tool, called SMPL Anthropometry, on GitHub [5]. This tool serves for measuring of SMPL or SMPL-X body models. User can choose from two input options. The first way is to use β parameters as input and the second is to use directly vertices of mesh. After inserting data, the tool will measure the input model. The result consists of 16 different measurements, for example height or head, chest, waist circumference and so on. There are also measurements like arm or leg length and other similar. The whole program is written in Python. It is also possible to run this tool in Docker [6] container [4].

SHAPY

There are many different researches dealing with human model extraction from photographs. Usually they take photograph as input and then output the estimated virtual avatar or predicted anthropometric measurements. As you might expect, estimating a 3D model from a 2D image is quite challenging.

One study in this category introduced SHAPY model [6]. It is a deep neural network that is capable of predicting pose and shape from a single RGB image. The big difference is that SHAPY estimates shape by combination of anthropometric measurements and semantic attributes. The term semantic attributes refer to linguistic shape attributes that they gathered using crowdsourcing. These attributes describe a characteristic of human body. There are three categories, the first contains common attributes for both sexes like short, big, long neck, long legs, short arms and so on. The second category is intended only for males and contain attributes like skinny arms, masculine, etc. On the other hand, the last third category is intended only for females. There are attributes like pear shaped, feminine, skinny legs and others. Each attribute has its own rating or weight that defines the body model.

[5] `https://github.com/`
[6] `https://www.docker.com/`

It offers two components to use linguistic attributes. The first is Attributes to Shape (A2S) and the second is Images to Attributes (I2A). As the name suggests, A2S is used to generate SMPL-X body model from input attributes with their values. As you might expect, I2A takes color image as input and generate attributes of estimated body model.

During the research, they created a dataset called Human Bodies in the Wild (HBW). They use this dataset and some others to experiment with the model in order to compare their model with at that time state-of-the-art model. They showed that SHAPY significantly outperform other models on the HBW dataset.

We see the only disadvantage of this that it is impossible to obtain body parameters in metric units, such as waist circumference and the like.

The Virtual Caliper

As we mentioned in the beginning of this section, there are two ways how tools in this category work. In 2019, researchers, predominantly from Max Plank Institute, develop a tool called the Virtual Caliper [22]. This tool belongs to the second category, what is generating a 3D avatar from input parameters.

The Virtual Caliper is a tool for generating a base SMPL body model from input anthropometric measurements. It has a variable number of input parameters from only two basic parameters, overall height and weight, up to six parameters. Other parameters completing the six are arm span fingers, inseam height, hip width and arm length. The whole application is developed in Unity game engine we mentioned earlier.

There are two options how you can input these parameters to the Virtual Caliper. The first option requires virtual reality headset HTC Vive [7] with his game controllers. These are used to capture the dimensions of the body by positioning controllers to different parts of the body. These are subsequently fed into the Virtual Caliper. The second option is to manually insert required parameters to the program through a graphic user interface. In the Figure 3.4 is depicted the interface of the Virtual Caliper GUI. In the bottom half, we can preview of generated avatars. In the upper half, we can see various values. At the very left are position sliders for input parameters. Above them, we can choose the number of input parameters for our model. Next to that, we can observe measurements of generated avatars with calculated error in comparison with required values. The last things that are depicted in the image are F Beta and M Beta. These represent a shape values of base SMPL model. Finally, there is an export button in the right upper corner.

It is interesting how authors work with weight. It is obvious that virtual avatar cannot be weighted, so they see it as a linear relation between weight and volume

