Human Body Dimension Estimation from Occluded Images

Marc Beiwinkler Martin Krimpelstätter Ilpo Viertola Thomas Wulz

***Abstract*—Human shape estimation is a growing field. Its applications include, for example, defining dress size from an image of a person. With *Human Body Dimensions Estimation* (HBDE) we refer to estimating certain body measurements, like arm lengths or inseam length, from images. Our aim is to research how accurately a convolutional neural network (CNN) can predict human body dimensions from a 2D image of a 3D human body mesh that has been partly occluded. We call such a network a Neural Anthropometer. The images used in this research are artificially generated 3D human body meshes, converted to 2D images and occluded with black rectangles with varying locations and sizes. The data set used for training and testing consists of such images along with the correct human body dimensions. We tested the network with occluded but also with non-occluded images. With non-occluded images, our network was able to estimate the body dimensions with a mean relative error of 6.41%. On the other hand, the network was able to predict the body dimension with a mean relative error of 3.87% from occluded images. Training the same network with non-occluded images resulted in good estimation performance of the body dimensions from non-occluded images, with a mean relative error of 2.17%. On the other hand, this model was not able to estimate the human body dimensions from images containing occlusions.**

I. Introduction

This work addresses the problem of estimating *Human Body Dimensions* (HBDs), such as arm length, height of the person or inseam length, from artificially generated images of humans using a *Convolutional Neural Network* (CNN).

It is based on the paper *A Neural Anthropometer Learning from Body Dimensions Computed on Human 3D Meshes* by *H. A. Mayer and Y. G. Tejeda* [1]. The data set used to train and test the network is the same as in the article. For our experiments, we produced a second version of the data set, from which we occluded the images using black rectangles to evaluate the network's performance on. Details about the method used for occlusion are given in chapter IV.

II. Problem Description

The goal of our research was to determine how the original Neural Anthropometer [1] would perform when instead of introducing clean, non-occluded images to it, we would test its performance against occluded images of human body meshes. Also, we wanted to test how the network's performance differed, when we trained it using only occluded or only non-occluded images. In both cases we tested the trained networks with occluded and non-occluded images.

Another was to simplify the model, but still achieve the same results with non-occluded training data set and with non-occluded test data set. One way to do this was to reduce the amount of HBDs estimated by the network.

III. Brief Summary of Original Paper

The paper [1] that this work is based on introduced a novel approach for determining human body dimensions using a convolutional neural network for deep regression. Instead of using images of real humans as the data set, the Skinned Multi-Person Linear [4] model was used to generate human body 3D meshes. Using body landmarks and the bounding box, several HBDs are computed on these models. Namely: shoulder width, right and left arm length, inseam; chest, waist and pelvis circumference and height. 2D images are synthesized from the 3D models, and these 2D images are then used as the input for the neural network, along with the calculated HBDs as targets.

In the original paper, a convolutional neural network is trained using these images. After training the network, its performance is tested using common evaluation measures described in the chapter VI-A.

IV. Occlusion

To perform the occlusion on the data set, we added a single randomly generated black rectangle to each image. We did this by first generating the coordinates of the upper left corner using a uniform distribution over the complete image. To ensure that we would end up with a large amount of images where these rectangles have some overlap with the body, instead of just being besides it, we used experimentally determined lower and upper bounds for these coordinates depending on the sex and pose of the body. The bottom right corner of the rectangle is then chosen such that the height and width of the rectangle are between 10 and 50 pixels, which makes the rectangle cover anywhere between 0.25% and 6.25% of the image. It should be noted that this is only an upper bound for the occluded portion of the human and does not guarantee that the human is occluded at all. An example of an occluded image is shown below.

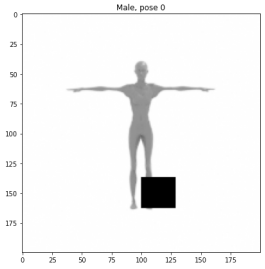


Fig. 1. Example for an occluded image.

