Human-shape classification with artificial backgrounds

Armin Niedermueller[1], Melanie Urban[1] and Ahmet Cihat Bozkurt[1]

***Abstract*— Human Body Dimensions (HBD) Estimation using images becomes increasingly important. Datasets consisting of 2d images of humanoids have been successfully generated and then used for training such networks. However, those images all have a single coloured grey background. The HBD estimation of such images generates very good results, but what if an artificial background is introduced into the dataset and the humans are visually not easily separable? In this work, we introduce random coloured backgrounds as well as textured backgrounds into an existing dataset of humanoids. We create and train a modified network architecture and then calculated the HBDs. Our modified network achieved far better results ($\sim 50\%$ improvement) as in the original paper, even with the texture and background dataset.**

I. INTRODUCTION

Calculating human dimensions from 3D human meshes is possible, but that is another research topic. In this study, we tried to develop a model that predicts body measurements from images of humans and monsters in different poses and related body measurements. Therefore we used the architecture of the *Neural Anthropometer* (NA) as basis and also the provided dataset [7].

Since the model trained with uniform background images will have a narrow usage area, we added artificial backgrounds to the images in order to train a model that can make accurate predictions in different scenarios. This setting is much more realistic and can be used in a variety of application areas. For example an online shop could provide a tool for uploading an image by oneself and afterwards announce the dress size.

The background we add are either a random color or texture images. We tested our model with plain images of humans or monsters and respectively images with RGB background or with texture background. We evaluate the results by calculating the average model accuracy, MAD (Mean Absolute Deviation) for each body dimensions and RPE (Relative Percentage Error) for each body dimensions.

In this paper will first provide an overview of related work and the used dataset. Then we introduce our model and show the differences to the original Neural Anthropometer. Afterwards, we evaluate and discuss the results of testing our model.

[1]Pattern Recognition II, Paris Lodron Universität Salzburg, armin.niedermueller@mailbox.org
melanie.urban@stud.sbg.ac.at
s1079921@stud.sbg.ac.at

II. HUMAN BODY DIMENSION ESTIMATION

There are plenty ways of estimating human body dimensions and just as many different input types. We orientated our model on the descriptions made in the paper *A Neural Anthropometer Learning from Body Dimensions Computed on Human 3D Meshes* [7]. They used images of 3D meshes synthesized with the Skinned Multi-Person Linear Model (SMPL) [5] together with the calculated HBDs as ground truth.

But our NA is modified and outperforms the original architecture by far. Our code is publicly available at https//github.com/nerovalerius/humanoids_cnn and the code of the original NA and also the dataset at: https://github.com/neoglez/neural-anthropometer

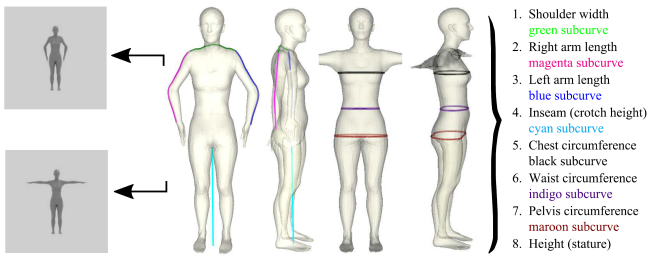


Fig. 1. Method of generating 2D images from 3D meshes which was used in the original paper [7]. On the right side, the HBDs are listed. While the humanoids are stored as 2D grey scale images, the HBD are provided as json data.

A. Dataset

The used dataset consists of 12.000 grey-scale 2D images of humans and 2.000 additionally provided 2D images of monsters, with a size of 200 x 200 each.

The underlying 3D meshes were formed by providing shape parameters to the SMPL. To get bodies with a huge diversity of measurements the parameters were uniformly varied. [7] But there is also the possibility to end up with so-called monster shapes. We also tested these monsters, however, separately from the humans.

Two different poses are used, pose 0 and pose 1 (see Fig. 2). Pose 0 shows the humanoid with the arms stretched out to both sides and pose 1 shows the hands hanging down. Furthermore, the dataset consists of females and males. We have 3500 females in pose 0, 3500 females in pose 1, 3500 males in pose 0, 3500 males in pose 1.