[7] `https://www.vive.com/`

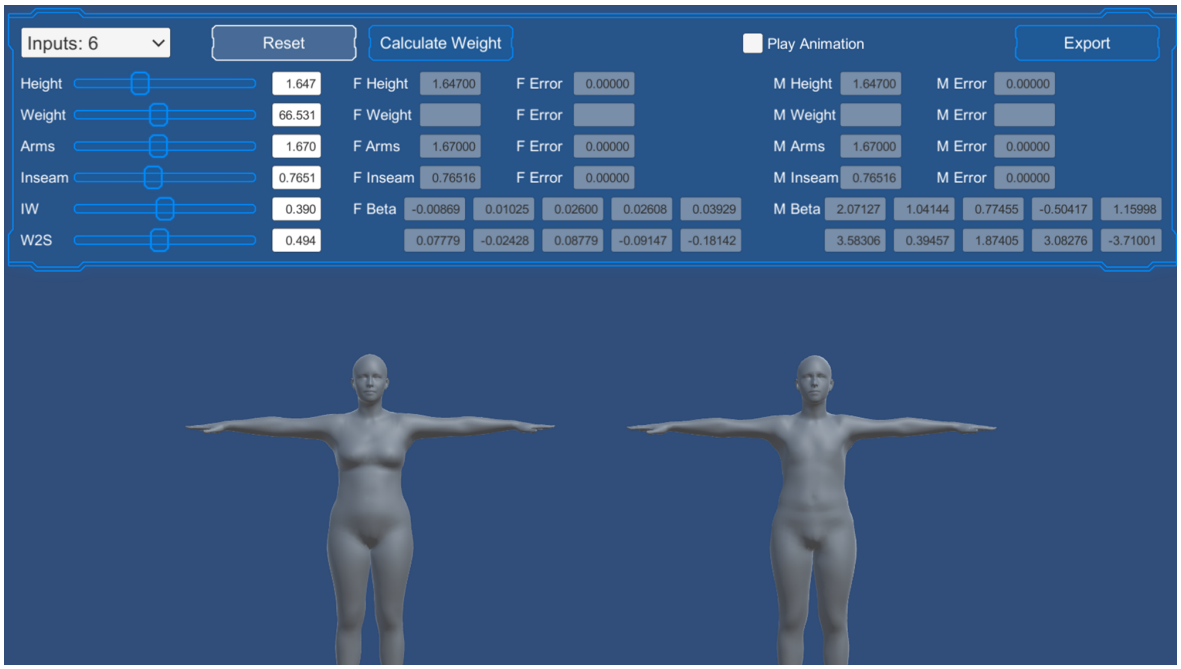


Figure 3.4: Main interface of Virtual Caliper.

of the mesh. They discovered this relation by observation on data from CAESAR database [24]. This database contains 2000 scans per gender, and it is also available the weight for every subject. For generating output parameters from input metrics, they use regressors to predict them. They also found out that the relation between shape parameters and input measurements is linear.

3.2.2 Commercial Solutions

As we stated before, there is also a great deal of commercial potential in this research area. That implies that there are also commercial subjects that offering their services to users. Some of them are focused only on creating a copy of human, but some of them are directly focused on fashion industry. Now we will describe some selected tools that are available right now.

Meshcapade

The Meshcapade [8] is a start-up that is relatively closely connected to Max Planck Institute, since it was founded by three researchers and employees from this institute. The company was founded in 2018.

Digidoppel [9] is a platform developed by Meshcapade. The entire application is browser-based, so all processing is done on the Meshcapade side. It offers two ways to