In this example, the rectangle covers up the left shin of the person in the image. In theory, this should not interfere with any of the body dimensions the network is trained to estimate, as the height can still be inferred from the other leg. Therefore, the accuracy of the estimation performed by the network on this image should not deteriorate because of the occlusion.

V. Approach

Our approach is widely based on the approach represented in the paper [1]. We use the same data set as in the paper, with a difference that we occlude the images with rectangles in a way described above. The code is available at www.github.com/ilpoviertola/pattern_recognition_2

A. Model Architecture

As stated before, we used a similar approach to the problem as used in [1]. Thus, the model architecture also mimics the one in the paper. We implemented the network using PyTorch [2]. The input layer takes a gray scale image of size `200x200x1`. This image contains a synthetic image of a human standing. This first layer processes the image with six, a five-by-five convolutional kernels. The layer outputs a feature map of size `196x196x6`. The feature map differs from its counterpart described in the paper in its last dimension, since we only chose six human body dimensions to estimate. One channel per estimated dimension is used. Our guess is that this will result in a good performance. Next we used a *Leaky Rectified Linear Unit* (LeakyReLU) as an activation function. LeakyReLU is chosen to avoid the Dying ReLU-problem which is a form of vanishing gradient-problem. This can occur during the learning process of the network. In this dead state, no gradients are flowing back through the network, and training will stall. Neurons of the network will get stuck in some state as their input weights will not get adapted any further [3]. Because we did not perform any normalisation of the data before feeding it into the network, batch normalisation is applied after the LeakyReLU layer. This is achieved with a batch normalisation layer. The resulting tensor is an input to a max pooling layer with a square kernel of size five and with a stride length of two.

After max pooling, a second two-dimensional convolutional layer is applied with a kernel size of five-by-five and with twelve output channels. This is followed by a LeakyReLU and a max pooling layer similar to that described earlier. After the max pooling, the tensor is flattened. This results in a vector of size 23232. The vector is then fed to a classifying part of the network, which outputs the estimations for the body dimensions.

For the classifying part of the model, we used two fully connected layers with the first containing a LeakyReLU activation function. The first fully connected layer takes an input with 23232 features and then reduces the feature amount down to 512. This is then passed to the LeakyReLU, which passes the vector on to the second fully connected layer. This final layer then takes this vector with 512 features and reduces the feature size down to six, which is the amount of body dimensions we were estimating. After experiments with a higher number of features we reduced the amount of features between the two fully connected layers in order to achieve a network with reduced amount of parameters. This measure significantly reduced the amount of time needed to train the model and the amount of storage space needed to save its parameters after training.

B. Model Training

We trained the model using occluded but also non-occluded images. Training and testing with non-occluded images yielded similar results as in the paper, so the networks can be considered equally efficient. In following chapter VI, we are going to explain the way the model was trained more precisely.

VI. Experiments and Results

One important goal in design, execution and evaluation of our experiments was the comparability with the experiment and the outcome described in the paper *A Neural Anthropometer Learning from Body Dimensions Computed on Human 3D Meshes* [1]. From the eight HBDs estimated in the original paper we chose six. These were shoulder width, right and left arm length, height, chest circumference and pelvis circumference. By leaving out inseam and waist circumference we aimed to reduce the complexity of our model in comparison to the original model. Our main goal was to find out, if a model resembling the Neural Anthropometer [1] was able to estimate those HBDs from images showing persons partially occluded by black rectangles of a certain size. We conducted two different experiments:

- Experiment 1: Train the model using **non-occluded** images and test it against non-occluded and occluded ones.
- Experiment 2: Train the model using **occluded** images and test it against non-occluded and occluded ones.

The data set used to train and test the model during the first experiment was similar to the one used to train the original Neural Anthropometer [1]. For the second experiment, we occluded every image with a single rectangle. Thus both data sets contain the same set of images, the first **without occlusions** and the second **with occlusions**. The structure and origin of this image set shall be described as follows.

