NormalGAN – Supplementary Material

Lizhen Wang[1], Xiaochen Zhao[1], Tao Yu[1], Songtao Wang[1], and Yebin Liu[1]

Tsinghua University, Beijing, China

1 Architecture Details

The detailed architecture of our generators are shown in Tab. 1, which is based on UNet. Some other details: 1) The output images of the front-view depth rectification network $\mathcal{G}_{df}$ actually have 2 channels. One channel is the rectified depth image and the other channel is the mask of the orthographic view. 2) The encoder for SH bases in $\mathcal{G}_{cf}$ is the same as the down-sampling layers of the generators. 3) To enhance the inference capability of the back-view inference networks G_{db} and G_{cb}, we remove the skipped connection of the "Int3" layers in Tab. 1.

Layer	**Layer type**	**Output shape**
Input	Input	424x424xC_{in}
Conv1	Conv 3x3 stride=1, ELU	424x424x32
	Conv 3x3 stride=2, ELU, InstanceNorm	212x212x64
Conv2	Conv 3x3 stride=1, ELU	212x212x64
	Conv 3x3 stride=1, ELU	212x212x64
	Conv 3x3 stride=2, ELU, InstanceNorm	106x106x128
Conv3	Conv 3x3 stride=1, ELU	106x106x128
	Conv 3x3 stride=1, ELU	106x106x128
	Conv 3x3 stride=2, ELU, InstanceNorm	53x53x256
Conv4	Conv 3x3 stride=1, ELU	53x53x256
	Conv 3x3 stride=1, ELU	53x53x256
	Conv 3x3 stride=2, ELU	27x27x512
	Conv 3x3 stride=1, InstanceNorm	27x27x512
Res1	Residual module block	27x27x512
Res2	Residual module block	27x27x512
Int1	Interpolation,skipped connection from Conv3	53x53x768
Conv5	Conv 1x1 stride=1, ELU	53x53x256
	Conv 3x3 stride=1, ELU	53x53x256
	Conv 3x3 stride=1, ELU, InstanceNorm	53x53x256
Int2	Interpolation,skipped connection from Conv2	106x106x384
Conv6	Conv 1x1 stride=1, ELU	106x106x128
	Conv 3x3 stride=1, ELU	106x106x128
	Conv 3x3 stride=1, ELU, InstanceNorm	106x106x128
Int3	Interpolation,skipped connection from Conv1	212x212x192
Conv7	Conv 1x1 stride=1, ELU	212x212x64
	Conv 3x3 stride=1, ELU	212x212x64
	Conv 3x3 stride=1, ELU, InstanceNorm	212x212x64
Int4	Interpolation	424x424x64
Conv8	Conv 1x1 stride=1, ELU	424x424x32
	Conv 3x3 stride=1, ELU	424x424x32
	Conv 3x3 stride=1, ELU, InstanceNorm	424x424xC_{out}

Table 1: The architecture of our generators.

The architecture of our discriminators are shown in Tab. 2, which is inspired by PatchGAN. The 4-channel input image consists of a 3-channel normal map or color image and a 1-channel mask, which is used to enhance the discernment of the discriminators.

Layer	Layer type	Output shape
Input	Input	424x424x4
Conv1	Conv 4x4 stride=2, LeakyRelu	212x212x64
Conv2	Conv 4x4 stride=2, LeakyRelu, InstanceNorm	106x106x128
Conv3	Conv 4x4 stride=2, LeakyRelu, InstanceNorm	53x53x256
Conv4	Conv 4x4 stride=2, LeakyRelu, InstanceNorm	27x27x512
Conv5	Conv 4x4 stride=1	27x27x1

Table 2: The architecture of our discriminators.

2 Additional Experiments

The Validity of SH Bases Module. As shown in our overview figure, to enhance the front-view albedo inference, we encode the spherical harmonics bases (SH bases) calculated from the refined normal maps. To demonstrate the validity of the SH bases module, we train $\mathcal{G}_{cf}$ without the SH bases for comparison. As shown in Fig.1, $\mathcal{G}_{cf}$ trained without the SH bases fails to remove shadows caused by clothes folds due to the lack of encoded SH information.