[8] `https://meshcapade.com/`
[9] `https://digidoppel.com/`

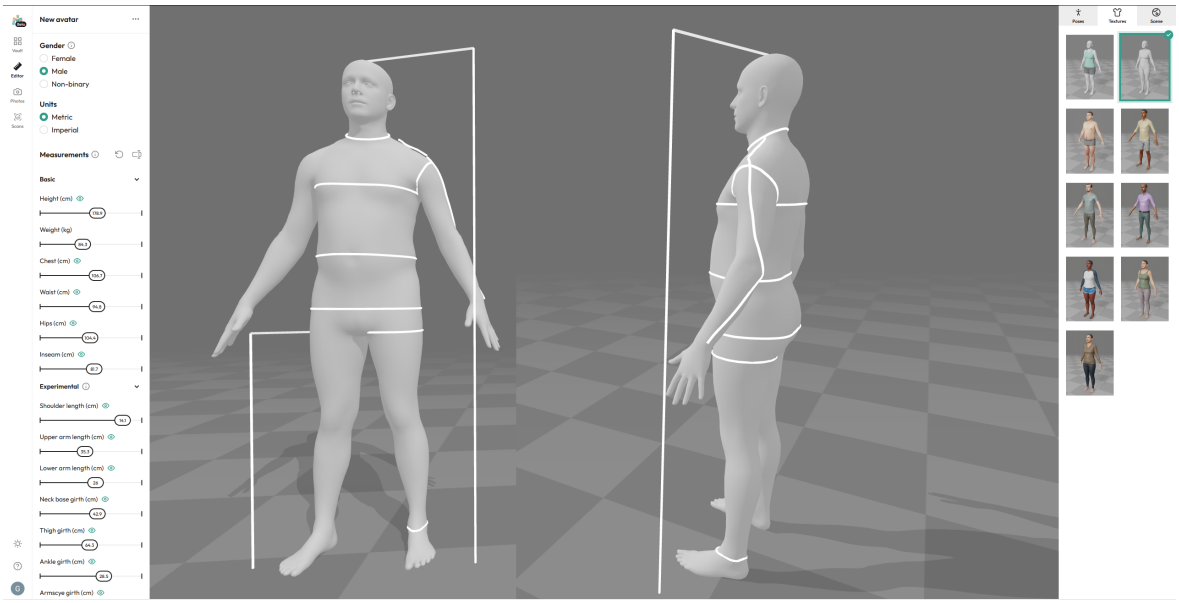


Figure 3.5: Modified image of Meshcapade Me main interface.

input data in order to get a generated avatar. The first way is to upload a 3D scans of your body. These will be subsequently processed, and the resulting avatar will be generated. The second way is to manually insert anthropometric measurements. The Digidoppel define five elemental measurements that are height, chest circumference at maximum, inseam height, shoulder breadth and waist circumference. These five basic measurements can be extended up to 18 parameters. These contain five mentioned attributes, waist height, chest circumference at scye, neck circumference at base, four measurements for arm and additional six measurements for leg.

The Meshcapade very recently introduce Meshcapade Me [10] tool. It is intended to be the successor of Digidoppel platform. It is a browser-based application exactly like Digidoppel. Although, they changed many things in comparison to its predecessor. Meshcapade Me offers three input ways of data instead of two. They preserve the first and second way from Digidoppel, so it is possibly upload 3D and manually fill input measurements. However, they significantly changed these measurements. The basic metrics are made of six different measurements instead of five. These are height, weight, chest circumference, waist circumference, hips circumference and inseam height. Except six basic measurements, Meshcapade Me offer additional seven attributes. This means that a total of 13 parameters can be entered, which is five less than in the case of its predecessor, Digidoppel.

You can see the interface of this new tool in a Figure 3.5, more precisely interface of the second mode that requires measurements for avatar generation. The white lines in the figure represent individual measurements. On the left, we can see a sidebar where

[10] `https://me.meshcapade.com`

the user can insert metrics by using the slider or manually typing. In addition, user can change sex of avatar in this part. The main part consists of 3D space where the user can see a preview of his avatar, rotate it or move. On the right, we can see the second sidebar. It offers three tabs that are poses, texture and scenes. As the name suggest, poses tab offer user to use one of preprepared poses. Texture tab is depicted in the picture and user can choose texture for his avatar. The last tab enables the user to change the background for his avatar. It is worth to note that the Figure 3.5 is adjusted in order to show the reader the avatar from two different angles in one image.

It is hard to say exactly which type of avatars Meshcapade is using since source code is not available. However, we can assume that they are using one of the SMPL avatars. This is also supported by the fact that if you want to use the SMPL, SMPL-H, SMPL-X or STAR model commercially, you need to obtain a license from Meshcapade. It is also possible that they are using a custom avatar derived from one of SMPL avatar version.

The last way of inserting data that was added, in comparison to Digidoppel, is generating avatar from 2D photos. User can upload one or more images, then these photos will be processed on Meshcapade side and user will get generated 3D avatar extracted from photos.

