BCNet: Learning Body and Cloth Shape from A Single Image(Supplementary Material)

Boyi Jiang[1], Juyong Zhang[1], Yang Hong[1], Jinhao Luo[1], Ligang Liu[1], and Hujun Bao[2]

[1] University of Science and Technology of China
[2] State Key Lab of CAD&CG, Zhejiang University

In the main paper, we propose the BCNet, a deep neural network model that takes a near front view color image of the clothed human body as input, and outputs reliable garments and body 3D geometries separately. In this supplemental material, we first describe some implementation details, then present more qualitative results and discuss the limitations of our approach.

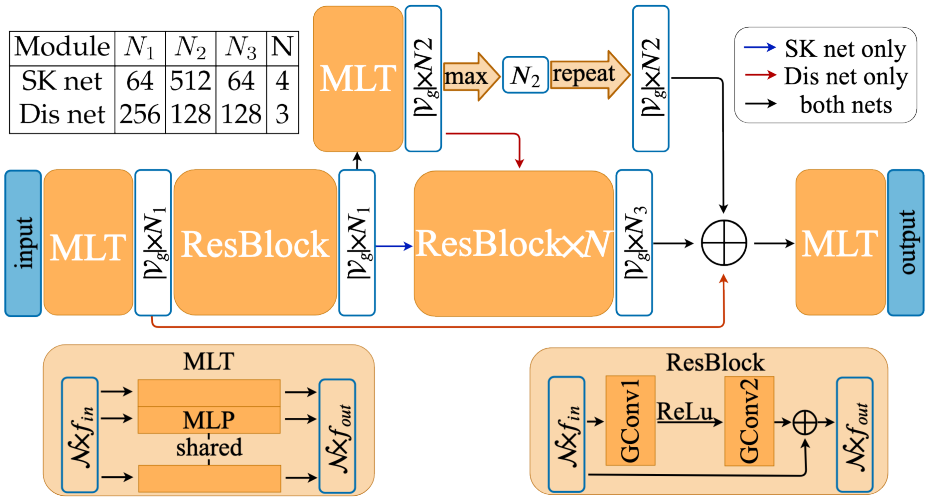

Fig. 1. The architecture of the skinning weight(SK net) and displacement network(Dis net). Two networks follow similar architecture with slightly different configuration, as presented in the figure. The Multi-Layer transform(MLT) is a module that utilizes a shared Multi-Layer perceptron(MLP) to change the feature dimension for each vertex. The ResBlock follows the design of the standard residual network.

1 Implementation Details

Detailed Architecture. In this section, we describe the detailed architecture design of the skinning weight network and displacement network. Both networks follow similar architecture presented in Fig. 1 while with slightly different

module configuration and computing flow. The architecture is a point-to-point feature calculation process that does not require upsampling or downsampling. We utilize Multi-Layer transform(MLT) and ResBlock to construct the network. The MLT uses a shared Multi-Layer perceptron(MLP) to change the feature dimension for each vertex. And the ResBlock is used to extract deeper features following the classical design. To obtain global information, we do max-pooling on the middle layer feature and concatenate it to the input of the final MLT. Optionally, we can also concatenate the shallow feature to enhance the inference ability of high-frequency details. The specific graph convolution, input features, and outputs for the two networks are different, and they have been described in the main paper.

Details of Rigged and Posed Registration. Before computation, we need to segment the rigged avatars. Some of the purchased rigged avatars provide accurate garment and skin segmentation. For segmentation that is inaccurate or unavailable, we manually segment the models with Blender utilizing the texture or color information. For BUFF[56] data, we segment similarly. The garment and skin segmentation of the Digital Wardrobe [7] is available. For the optimization, we implement the whole objective energy with Pytorch framework, and minimize the loss with the Adam iteration method.

2 More Results

In this section, we show more evaluations and results of our approach and comparisons with the state-of-the-art methods. First, we show samples of our constructed dataset in Fig. 2. Then, we show more evaluations of our skinning weight network. Next, we show a qualitative comparison with MGN [7] in Fig. 5 and more reconstruction results in Fig. 6, Fig. 7, and Fig. 8. Finally, more garment transfer and switching results are presented in Fig. 9, Fig. 10 and Fig. 11.

Samples of Constructed Dataset. In Fig. 2, we present more samples of our constructed Synthetic Dataset and HD Texture Dataset. Our constructed datasets include various kinds of postures and garments, and the garment geometries match quite well with the corresponding images.