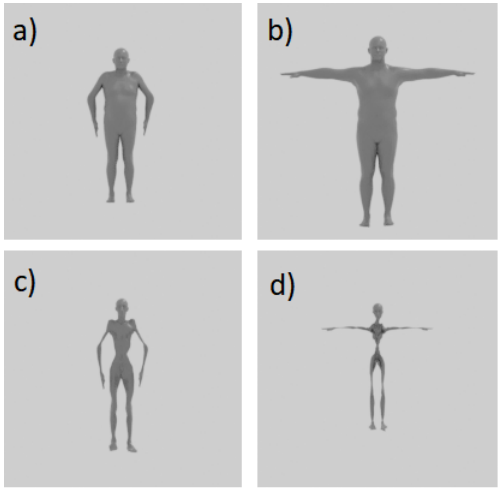


Fig. 2. Different Humanoids from the original dataset [7]. a) shows a human male in pose 0. b) shows a human male in pose 1. c) shows a monster male in pose 0. c) shows a monster female in pose 1.

Furthermore, the dataset provides the correct annotations for the HBDs of each image in order to let our network be able to learn. The eight HBDs are as follows:

Shoulder width	Right arm length
Left arm length	Inseam (crotch height)
Chest circumference	Waist circumference
Pelvis circumference	Height (stature)

B. Artificial Backgrounds

We augment the original gray scale dataset by adding two types of backgrounds and thus generating two new datasets. For the first new dataset, we replaced the single color background by inserting a background where all pixels have the same random RGB value. At the second dataset, we replace the background with textures of the DTD-dataset, which consists of 5640 different images [3]. Since our dataset consists of 14.000 images, the DTD-images will be used multiple times. However, this should not affect the results in any way, when a DTD-image is used twice, it is attached to a different humanoid. A subset of the resulting images are shown in Fig. 3.

At the end of the process, we have 3 different datasets: human/monster images with the original gray background, a second data set with grey texture backgrounds and the third dataset with random grey scale backgrounds.

III. APPROACH

A. Pre-Processing

To keep the overall CNN architecture simple we restrict the input channels to 1. Since we introduced RGB backgrounds to the datasets in Section II-B, we converted them back to grey scale before training using OpenCV [2].

Fig. 3. Artificial Backgrounds which are introduced into the original dataset. Row 1 and 2 show texture backgrounds from the DTD-dataset

B. Network Architecture

Our neural network is shown in Fig. 4 and implemented with Pytorch [6]. We used the original architecture as in [7] which is described as follows:

The input layer processes 200 x 200 x 1 images. A 2D-convolution with a 5-pixels square kernel produces a feature map of size 196 x 196 x 8, which is afterwards passed through a rectified linear unit (ReLU) [1] and then is batch normalized. After a max pooling with stride 2, a second 2D-convolution is applied (again squared kernel of size 5), which results in a tensor of size 94 x 94 x 16. After another max pooling, the output is flattened to a tensor of size 35344. At last, this tensor is passed to a fully connected layer and trough a ReLU. The last layer is a regressor that outputs the eight human body dimensions in meters [7, p. 6].

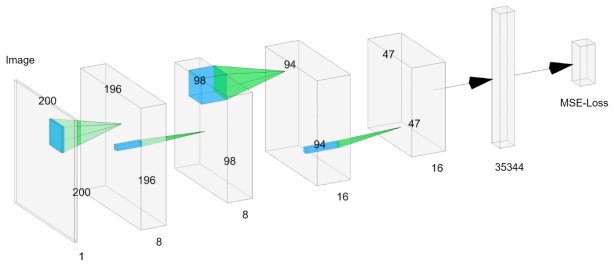


Fig. 4. Our CNN architecture based on the original NA framework [7]

Once we implemented our network and selected hyperparameters providing the best results, we also compared our model to the meanwhile published original Neural Anthropometer. There are several differences which lead to better results:

LayerNorm

The original NA uses a BatchNorm2d-Layer [4] right after the output of the first convolution is passed through the ReLU-Layer. We instead normalize with LayerNorm [9]. As shown in Fig. 5 the BatchNorm normalizes the

tensor across the batch-size and the spatial dimensions for each channel while the LayerNorm uses for normalization the values computed across all channels for each sample. Because at this stage of the NA we already have eight different channels it might be useful to normalize per channel. But these channels were derived from a single input image without channels at all in the first place. If the eight derived channels do not include all the informations needed to compute accurate statistics for the final eight output dimensions a layer normalization might handle this.