As you can expect, this is a paid tool. It works on the principle of credits. Creating a one avatar from photos or metrics costs 100 credits. The most expensive option is to create an avatar from 3D scans, which costs 500 credits. Surprisingly, the cost of one credit is constant regardless of the amount of credits purchased. 500 credits costs 6 € including VAT, so generating one avatar from measurements or photos costs exactly 1.20 € and one avatar generated from 3D scans will cost you 6 €.

3DLook

3DLook [11] is a company that is focused primarily on a fashion industry. They offer two solutions in this area. The first is called YourFit. This application belong to category that we refer earlier as virtual dressing room. In this case, the mobile application capture two photos, one from front view and another from side view. Then the user has to enter their height and weight into the app. Once this process is complete, the user can select a piece of clothing that they want to try on virtually. The disadvantage is that the application is now available only on Apple's iOS [12]. The second product they developed is Mobile Tailor. Same as YourFit, Mobile Tailor takes as input two photos and produces a set of 80 measurements of user body. Compared with Meshcapade Me, there are 64 more measurements produced. However, we do not have proper

[11] `https://3dlook.ai`
[12] `https://www.apple.com/ios`

information about these measurements, so it is hard to say how useful they are or if they are additionally calculated from some basic attribute subset.

3.3 Synthetic Data

As we pointed out before, synthetic data can improve training process of neural network and result in more precise and overall better final model. Using synthetic data for training a neural network has many advantages. Because we know that in order to train an accurate and reliable model, we need to provide a huge amount of training data during the training process. This could be a problem in some areas where we do not have enough data available, or it is very complicated to collect data to create a dataset of reasonable size. This could be the case in medical area when we talk for example about tomography scans or magnetic resonance images. If we want to train a neural network for a segmentation task, we will run into another problem. The problem is that in many cases the data needs to be segmented manually, which is really time-consuming and sometimes even requires experts in the field to do it. One way how to solve this issue is to teach volunteers how to segment these data. This way was used, for example by NASA [13], while collecting training data for NASA NeMO-Net convolutional neural network through their mobile application. In this research, volunteers hand-mark marine corals [13]. These issues usually does not apply to synthetic data because they are commonly generated together with their segmented images. Now, let's take a more detailed look at some researches from this area.

SURREAL

In order to increase the quality of the training process and reduce the time required for segmentation and data collection, the researchers decided to expand the currently available training data sets at the time by creating synthetic data, the researchers decided to create SURREAL. Their goal was to created as realistic data as possible. Its nickname is derived from its full name, Synthetic humans for real tasks [32].

SURREAL is a wide-ranging dataset that primarily consists from synthetic data. They generated more than six million images, including ground truth positions, depth maps and segmentation information for each frame. To represent the person in image, they used a basic SMPL model. The pipeline they created is made of several steps. The first step is to render a SMPL avatar with random pose and shape parameters. To obtain a random shape, they randomly selected one person from the CAESAR dataset [24] and approximated their body shape using SMPL, performing a similar process with the pose parameter chosen from the CMU motion capture dataset [30]. Next,

[13] `https://www.nasa.gov/`

they apply random texture to this avatar. After completing the avatar model, they render the whole image including background together with all data we mentioned. Rendered frame also have random lighting or camera position. Using this procedure, they generated data for the entire dataset.

Authors also evaluate the results they managed to achieve. They created a convolutional neural network that they trained with different training data. They compared the resulting model that was trained on only real data, only synthetic data, combination of real and synthetic data. They evaluate the model on multiple datasets like Freiburg Sitting People [19], Human3.6M [9] and MPII Human Pose dataset [2]. Evaluation of the convolutional neural network model on these datasets showed that the best results were achieved when the model was trained using a combination of synthetic and real data. These results further support what we mentioned above, that adding synthetic data to the training set can improve the resulting models.

SURREACT

SURREACT stands for Synthetic humans for real actions. As you can notice, it uses very similar naming conventions to previously mentioned SURREAL. This is caused by the fact that major part of researchers that worked on SURREACT previously worked on SURREAL. This also made SURREACT based on SURREAL. It also has a similar goal of generating synthetic data that will increase the performance of a convolutional neural network, but in this case not in the segmentation task, but in the action recognition task [31].