It comprises 12000 images produced from 3D human body meshes originating from 6000 subjects generated with the *Skinned Multi-Person Linear* (SMPL) generative human body model Loper et al. (2015) [4]. These images show 3000 female and 3000 male subjects, where each subject is depicted in two different poses. Thus, the image set consists of 12000 synthetic gray-scale images of size `200x200x1`. For each of the 6000 subjects ground truth values for the eight HBDs named above are given. These values have been determined using methods that are described in the original paper [1]. We chose six (shoulder width, right and left arm length, height, chest circumference and pelvis circumference) from these eight HBDs.

In analogy to the original training and evaluation method [1], we trained and evaluated the network with the whole data set using k-fold cross-validation with k equal to five. During each of the five iterations we trained the model for 20 epochs using mini-batches containing 100 images each. We minimized *Minimum Squared Error* (MSE) between ground truth and estimated HBDs using stochastic gradient descent with a momentum of 0.9. Compared to the learning rate of 0.01 used in the original experiment [1], we chose to decrease the learning rate to 0.001. We ended up alternating it because we had experienced numerical and dying ReLu problems [3] using the same learning rate as in the original experiment. For training the model, we utilised Google Colab in order to optimise the time needed to train the model with GPU acceleration. Training the model, using 80% of the whole data set as training data, took 10 to 20 minutes. The time needed to train the model did not depend on whether occluded or non-occluded images were used.

To evaluate the network's performance on occluded images in the first experiment, we conducted additional tests after each of the five training phases. The data samples used for these additional tests comprised the same images contained in the original, non-occluded test samples used for this experiment, but with every image additionally containing one black rectangle. To evaluate the network's performance on non-occluded images in the second experiment, we conducted the same additional tests like in the first experiment. The data samples used for these additional tests comprised the same images contained in the original test samples.

A. Quantitative Evaluation Formulas

The formulas and identifiers shown in this section have been adopted to a great extent from the original paper [1] and are given to explain how the statistics presented in this section have been derived from the data resulting from model evaluation.

As stated above, for model evaluation we used k-fold cross validation (with $k = 5$), i.e., during each iteration j, $a = 2400$ instances were used for evaluation. We used our model to infer $n = 6$ HBDs, resulting in each of our two experiments producing two vectors. One vector contained data resulting from testing the model with non-occluded images, whereas the second vector contained data resulting from testing the model with their occluded counterparts. Both vector had the shape

$$k \times a \times |\{\hat{D}_i, D_i\}| \times n = \texttt{5 x 2400 x 2 x 6}. \quad (1)$$

The estimation error introduced by the model is represented by two measures. The first measure is the Mean Absolute Difference (MAD) e^i_{MAD} between estimated and given value for each HBD indexed by i, i.e. $\hat{D}_i$ and D_i, over the k iterations. Variable a denotes the number of estimations per a fold. Here $a = 2400$.

$$e^{ij}_{MAD} = \frac{1}{a}\sum_{l=1}^{a}|\hat{D}^l_i - D^l_i|, \quad e^i_{MAD} = \frac{1}{k}\sum_{j=1}^{k} e^{ij}_{MAD} \quad (2)$$

This measure is complemented by its average, Average Mean Absolute Difference e_{AMAD} (AMAD), over all $n = 6$ HBDs,

$$e_{AMAD} = \frac{1}{n}\sum_{i=1}^{n} e^i_{MAD} \quad (3)$$

The second measure used to illustrate the estimation error is the Relative Percentage Error (RPE) e^i_{RPE} for each HBD D_i

$$e^{ij}_{RPE} = \frac{1}{a}\sum_{l=1}^{a}|\frac{\hat{D}_l - D_l}{D_l}|, \quad e^i_{RPE} = \frac{1}{k}\sum_{j=1}^{k} e^{ij}_{RPE} \quad (4)$$

Together with RPE, we also use its average, Average Relative Percentage Error e_{ARPE} (ARPE). It is calculated over all $n = 6$ HBDs

$$e_{ARPE} = \frac{1}{n}\sum_{i=1}^{n} e^i_{RPE} \quad (5)$$

With the measures defined above we are able to get comparable results with the original *Neural Antrophometer* (NA) [1]. In the following subchapter, we will present our results utilising these measures.

