ACCV2020
#85

Supplementary Material Multi-View Consistency Loss for Improved Single-Image 3D Reconstruction of Clothed People

Anonymous ACCV 2020 submission

Paper ID 85

1 Network Details

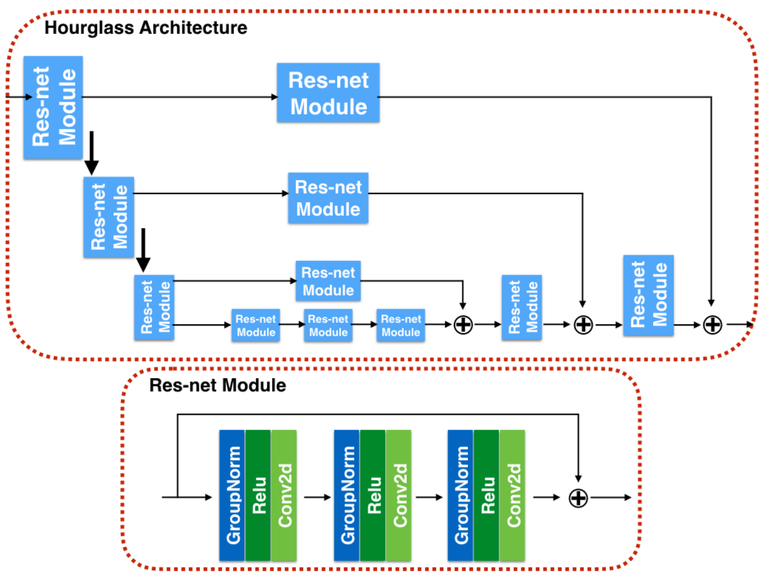

Fig. 1. This figure illustrates the hourglass architecture used in the proposed network together with associated res-net module.

In the proposed network (Main text Sec. 3.1), shared stacked hourglass architecture is used to predict single-image 3D reconstruction. In particular, the hourglass part of the network is inspired from a previous work proposed in [1]. However, in each hourglass network, we propose a different res-net module, which consists of group normalization [2], activation function and convolution layer, Fig. 1. With this res-net module, the proposed network is able to converge during training even if the batch size is small. In addition to overall architecture,

implementation detail of the proposed hourglass network is provided in Table 1. The input image size is $512 \times 512 \times 3$, and the size of output occupancy volume is $128 \times 128 \times 128$.

Table 1. Network Architecture Details : **conv** - 2D convolution, **GN** - Group Normalization, **tconv** - transpose convolution.

Layer	Kernel	Stride	Output
conv + GN + relu	3	1	[64, 256, 256]
conv + GN + relu	3	1	[256, 128, 128]
conv + GN + relu	3	1	[256, 64, 64]
conv + GN + relu	3	1	[256, 32, 32]
conv + GN + relu	3	1	[256, 16, 16]
conv + GN + relu	3	1	[256, 8, 8]
conv + GN + relu	1	1	[128, 8, 8]
tconv + GN + relu	1	1	[256, 8, 8]
tconv + GN + relu	3	1	[256, 16, 16]
tconv + GN + relu	3	1	[256, 32, 32]
tconv + GN + relu	3	1	[256, 64, 64]
tconv + GN + relu	3	1	[256, 128, 128]
conv + sigmoid	1	1	[128, 128, 128]

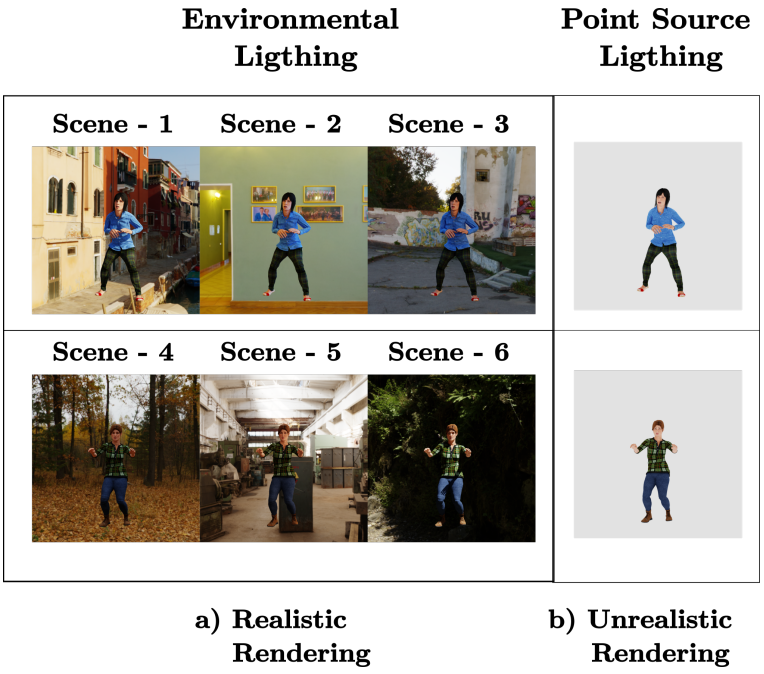

Fig. 2. This figure shows the different rendering of 3D human models with environmental and point source lightning, which results in realistic **(a)** and unrealistic **(b)** synthetic human images.

ACCV2020
#85

2 Improved Photo-Realism in 3DVH Dataset

The generation stages of 3DVH dataset is explained in the Main Text (Sec. 3.4). In 3DVH, the realistic images are rendered using environmental lighting and ray casting technique with surface properties of 3D models (gloss, diffuse, specular and normal maps). The proposed data generation method outperform the previous synthetic data generation approaches [3] with respect to photo-realism of the rendered images. Figure 2 shows the main difference between rendering with point source lighting and with the proposed environmental lighting. If a point source lighting is used (Fig. 2 - b), it is observed that the photo-realism of the rendered image is limited. However, the proposed data generation method of 3DVH dataset uses environmental illumination with the surface properties and apply ray casting technique to render realistic images. In Fig. 2 - (a), it is illustrated that different realistic images of the same 3D model can be generated using environmental illumination from different scenes. The improved photo realism in 3DVH dataset leads to learning a generalized method for single image 3D human reconstruction, which can be applied to real images.

References

1. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision, Springer (2016) 483–499
2. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 3–19
3. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: Modeling the Geometry of Dressed Humans. In: International Conference on Computer Vision (ICCV). (2019)