Action recognition is a computer vision task where, usually a convolutional neural network, is given a single photo as input. The goal is to correctly classify the activity that the depicted person performs. These can be activities such as walking, walking backwards, drinking water, sitting, standing up and many others. This can be used in smart surveillance systems that can detect if a person is carrying a gun or a knife, for example, and alert the police early enough to prevent a tragedy.

The paper is targeting to improve action recognition from unseen viewpoints by using synthetically generated data. As in SURREAL, they used the base SMPL model here as well, with ten shape parameters. For motion estimation, they used two different methods, HMMR [11] and VIBE [12]. They are used for finding motion vectors in 2D photo. After that they create SMPL avatar with random texture, lighting and shape values in the same way as they have done in SURREAL. Next, they animate the model with motion data. As a resource for motion data for action recognition, they used NTU [27] and UESTC [10] datasets. With the addition of a background, they create a dataset consisting of videos for action recognition.

Experiments were performed very similarly like they were done in SURREACT.

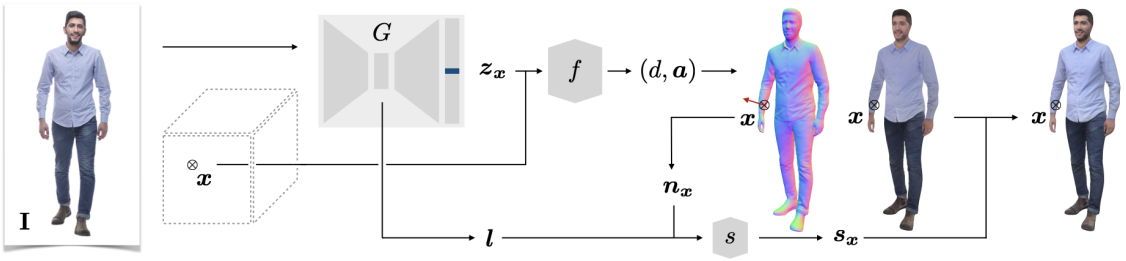


Figure 3.6: Pipeline of Google PHORUM [1].

They once again implement a model of convolutional neural network that subsequently trained on only synthetic data, real data and combination of real and synthetic data. The result was also the same as in the previous research. The best result was achieved by using a combination of synthetic and real data. Moreover, in some cases training only on synthetic data outperform model trained only on real data. They also compared motion estimation methods HMMR and VIBE and find out that VIBE produced more accurate pose estimations than HMMR, which led to better overall results.

3.4 Generative Neural Networks

Generative neural networks have a good deal of utilization in computer vision or computer graphic field. They are used for generating synthetic data, producing synthetic humans and many other uses. Sometimes they are used for creating a 3D object out of an input 2D image that is capturing a person or any other object. Now, we'll investigate further some works from this category.

PHORUM

Google [14] is the company most likely known for its search engine. Except that, it is also known for their various products like Gmail, Drive, Maps and so on. Moreover, Google is developer of Android, what is the most used operational system for mobile phones in the world. If Google wants to maintain its dominant position in the market and at the same time be one of the largest companies in the world, it must also do research in order to continue to maintain its competitive advantage. PHORUM is a method from one of Google researches in computer vision field, and we will now analyze it [1].

The goal of research is photorealistic reconstruction of humans together with their clothes from a single 2D RGB image. In order to achieve this task, they created a method called PHORUM. It uses deep neural network to achieve this goal. Their objective was to use an alternative way of generating 3D models of a person compared

[14] `https://about.google/`

Figure 3.7: Results of En3D [17].

to approaches that use a multi-camera setup followed by post-processing by an artist.