B. Experiments and Discussion

Table I and table II show the results of our first experiment, in which our model has been trained on non-occluded images only. From table I showing results from tests with non-occluded images it can be seen that our model introduces smaller errors than the original Neural Antrohpometer for all estimated HBDs. In particular, height (0.94% and 1.58%, respectively), pelvis circumference (1.49% and 2.40%, respectively) and chest circumference (2.00% and 2.51%, respectively) show lower estimation errors. Both models show the highest relative estimation error (4.51% and 4.93%, respectively) in their predictions of the shoulder width. Both models estimate the height with the lowest relative estimation error (0.94% and 1.58%, respectively). The difference between the prediction errors for left and right arm length, that have been already closer to each other than to any other estimation error for the original NA (2.34% and 2.22%, respectively), was even smaller for our model (2.09% and 2.00%, respectively). On average, our model was able to predict the six HBDs with a relative error of 2.17%, whereas the original NA introduced an error of 2.66%. We think, that our model's improved estimation performance in comparison to the original NA has been caused by the reduced number of HBDs our model had to estimate. A reduction of the number of outputs by one fourth compared to the original Neural Anthropometer's eight outputs could have increased the effectiveness of the MSE over all outputs used as a learning criterion for both models.

HBD	**MAD**		**RPE**	
	Ex. 1	***NA***[a]	***Ex. 1***	***NA***[a]
Height	16.14	27.34	0.94	1.58
Shoulder width	10.41	12.54	4.51	4.93
Left arm length	11.96	13.48	2.09	2.34
Right arm length	11.57	12.98	2.00	2.22
Pelvis circumference	15.80	25.85	1.49	2.40
Chest circumference	20.60	25.22	2.00	2.51
AMAD	14.41	19.57		
ARPE			2.17	2.66

[a] All values for the original NA except AMAD and ARPE were taken from [1].

TABLE I

ESTIMATION ERRORS INTRODUCED BY OUR MODEL AFTER TRAINING WITH NON-OCCLUDED IMAGES DURING FIRST EXPERIMENT AND BY ORIGINAL NEURAL ANTHROPOMETER (NA), BOTH MODELS TESTED WITH NON-OCCLUDED IMAGES. DISPLAYED IS THE MAD FOR EACH HBD (FIRST COLUMN) OVER FIVE FOLDS (SECOND COLUMN), IN COMPARISON WITH THE MAD FOR THE SAME HBD OVER FIVE FOLDS CAUSED BY THE NA DESCRIBED IN THE ORIGINAL PAPER [1] (THIRD COLUMN). MADS ARE EXPRESSED IN MILLIMETERS. THE FOURTH AND THE FIFTH COLUMN SHOW THE RPE FOR EACH HBD INTRODUCED BY OUR MODEL AND BY THE ORIGINAL NA, RESPECTIVELY. THE LAST TWO ROWS DISPLAY AMADS AND ARPES OVER ALL HBDS AND FOLDS FOR BOTH MODELS.

From table II it is evident that our model can not be used to estimate HBDs from images containing partially occluded humans after having been trained on non-occluded images only. Relative errors introduced for each of the six HBDs were larger than 100%, if estimations were based on partially occluded images; relative error for shoulder width, for which the model had already shown the highest relative error when predicting HBDs from non-occluded images, was even above 150%. Correspondingly, the average relative error over all six HBDs increased by more than the 60 times (2.17% and 134.34%, respectively).