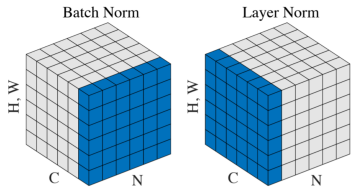


Fig. 5. Different normalization methods for a tensor with batch axis N, channel axis C, the height H and the width W. The blue pixels were normalized by calculating the needing values from these pixels. [8]

Pooling Layer

Both NAs use a MaxPool2D-Layer [6] twice. But instead of a kernel size of 2 we reduced it to size 1. This means the amount of pixels is still shrunk to the half but without max pooling.

Fully Connected Layer

After the second max pooling the output is flattened to size 1x35344 and passed to a fully connected layer. This might now maintain this dimension or already shrink the output. Shrinking the output also significantly decrease the size of the model. After cross-validating different output sizes we fixed it to 512. The original NA uses an output size of 84. It seems that the fast reduction of size before passing the tensor to the last ReLu leads to too much information loss.

IV. EXPERIMENTS AND RESULTS

We take the whole dataset (See II-A) once for humans and once for monsters and use a k-fold cross-validation ($k = 5$) for each separately. We randomly select images from the dataset and train our network for 20 epochs with a mini-batch size of 100 and a learning rate of $\eta = 0.01$. **MSE** between the actual and predicted HBDs is minimized during training. Furthermore we use stochastic-gradient-descent with momentum term ($\alpha = 0.9$) as optimization method. Those hyperparameters are identical to those provided in the paper of the original NA [7].

The experiment is performed on a Computer with an AMD Ryzen 9 5950X CPU and a NVIDIA RTX 3090 with 24 GB of GPU Memory.

A. Quantitative Evaluation

We evaluate our model with a k-fold cross validation ($k = 5$). Each fold j consists of $a = 2400$ images for the humans dataset and $a = 400$ for the monsters dataset. We estimate the eight HBDs using our altered model and use the same metrics as in the original paper [7]:

Statistics (here for the humans dataset) are based on a results tensor of shape:

$$k \times a \times |\hat{D}_i, D_i| \times 8 = 5 \times 2400 \times 2 \times 8 \tag{1}$$

The estimation error e^j_{MAD} for each HBD i is the Mean Absolute Difference over the j folds between the predicted and actual HBDs $\hat{D}_i, D_i$:

$$e^j_{MAD} = \frac{1}{a}\sum_{l=1}^{a}|\hat{D}_l - D_l|, \qquad e^j_{MAD} = \frac{1}{k}\sum_{j=1}^{k} e^j_{MAD} \tag{2}$$

Furthermore, we consider the Relative Percentage Error (RPE) e^j_{RPE} for each HBD i and its average (ARPE).

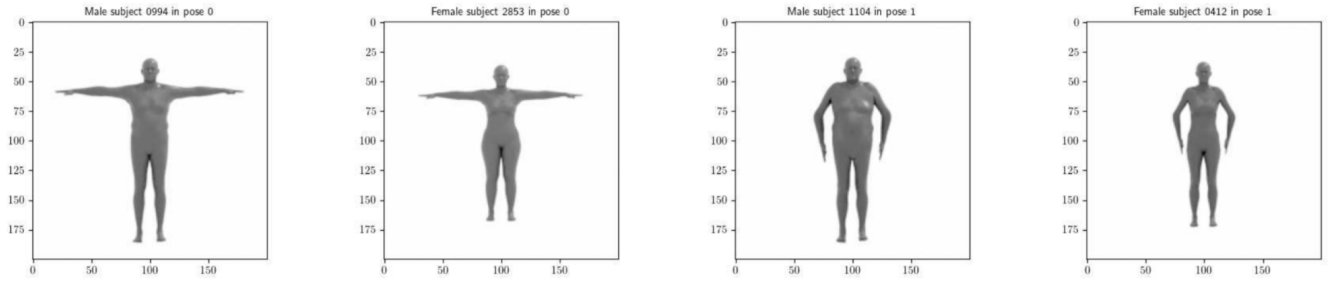
$$e^j_{RPE} = \frac{1}{a}\sum_{l=1}^{a}|\frac{\hat{D}_l - D_l}{D_l}|, \qquad e^j_{RPE} = \frac{1}{k}\sum_{j=1}^{k} e^j_{RPE} \tag{3}$$