In the Figure 3.6, we can see the entire pipeline of Google PHORUM method. As an input, we have a photograph I that captures our target person. Then the feature extractor G is applied. It is a convolutional neural network with decoder-encoder architecture that produce a feature map z_x from an input picture for points in space x. Distance function f is a multi-layer perceptron with eight layers that is responsible for estimating 3D geometry of the input image. Output from it is distance d and color reflected by the surface a. The cube on the left side of the photo represent space x with its points. The l is illumination estimated from RGB input. Next, it is inputted to s together with surface normals n_x. The s is a neural network that predicts resulting shading effects s_x for the final model of human. On the right of the picture, we can see the resulting 3D model of a human produced from a single input RGB photo.

They have also done various experiments to compare performance of PHORUM with other competitors in this area. Experiments showed that Google PHORUM method outperforms its competitors in various fields.

Authors also mentioned some limitation of their method. There were issues in cases that were not covered by training set. It has wrong results for people photographed for example in loose or baggy clothes.

En3D

En3D is a generative schema for generating 3D data [17]. It uses different approach than it is usual because it does not use any real data during the training. Instead, it is using 2D synthetic data that contains life-like human models. By using these synthetic data, they manage to overcome most issues that are connected to traditional datasets with real data like image quality or diversity of resulting dataset. In this way, the model learns to correctly generalize the human body.

The pipeline of En3D is made of three major components. The first module is 3D Generative Module (3DGM). Its task is to first synthesize human pictures using known parameters of the camera, and then learn from these data in order to learn how to create a believable human model. The second component is Geometric Sculpting (GS). Its responsibility is to enhance the quality of human shape produced by 3DGM. The last component called Explicit Texturing (ET). It is the last component in their pipeline, and it is in charge of adjustments of textures.

The Figure 3.7 is depicting the output sample from En3D. As you can see, the model is capable of producing really realistic models of humans with correct geometry. In comparison to others, En3D is very flexible in terms of animating generated avatars and in experiments overcome its competitors in many areas.

Chapter 4

Research

In this principal chapter of our work, we will take a more detailed look at our research and the findings we came to. First, we look to generating synthetic data in order to improve training of convolutional neural network.

4.1 Generating Synthetic Data

As we talked about earlier in previous Chapter 3 in Section 3.3, using synthetic data can be very beneficial for training neural networks. As a part of our work, we also focused on generating synthetic data in order to improve training process of convolutional neural network.

BodyM dataset was a dataset introduced in a research paper that was released in 2022 [26]. This dataset consists of frontal and lateral silhouettes photos of people. There is a total of 8324 photos but only 2779 with anthropometric measurements of captured people. There are 14 various measurements for these people. Unfortunately, this dataset has many disadvantages. The most significant one is that there are not available original RGB photos of participants. They only make available dataset containing silhouettes in order to preserve privacy of participants. This is fully understandable, but not ideal for the research community. Another big problem is the size of a given dataset. On the one hand, it is impressive how many people they were capable to capture. On the other, the dataset is still too small to train a neural network on it. It is even worse if we would want to train a convolutional neural network with less than three thousand photos. Based on facts we stated, we decided to extend this dataset by using synthetically generated data.

In the Figure 4.1, we can see a sample from BodyM dataset. As highlighted earlier, the pictures depicted in the first column are not available in the final dataset. Next to them, we can see frontal and lateral silhouettes of people. On the left, we can see a graph that captures anthropometric measures of these two persons.