HBD	**MAD**		**RPE**	
	Nonocc.	***Occ.***	***Nonocc.***	***Occ.***
Height	16.14	2.02e3	0.94	118.77
Shoulder width	10.41	548.78	4.51	152.96
Left arm length	11.96	733.99	2.09	128.97
Right arm length	11.57	791.88	2.00	137.67
Pelvis circumference	15.80	1.41e3	1.49	137.30
Chest circumference	20.60	1.30e3	2.00	130.36
AMAD	14.41	1.14e3		
ARPE			2.17	134.34

TABLE II

ESTIMATION ERRORS INTRODUCED BY OUR MODEL IN TESTS WITH NON-OCCLUDED AND OCCLUDED IMAGES AFTER TRAINING WITH NON-OCCLUDED IMAGES DURING FIRST EXPERIMENT. DISPLAYED IS THE MAD FOR EACH HBD (FIRST COLUMN) OVER FIVE FOLDS IN A TEST WITH NON-OCCLUDED IMAGES (SECOND COLUMN), IN COMPARISON WITH THE MAD FOR THE SAME HBD OVER FIVE FOLDS IN A TEST WITH OCCLUDED IMAGES (THIRD COLUMN). MADS ARE EXPRESSED IN MILLIMETERS. THE FOURTH AND THE FIFTH COLUMN SHOW THE RPES FOR EACH HBD CAUSED BY THE MODEL IN THESE TESTS. THE LAST TWO ROWS DISPLAY AMADS AND ARPES OVER ALL HBDS AND FOLDS FOR BOTH TESTS.

Table III and table IV show the results of our second experiment, during which our model has been trained on occluded images only. Table III represents results from tests with non-occluded images. Using a training set consisting of occluded images resulted in worse HBD estimations from non-occluded images, compared to the original NA, which had been trained with non-occluded images. Relative estimation errors for most HBDs ranged around 6.5%, with relative estimation errors for pelvis and chest circumference 6.09% and 6.15%, respectively) lying a bit lower than the other errors. On average, our model increased its relative error almost three times (2.17% and 6.41%, respectively). The original Neural Anthropometer introduced an error of 2.66% on the same task.

HBD	MAD		RPE	
	Ex. 2	*NA*[a]	*Ex. 2*	*NA*[a]
Height	107.75	27.34	6.33	1.58
Shoulder width	19.94	12.54	6.76	4.93
Left arm length	38.26	13.48	6.64	2.34
Right arm length	37.42	12.98	6.47	2.22
Pelvis circumference	62.21	25.85	6.09	2.40
Chest circumference	62.39	25.22	6.15	2.51
AMAD	54.66	19.57		
ARPE			6.41	2.66

[a]All values for the original NA except AMAD and ARPE were taken from [1].

TABLE III

ESTIMATION ERRORS INTRODUCED BY OUR MODEL AFTER TRAINING WITH OCCLUDED IMAGES DURING SECOND EXPERIMENT AND BY ORIGINAL NEURAL ANTHROPOMETER (NA), BOTH MODELS TESTED WITH NON-OCCLUDED IMAGES. DISPLAYED IS THE MAD FOR EACH HBD (FIRST COLUMN) OVER FIVE FOLDS (SECOND COLUMN), IN COMPARISON WITH THE MAD FOR THE SAME HBD OVER FIVE FOLDS CAUSED BY THE NA DESCRIBED IN THE ORIGINAL PAPER [1] (THIRD COLUMN). MADS ARE EXPRESSED IN MILLIMETERS. THE FOURTH AND THE FIFTH COLUMN SHOW THE RPE FOR EACH HBD INTRODUCED BY OUR MODEL AND BY THE ORIGINAL NA, RESPECTIVELY. THE LAST TWO ROWS DISPLAY AMADS AND ARPES OVER ALL HBDS AND FOLDS FOR BOTH MODELS.