Fig. 1: The results of our method with/without the SH bases module.

Multi-frame RBG-D methods. We extensively evaluate our method with the volumetric fusion methods DoubleFusion [2] and Guo et al. [1] using RGB-D sequences captured by a Kienct v2 camera. As shown in Fig.2, most volumetric fusion methods

need to start with A pose or T pose and maintain a fixed geometric topology. Moreover, important dynamic geometric details like facial expressions and clothing wrinkles are lost due to the incremental fusion procedure. More detailed results are presented in the supplementary video.

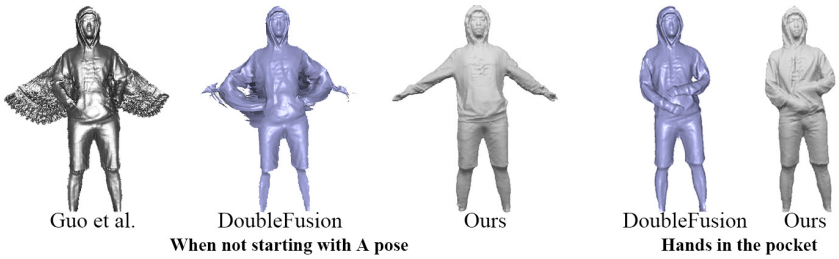

Fig. 2: Comparison with typical volumetric fusion methods on RGB-D sequences. None of the volumetric fusion methods can reconstruct accurate dynamic geometric details and handle topological changes.

Comparison with Single-image methods. As shown in the comparison section of our paper, we have presented the results of retained single-image methods retrained with RBG-D images of our dataset. As shown in Fig.3, we present the results of DeepHuman, Moulding Humans and PIFU retrained with only RGB images of our dataset.

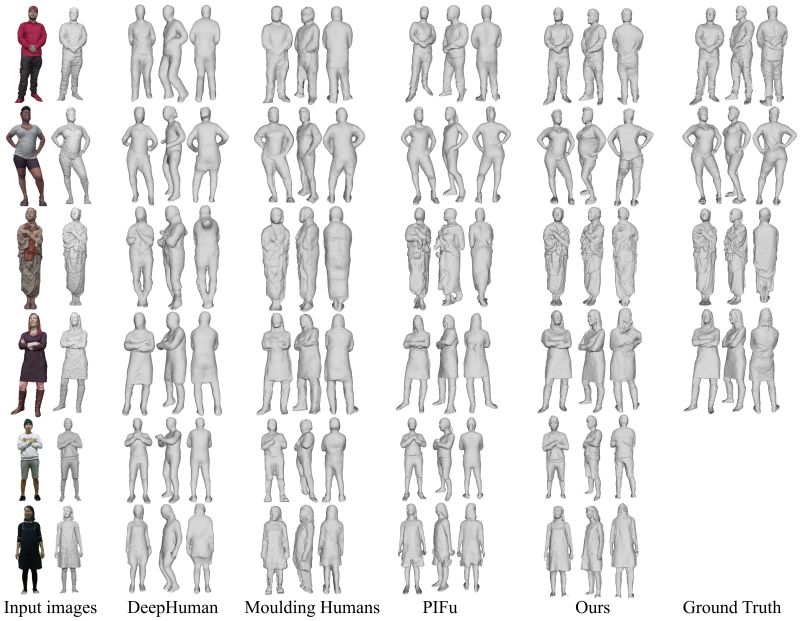

Fig. 3: Results of DeepHuman, Moulding Humans and PIFU retrained with only RGB images of our dataset and the corresponding NormalGAN results.

References

1. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. ACM Trans. Graph. **36**(3), 32:1–32:13 (Jun 2017). https://doi.org/10.1145/3083722, http://doi.acm.org/10.1145/3083722
2. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7287–7296 (2018)