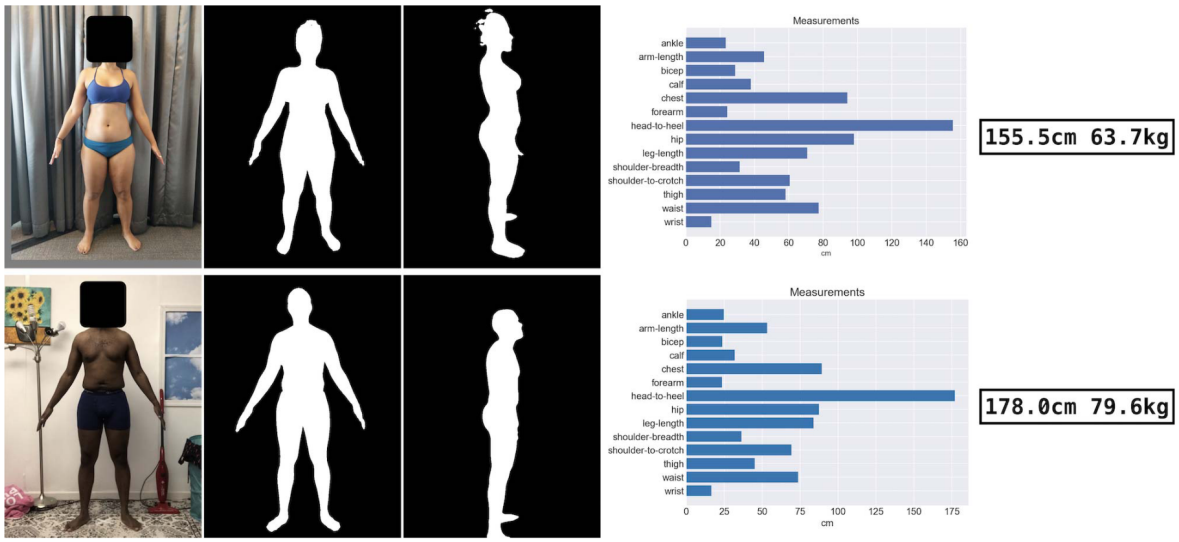


Figure 4.1: Example of BodyM dataset [26].

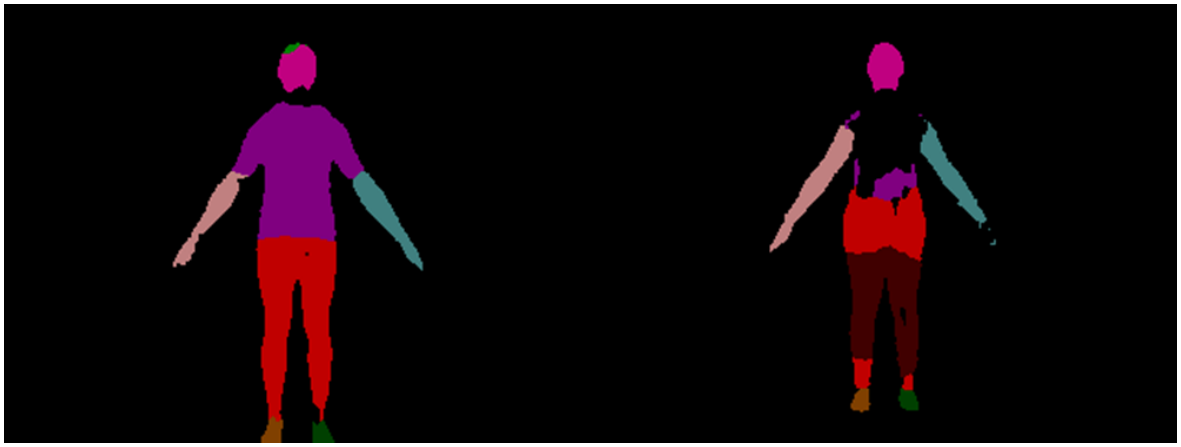


Figure 4.2: Example of using segmentation tool we chose for our research.

4.1.1 Segmentation

Segmentation is very important and highly used in computer vision. It is a process of dividing the picture into different regions based on various criteria. Basic segmentation can segment image into different part based for example on color or intensity of pixels in given region.

Semantic segmentation is a more advanced than basic segmentation. Like basic segmentation, it divides the input image into different regions, in this case based on the semantic class to which the pixel belongs. Each segment then represents one class.

During the years, scientist and researchers developed many different segmentation tools based on various ideas and focused on different areas. We will zero in on segmentation tools, sometimes also called human parsers. These are segmentation tools that are designed to segment different parts of human body. For our research, we

chose a tool called Self-correction for human parsing [14]. It comes pretrained on three different datasets, thus it can be directly use after installation. The tool is capable of distinction of left or right arm, face, hair and many more classes.