Table IV shows estimation results after training the model with occluded images. Test results achieved on non-occluded images are compared to those achieved on occluded images. Using occluded training data yields a much more balanced estimation performance between occluded and non-occluded images compared to the model's performance after being trained on non-occluded images. Nevertheless the numbers in the table prove that our model achieves much better results on occluded images than on non-occluded images when trained on occluded images. Relative estimation errors for height (1.79%), left and right arm length (3.77% and 3.62%, respectively) as well as pelvis and chest circumference (3.02% and 3.86%, respectively) are lower for occluded than for non-occluded images. For height (1.79%) a value could be reached that is comparable to that caused by the NA on non-occluded images (1.58%). In contrast to that, relative estimation errors for shoulder width on non-occluded as well as occluded images (6.76% and 7.19%, respectively) are higher than any other error. Whereas the overall error values caused by our model after having been trained on occluded images only will almost certainly be too large for an application like distance tailoring, they might still be acceptable in other areas of application, e.g. for determining a person's risk of being overweight to an extent relevant for her or his health.

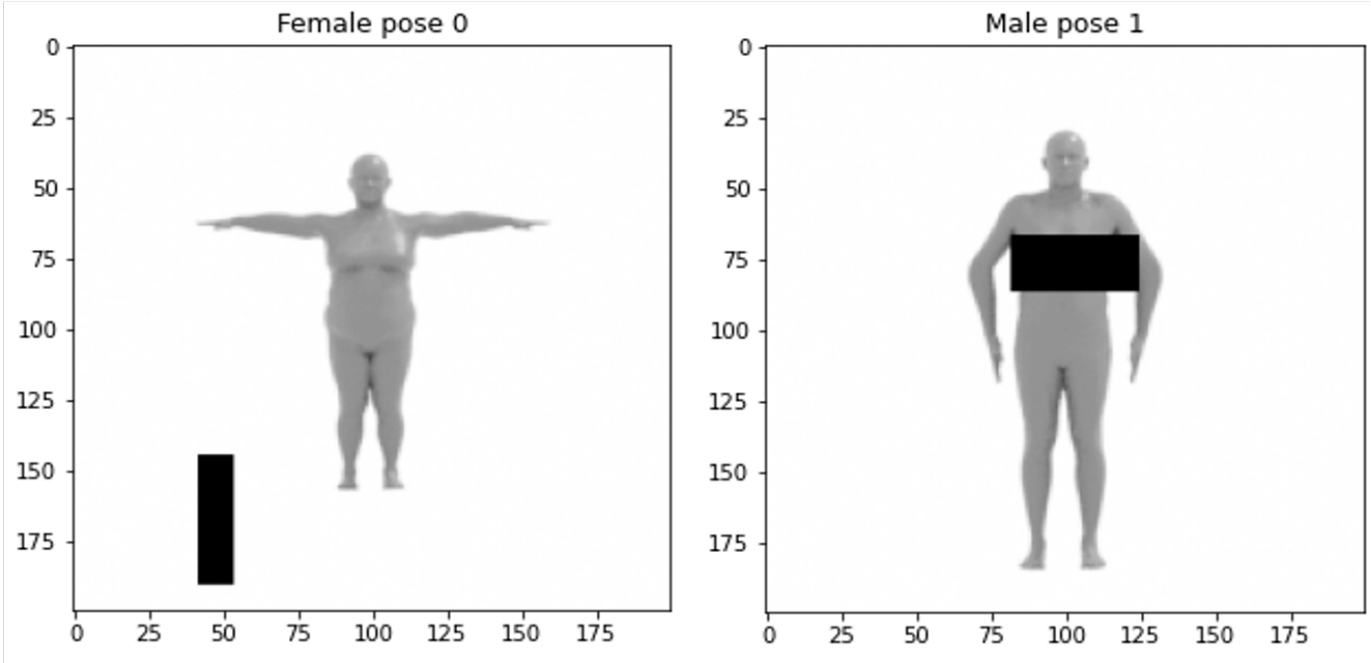
HBD	MAD		RPE	
	Nonocc.	*Occ.*	*Nonocc.*	*Occ.*
Height	107.75	30.64	6.33	1.79
Shoulder width	19.94	19.85	6.76	7.19
Left arm length	38.26	21.51	6.64	3.77
Right arm length	37.42	20.88	6.47	3.62
Pelvis circumference	62.21	31.32	6.09	3.02
Chest circumference	62.39	39.33	6.15	3.86
AMAD	54.66	27.25		
ARPE			6.41	3.87