Table 1. The skinning weight ℓ_1 errors($\times 10^{-3}$) between the predicted weight by our skinning weight network and ground truth, and the Euclidean distance(ED) in mm between deformed meshes with predicted and ground truth weights.

type	ℓ_1 mean	ℓ_1 std	ED mean	ED std
l-shirt	0.81	2.83	0.40	0.57
s-shirt	0.84	2.82	0.43	0.54
l-pant	0.45	1.92	0.30	0.36
s-pant	0.51	2.23	0.34	0.61
l-skirt	0.71	2.88	0.64	2.00
s-skirt	0.66	2.55	0.42	0.90

Fig. 2. Some examples of Synthetic Dataset and HD Texture Dataset. We show color image and geometry in each group. The first two rows are samples from Synthetic Dataset, and the last two rows are samples from HD Texture Dataset. Our constructed dataset includes various kinds of postures and garments.

Evaluation of Skinning Weight Network. We test the reconstruction ability of our skinning weight network. For each neutral garment of the whole test set, we reconstruct the skinning weights with our network and compute the ℓ_1 error relative to the ground truth weights. Then, to evaluate skinning deformation, we compute the Euclidean distances between deformed garments with predicted and ground truth weights, respectively. Twenty random postures from the Mocap dataset are used to deform these neutral garments. From second to fifth columns of Table 1, the average and standard deviation of ℓ_1 error and Euclidean distance for each garment type are given in turn. As can be observed from the results, our model can generate highly accurate results. In Fig. 3, we visualize the error maps of several deformed garments of different types.

For some specific garments, close vertices in some parts may have very different skinning weights. For example, the vertices on either side of the crotch have

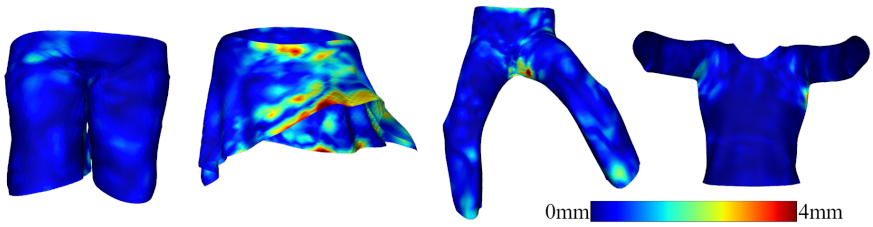

Fig. 3. Error maps between deformed clothes with predicted skinning weights and ground truth weights.

close vertex positions, while the skinning weights are very different because they belong to different legs. In this case, the vertex normals of input could supply useful information to distinguish these ambiguous vertices. As shown in Fig. 4, the network without normals as input can not predict accurate weights of some vertices, leading to significant errors and artifacts.

Qualitative Comparison with MGN. We show two comparison results on BUFF [56] dataset with MGN [7] in Fig. 5. The results by MGN are generated by 8 views input without postprocessing optimization. We can see that the predicted body shapes by our method match the input image more closely. Besides, our predicted garments can capture size variation, such as the length of trouser legs. For the MGN method, the predicted garments are bound with SMPL, which limits its expression ability of large displacement. Therefore, although the input images are different, the results of MGN tend to maintain the same style and size for the same garment type.

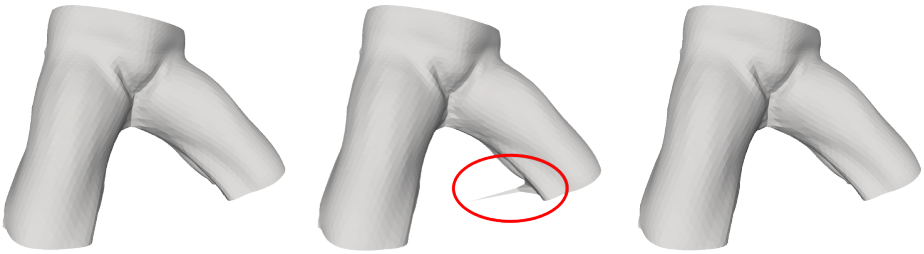

Fig. 4. Ablation study of the vertex normals as input for skinning weight network. On the left is the deformed pant with GT weights, the middle is the result of the network without normals as input, and the right is the result of the network with full input. Predictions without normals as input would produce artifacts in ambiguous vertices.