In the Figure 4.2, you can see results of segmentation using that tool performed on our synthetic data that we will closely talk about later. On the left, we can see a relatively good result of segmented parts of the body of our avatar. There are clearly visible and easily distinguishable individual arms and legs, upper and lower part of the clothes. On the right, we can see that even this tool has some limitations. In this case, avatar in input photo has the shirt with same color as the background what caused some issues for segmentation of upper part of the clothes.

4.1.2 SURREACT implementation

How we used the source code of SURREACT to generate data for our purpose.

4.1.3 Our Data

Description of our generated data we get from adjusted implementation of SURREACT

4.2 Anthropometric Avatar Generator

Description of how we transform the input anthropometric measurements to SMPL beta parameters

4.3 Augmenting 3D Avatars

4.3.1 Coloring Avatars Meshes

Coloring meshes of avatar based of segmentation we get before.

4.3.2 Adding clothes to avatars ??

4.3.3 Adding hair to avatars ??

Chapter 5

Specification

Chapter 6

Implementation

Chapter 7

Evaluation

Conclusion

Bibliography

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[3] Michael Black. Smpl made simple. `https://youtu.be/rzpiSYTrRU0?si=KpMrt2rhZSmihhAf`, 2021. Last accessed 20 November 2023.

[4] David Boja. SMPL Anthropometry. `https://github.com/DavidBoja/SMPL-Anthropometry`, 2023. Last accessed 31 December 2023.

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[6] Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3d body shape regression using metric and semantic attributes, 2022.

[7] Unreal Engine. How to use metahuman animator in unreal engine. `https://youtu.be/WWLF-a68-CE?si=gZ0_Cn8fhdtEAIPS`, 2023. Last accessed 31 December 2023.

[8] Virtual Humans. Unreal Engine's MetaHuman Creator, analyzed and explained. `https://www.virtualhumans.org/article/unreal-engines-metahuman-creator-analyzed-and-explained`, 2021. Last accessed 31 December 2023.

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural

environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[10] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *Proceedings of the 26th ACM international Conference on Multimedia*, pages 1510–1518, 2018.

[11] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019.

[12] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[13] Alan S. Li, Ved Chirayath, Michal Segal-Rozenhaimer, Juan L. Torres-Pérez, and Jarrett van den Bergh. Nasa nemo-net's convolutional neural network: Mapping marine habitats with spectrally heterogeneous remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5115–5133, 2020.

[14] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[15] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[17] Yifang Men, Biwen Lei, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. *arXiv preprint arXiv:2401.01173*, 2024.

[18] Meshcapade. Smpl: A skinned multi-person linear model. `https://meshcapade.wiki/SMPL`, 2021. Last accessed 20 November 2023.

[19] Gabriel L Oliveira, Abhinav Valada, Claas Bollen, Wolfram Burgard, and Thomas Brox. Deep learning for human part discovery in images. In *2016 IEEE International conference on robotics and automation (ICRA)*, pages 1634–1641. IEEE, 2016.

[20] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[21] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[22] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H. Bülthoff, and Michael J. Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3D measurements. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1887–1897, 2019.

[23] Andrey N. Rabchevsky and Leonid N. Yasnitsky. The role of synthetic data in improving neural network algorithms. In *2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, pages 316–320, 2022.

[24] Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian american and european surface anthropometry resource (caesar), final report, volume i: Summary. *Sytronics Inc Dayton Oh*, 2002.

[25] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[26] Nataniel Ruiz, Miriam Bellver, Timo Bolkart, Ambuj Arora, Ming C. Lin, Javier Romero, and Raja Bala. Human body measurement estimation with adversarial augmentation, 2022.

[27] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[28] Unity Technologies. Synthetic Homes. `https://github.com/Unity-Technologies/SyntheticHomes`, 2022. Last accessed 30 December 2023.

[29] Unity Technologies. Synthetic Humans package (Unity Computer Vision). `https://github.com/Unity-Technologies/com.unity.cv.synthetichumans`, 2022. Last accessed 30 December 2023.

[30] Carnegie Mellon University. Cmu graphics lab motion capture database. `http://mocap.cs.cmu.edu/`. Last accessed 4 January 2024.

[31] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. In *IJCV*, 2021.

[32] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.