TABLE I
COMPARISON ORIGINAL NA AND OUR NETWORK

HBD	MAD			RPE (%)		
	Org.	Ours	Diff. (%)	Org.	Ours	Diff. (%)
Shoulder width	12.54	7.33	- 0.42	4.93	3.41	- 0.31
Right arm length	12.98	7.33	- 0.44	2.22	1.26	- 0.43
Left arm length	13.48	7.64	- 0.43	2.34	1.33	- 0.43
Inseam/crotch h.	22.17	12.20	- 0.45	3.12	1.78	- 0.43
Chest circumf.	25.22	10.95	- 0.57	2.51	1.09	- 0.57
Waist circumf.	27.53	10.58	- 0.62	3.67	1.26	- 0.66
Pelvis circumf.	25.85	8.88	- 0.66	2.40	0.84	- 0.65
Height	27.34	6.70	- 0.75	1.58	0.39	- 0.75
AMAD	20.89	8.95	- 0.57			
ARPE				2.84	1.42	- 0.50

TABLE II
RESULTS HUMANS

HBD	Plain		RGB		Texture	
	MAD	RPE (%)	MAD	RPE (%)	MAD	RPE (%)
Shoulder width	7.33	3.41	9.59	3.89	17.02	5.85
Right arm length	7.33	1.26	9.37	1.62	17.90	3.10
Left arm length	7.64	1.33	9.77	1.71	17.89	3.14
Inseam/crotch height	12.20	1.78	15.27	2.22	29.99	4.21
Chest circumference	10.95	1.09	15.25	1.51	36.30	3.61
Waist circumference	10.58	1.26	15.31	1.80	38.46	4.65
Pelvis circumference	8.88	0.84	11.96	1.14	31.01	3.01
Height	6.70	0.39	10.28	0.60	31.78	1.86
AMAD	8.95		12.10		27.54	
ARPE (%)		1.42		1.81		3.68



HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	122,25	124,1	**1,51**	1,52
H	193,99	193,84	**-0,08**	-0,35
I	90,27	89,26	**-1,12**	-1,46
LAL	65,79	63,68	-3,21	**-1,76**
PC	114,6	115,03	0,38	**0,25**
RAL	65,95	65,98	**0,05**	-0,19
SW	48,33	47,41	**-1,90**	-2,64
WC	112,82	110,99	**-1,62**	1,52

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	99,61	99,61	**0,00**	1,16
H	164,59	164,7	**0,07**	-0,17
I	67,86	67,58	**-0,41**	1,53
LAL	58,94	58,61	**-0,56**	-1,50
PC	114,25	113,04	**-1,06**	-1,72
RAL	59,16	59,89	1,23	-1,23
SW	36,25	37,57	3,64	**2,41**
WC	86,91	86,07	**-0,97**	1,16

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	134,4	135,59	**0,89**	-1,37
H	195,12	194,29	-0,43	**0,12**
I	88,61	85,81	-3,16	**-0,28**
LAL	65,17	65,35	**0,28**	2,29
PC	119,96	118,93	**-0,86**	-1,32
RAL	67,3	68,38	1,60	**-0,88**
SW	46,49	48,49	4,30	**0,44**
WC	120,46	121,68	**1,01**	-1,37

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	93,08	92,76	**-0,34**	1,10
H	174,23	174,04	-0,11	**0,10**
I	75,02	74,77	**-0,33**	2,79
LAL	59,62	57,24	-3,99	**-3,18**
PC	98,04	97,65	-0,40	**-0,31**
RAL	58,31	57,33	**-1,68**	-1,93
SW	36,17	36,13	**-0,11**	-0,18
WC	79,17	78,37	**-1,01**	1,10

Fig. 6. Estimation results of the same four subjects which are also selected in the original paper. The error of our net and the original net is also given. Bold numbers indicate a better value compared to the Two females and two males. Abbreviations: CC: chest circumference, H: height, I: inseam, LAL: left arm length, PC: pelvis circumference, RAL: right arm length, SW: shoulder width, WC: waist circumference.