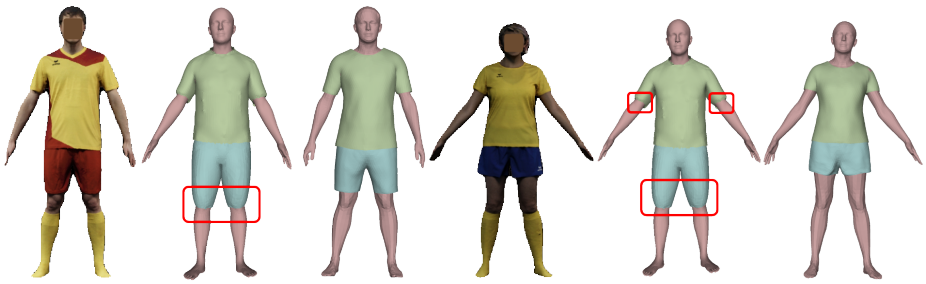

Fig. 5. Two comparison results with MGN on BUFF. From left to right of each group, the reference image, predicted shapes(without post-optimization) by MGN, and predicted shapes by our method are presented. We can observe that the predicted body shapes by our method match the input images better than the ones by MGN. Moreover, our garments can capture the size variation of the input image, while predicted garments of MGN tend to maintain similar size for the same garment type even with different input images. The mismatched parts between predicted garment size and the input image are marked with red boxes.

Qualitative Results. To demonstrate the reconstruction ability of our method, we show more results on different datasets. In Fig. 6, we show some reconstruction results from our test set. In Fig. 7, we show some reconstruction results from the Digital Wardrobe [7]. In Fig. 8, reconstruction results on real images are given. From these results, we can see that our method can capture the body and garment shapes from the input images quite well.

Garment Transfer. Fig. 9 shows a garment switching example and two garment transfer examples between images. All the reconstructed shapes are rendered with texture to improve the visual authenticity. In Fig. 10 and Fig. 11, we present more results of garment transfer and switching on HD Texture test dataset. Based on our method, we can easily transfer or switch garments geometry and texture between two images, even with different garment types. These applications further confirm the ability of our method to correctly predict the 3D shapes of garment and human body.

3 Limitations

Our method still has some limitations which deserve further study.

- Our method currently supports six garment types. Therefore, current trained model can not correctly predict the garment type which does not belong to the six garment types. One example is shown in Fig. 12 A. However, our method can be extended to support new garment type with the same strategy in the paper.
- In this work, our method can recover the body and garment shapes while we does not consider the hair, shoes, hats, and multi-layered clothing. We show an example of multi-layered clothing in Fig. 12 B.

Fig. 6. Reconstructed results by our method on test set. Each group includes the input image and our result.

– Clothed body images contain enormous diversities in the aspect of cloth types, textures, body shapes, lighting conditions, background, and camera angles. The trained model might produce over-smooth results for test images that have very different styles with the training dataset. Two examples are shown in Fig. 12 C. However, synthesizing more realistic data and utilizing more real data in our proposed framework can alleviate this problem.

Fig. 7. Reconstructed results by our method on DW Dataset [7]. Our method can recover the body shape, posture, and cloth type and shape quite well.

Fig. 8. Reconstructed results by our method on real images. Each group includes the input image and our result. First row is three images from the internet, and second row is two images from PeopleSnapshot [3] dataset.

Fig. 9. On the left, with two input images of the first row, we can predict their body shapes and exchange garments geometries and textures with BCNet. In the right, two garment transfer examples are given. In each row, we transfer the garments of the first image to the second image and show the reconstructed shapes with texture on the third column.

Fig. 10. More garment transfer results on test dataset. The target garment images are shown in the first column, and the garment in the target image is transfered to four different source clothed body images. For each result, the predicted body and transfered garments with texture are presented.

Fig. 11. Examples of garment switching. We supply pairs clothed body images in the first row and present predicted bodies and switched garments with texture in the second row.

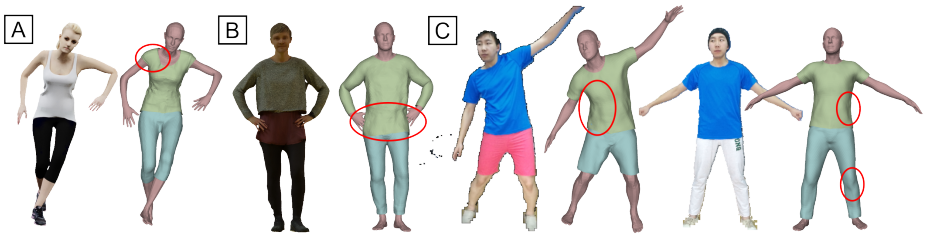

Fig. 12. The proposed method has shortcomings. We present three challenges for future work in the figure. A) Incorrect garment reconstruction due to unsupported garment type. B) Multi-layered garments are treated as a single layer garment. C) Over-smooth results for two images captured by Kinect v2 camera [55].