TABLE IV

ESTIMATION ERRORS INTRODUCED BY OUR MODEL IN TESTS WITH NON-OCCLUDED AND OCCLUDED IMAGES AFTER TRAINING WITH OCCLUDED IMAGES DURING SECOND EXPERIMENT. DISPLAYED IS THE MAD FOR EACH HBD (FIRST COLUMN) OVER FIVE FOLDS IN A TEST WITH NON-OCCLUDED IMAGES (SECOND COLUMN), IN COMPARISON WITH THE MAD FOR THE SAME HBD OVER FIVE FOLDS IN A TEST WITH OCCLUDED IMAGES (THIRD COLUMN). MADS ARE EXPRESSED IN MILLIMETERS. THE FOURTH AND THE FIFTH COLUMN SHOW THE RPES FOR EACH HBD CAUSED BY THE MODEL IN THESE TESTS. THE LAST TWO ROWS DISPLAY AMADS AND ARPES OVER ALL HBDS AND FOLDS FOR BOTH TESTS.

In figure 2 below, we present two examples from the occluded data set together with the CNN's estimations. Same estimation errors are used as in the tables before. In the image on the left side, the occlusion is not covering the subject. Still, the CNN can not produce as good estimations as it can for the subject on the right side of the figure. Also, the errors between arm lengths are relatively large with the female subject.

These results imply clearly that the model can predict HBDs the best from the type of images that it has seen before. In other words, if the model is trained with occluded images, it will generally be more precise estimating HBDs from occluded images than from non-occluded images, and vice versa. This provokes an idea for further investigation. The model could be trained with a data set containing occluded and non-occluded images with some weighting. A further direction of investigation could be to examine the effects of location and size of the occluding shape in an image on the model's estimation performance. The use of realistic backgrounds, together with occluding objects, showing a more realistic variability in shape and texture, could increase the practical relevance of the insights achieved by such research.



HBD	MAD (mm)	RPE (%)
Height	16.4	1.1
Shoulder width	21.7	5.57
Left arm length	4.52	0.93
Right arm length	34.91	7.55
Pelvis circumference	66.16	5.73
Chest circumference	72.55	5.93

HBD	MAD (mm)	RPE (%)
Height	7.53	0.39
Shoulder width	13.47	2.95
Left arm length	15.07	2.31
Right arm length	6.88	1.04
Pelvis circumference	27.47	2.35
Chest circumference	9.14	0.76

Fig. 2. Comparison between two occluded images and their body dimension estimations.

C. Summary

We trained a neural network to estimate various human body dimensions on synthesized images of human bodies. For training and testing, we used both the initial (non-occluded) data set and an occluded data set that we generated by adding randomly generated rectangles. Using the non-occluded data set for training and testing yielded great results, as did using the occluded data set for both. The network trained with the occluded data set performed worse when tested with non-occluded data, while the network trained with non-occluded data was unable to produce meaningful estimates when tested on occluded images.

REFERENCES

[1] H. A. Mayer and Y. G. Tejeda, "A Neural Anthropometer Learning from Body Dimensions Computed on Human 3D Meshes", unpublished.

[2] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep

[3] K. Leung, "The Dying ReLU problem, Clearly Explained", [Online]. Available: https://towardsdatascience.com/the-dying-relu-problem-clearly-explained-42d0c54e0d24

[4] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model" *ACM Trans. Graph.*, vol. 34, pp. 248:1–248:16, 2015.