TABLE III
RESULTS MONSTERS

HBD	Plain		RGB		Texture	
	MAD	RPE (%)	MAD	RPE (%)	MAD	RPE (%)
Shoulder width	22.50	7.10	26.10	8.23	32.19	9.93
Right arm length	32.87	12.92	37.15	13.43	42.21	13.01
Left arm length	37.75	16.73	42.59	13.95	48.27	21.21
Inseam/crotch height	30.99	inf	37.32	inf	61.72	inf
Chest circumference	36.96	3.72	46.60	4.59	68.96	6.75
Waist circumference	31.67	4.43	41.76	5.94	72.68	10.74
Pelvis circumference	25.80	2.59	35.42	3.59	60.92	6.38
Height	21.31	1.26	29.21	1.74	64.75	3.87
AMAD	29.98		37.02		56.46	
ARPE (%)		inf		inf		inf

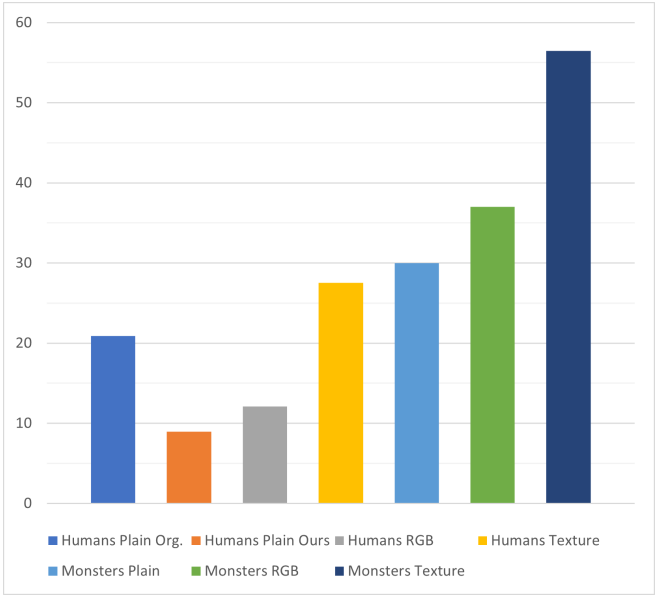


Fig. 7. AMAD on the test set after 20 epochs of training

We report the Average Mean Absolute Difference (AMAD) over the HBDs:

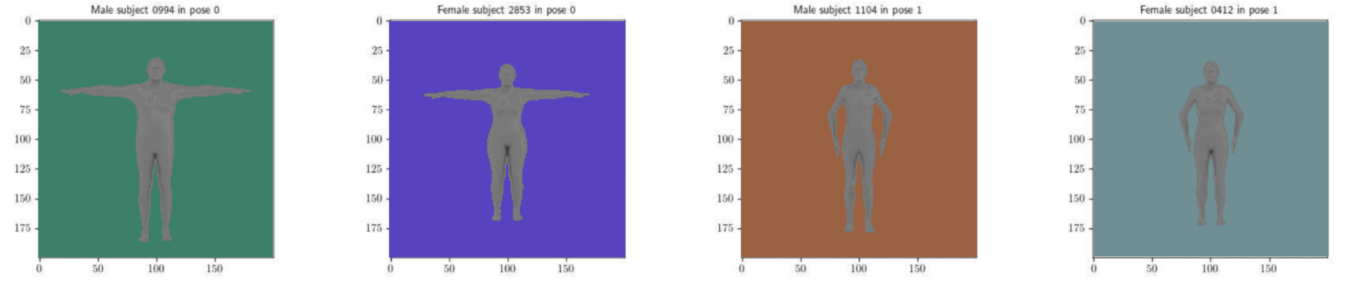
$$e_{AMAD} = \frac{1}{8}\sum_{j=1}^{8} e^{j}_{AMAD} \tag{4}$$

For having a baseline and compare our modified NA to the original NA we first evaluated the human dataset with plain background. The results are shown in Table I. It shows that the AMAD declines from 20.89 to 8.95 millimeters using our NA, an improvement of 57%. Furthermore, ARPE improves by 50%, showing a fall from 2.82 to 1.42 mm.

The tables II and III show the MAD and RPE for each HBD and in average for the humans dataset and the monsters dataset respectively. They also contain all three different setting: plain, RGB or texture background.

Also for the humans dataset the AMAD for the images with RGB background is about 25% higher than that of the ones with plain background it is still better than the results of the original NA. The same applies to the ARPE. The ARPE for the texture background is twice at high than the one achieved with RGB background.Shoulder width, left and right arm length can be estimated always quite well while the pelvis circumference and height also was estimated considerable well at the images with plain and RGB background.

The monsters dataset instead was indeed difficult for the



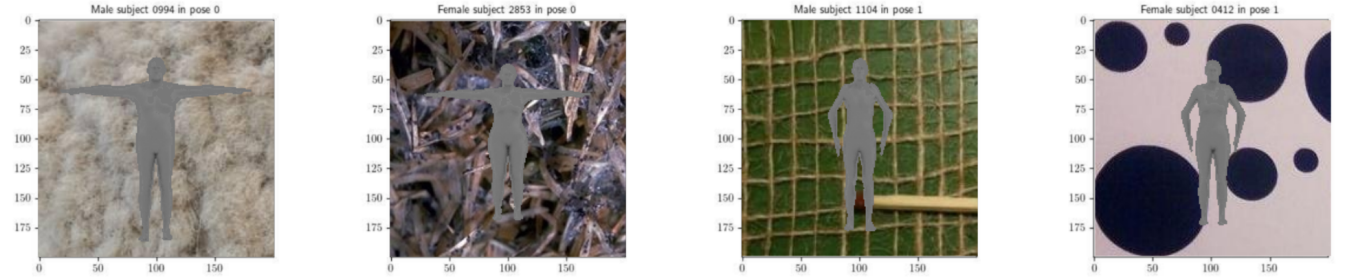
HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	122,25	119,45	-2,29	**1,52**
H	193,99	191,65	-1,21	**-0,35**
I	90,27	87,3	-3,29	**-1,46**
LAL	65,79	63,86	-2,93	**-1,76**
PC	114,6	112,17	-2,12	**0,25**
RAL	65,95	65,25	-1,06	**-0,19**
SW	48,33	47,84	**-1,01**	-2,64
WC	112,82	112,16	**-0,59**	1,52

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	99,61	100,38	**0,77**	1,16
H	164,59	165,73	0,69	**-0,17**
I	67,86	68,84	**1,44**	1,53
LAL	58,94	58,73	**-0,36**	-1,50
PC	114,25	113,38	**-0,76**	-1,72
RAL	59,16	59,11	**-0,08**	-1,23
SW	36,25	36,53	**0,77**	2,41
WC	86,91	36,53	-57,97	**1,16**

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	134,4	132,97	**-1,06**	-1,37
H	195,12	193,92	-0,62	**0,12**
I	88,61	85,66	-3,33	**-0,28**
LAL	65,17	64,64	**-0,81**	2,29
PC	119,96	119,11	**-0,71**	-1,32
RAL	67,3	67,45	**0,22**	-0,88
SW	46,49	48,11	3,48	**0,44**
WC	120,46	121,62	**0,96**	-1,37

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	93,08	92,54	**-0,58**	1,10
H	174,23	174,15	**-0,05**	0,10
I	75,02	75,75	**0,97**	2,79
LAL	59,62	58,35	**-2,13**	-3,18
PC	98,04	97,59	-0,46	**-0,31**
RAL	58,31	57,6	**-1,22**	-1,93
SW	36,17	35,53	-1,77	**-0,18**
WC	79,17	79,24	**0,09**	1,10

Fig. 8. Our input mini batch now with an added RGB background. Via preprocessing the image is converted from RGB to grey, which results in images with different grey values.



HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	122,25	120,44	**-1,48**	1,52
H	193,99	195,51	0,78	**-0,35**
I	90,27	87,48	-3,09	-1,46
LAL	65,79	66,06	0,41	**-1,76**
PC	114,6	114,05	-0,48	**0,25**
RAL	65,95	66,02	**0,11**	-0,19
SW	48,33	47,86	**-0,97**	-2,64
WC	112,82	111,1	**-1,52**	**1,52**

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	99,61	98,99	**-0,62**	1,16
H	164,59	165,21	0,38	**-0,17**
I	67,86	65,92	-2,86	**1,53**
LAL	58,94	58,72	**-0,37**	-1,50
PC	114,25	113,08	**-1,02**	-1,72
RAL	59,16	58,99	**-0,29**	-1,23
SW	36,25	35,98	**-0,74**	2,41
WC	86,91	86,28	**-0,72**	1,16

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	134,4	131,98	-1,80	**-1,37**
H	195,12	195,47	0,18	**0,12**
I	88,61	86,5	-2,38	**-0,28**
LAL	65,17	65,08	**-0,14**	2,29
PC	119,96	120,61	**0,54**	-1,32
RAL	67,3	67,01	**-0,43**	-0,88
SW	46,49	47,68	2,56	**0,44**
WC	120,46	118,62	-1,53	**-1,37**

HBD	Actual	Predicted	Error Ours (%)	Error Original Paper (%)
CC	93,08	97,15	4,37	**1,10**
H	174,23	173,82	-0,24	**0,10**
I	75,02	76	**1,31**	2,79
LAL	59,62	58,74	**-1,48**	-3,18
PC	98,04	98,03	**-0,01**	-0,31
RAL	58,31	57,66	**-1,11**	-1,93
SW	36,17	36,11	**-0,17**	-0,18
WC	79,17	80,84	2,11	**1,10**

Fig. 9. The mini batch shown with the added DTD backgrounds.

modified NA. Also the RPE for the inseam could not be calculated at all. That might be because of a hugh difference between the real and the estimated value in some cases. This difference grows to infinity while passed on during training.

B. Visualisation and Discussion

Figure 7 shows the AMAD of all six different datasets and also the one achieved by the original NA for the dataset with humans without background. It shows the modified NA can take on the results of the original one even with images of humanoids with textured background.

As already stated out, the monster HBDs are difficult to estimate. But we do recommend to keep a hugh varity of different body shapes due to diversity. Nevertheless one should only include images showing survivable humans.

Fig 6 shows $k = 4$ instances from a mini batch of the humans dataset with plain background. The models match the very same images discussed in the original paper [7] to give a comparison between the two implementations. The tables beneath the images contain the real and predicted value and the resulting relative percentage error. In the last column there is the achieved relative percentage error reported in the original paper. The shoulder with was also hard to estimate for the modified network but not only in the case of an human in pose 0. The height also achieved one of the lowest errors across all subjects and HBDs

The same humans with an RGB background achieve only slightly worse results as with the plain background dataset

and even the images with added textures show very good results. When looking at the height, the plain background dataset showed an error range from -0.43 % to 0.07 %, the RGB background dataset -1,21 % to 0.69 % and the texture background dataset showed -0,24 to 0,78.

Taking the shoulder width, we achieve with the plain background dataset -1,90 % to 4,30 %, with the RGB background dataset -1,77 % to 3,48 % and with the texture background dataset we reach an error of -0,97 % to 2,56 %.

Also some errors might seem not irrelevant at first glance a look at the actual and predicted values relativizes it. Almost no errors of the four presented examples across all different background types exceeds 3 cm.

V. CONCLUSION

Our goal was to create datasets with artificial backgrounds and evaluate the performance of the Neural Anthropometer [7] when using those augmented datasets. We changed the number of connections in the last linear layer, the normalization method and the pooling strategy. The rest of the CNN Architecture remained the same. However, we achieved far better results as in the original paper, even with the texture and background dataset. AMAD was improved by 57% and ARPE by 50%.

REFERENCES

[1] A. F. Agarap, "Deep learning using rectified linear units (relu)," 2018.

[2] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[6] Paszke, et. al., "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'Alché-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-learning-library.pdf

[7] Y. G. Tejeda and H. A. Mayer, "A neural anthropometer learning from body dimensions computed on human 3d meshes," 2021.

[8] Y. Wu and K. He, "Group normalization," 2018.

[9] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," 2019.