Supplemental Document: Neural Re-Rendering of Humans from a Single Image

Kripasindhu Sarkar[1] Dushyant Mehta[1] Weipeng Xu[2]
Vladislav Golyanik[1] Christian Theobalt[1]

[1]MPI for Informatics, SIC [2]Facebook Reality Labs

This document accompanies the main manuscript, and provides details of the network architectures employed, as well as snippets from the user study. Please refer to the main document, the accompanying video, and the project webpage[1] for further details.

1 Network Architecture

1.1 FeatureNet

FeatureNet is a U-Net based network that construct the full *UV Feature-map* from the partial RGB UV Texture-map. The network architecture is shown in Figure 1. FeatureNet comprises of four down-sampling blocks followed by four up-sampling blocks. We use the texture of resolution 256 × 256. At the middle most layer the intput is transformed to an activation volume of spatial dimension of 16 × 16. In all the figures, the 16 dimensional feature image is visualized by projecting it to 3 dimensions by a fixed random matrix.

1.2 RenderNet

Rendernet translates the rendered *Feature Image* (from source Feature Map and target pose) to a photorealistic image. The network architecture is shown in Figure 2. Both RendereNet and FeatureNet are trained *together* end-to-end following the full Pipelilne (Figure 2, Main paper).

2 Further results

In Figure 3 and 4 we show our results and its comparison to Coordinate Based Inpainting (CBI) [1] and DensePose Transfer (DPT) [2]. Here we present the list of the figures that was used in the user study.

[1] http://gvv.mpi-inf.mpg.de/projects/NHRR/

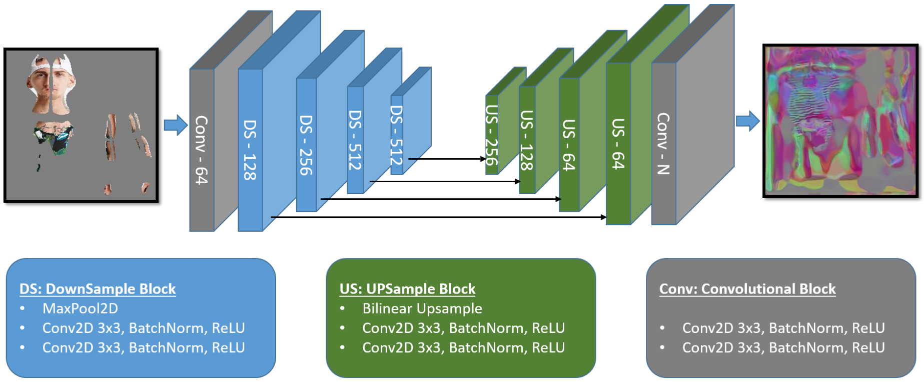
Conv - 64
DS - 128
DS - 256
DS - 512
DS - 512
US - 256
US - 128
US - 64
US - 64
Conv - N
DS: DownSample Block
• MaxPool2D
• Conv2D 3x3, BatchNorm, ReLU
• Conv2D 3x3, BatchNorm, ReLU
US: UPSample Block
• Bilinear Upsample
• Conv2D 3x3, BatchNorm, ReLU
• Conv2D 3x3, BatchNorm, ReLU
Conv: Convolutional Block
• Conv2D 3x3, BatchNorm, ReLU
• Conv2D 3x3, BatchNorm, ReLU

Fig. 1. FeatureNet (f): Network 1 of our full pipeline *(Figure 2 - Main Paper)*. FeatureNet converts the partial UV texture map to a full UV feature map, which encodes a richer N-dimensional representation at each texel. DS-<M>Denotes a DownSampling block containing MaxPool2D and double convolution with M ouput features. Similar is the case for *US* and *Conv* block.

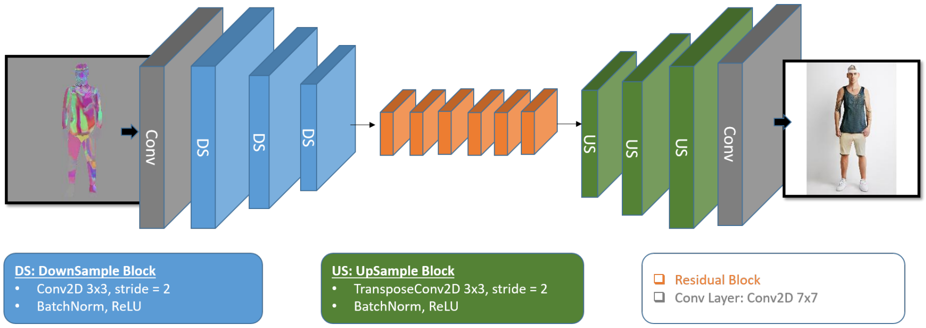

Fig. 2. RenderNet (g): Network 2 of our full pipeline *(Figure 2 - Main Paper)*. RenderNet is a translation network that translates the rendered d-dimensional *Feature Image* to a photorealistic image. The network comprises of (a) 3 down-sampling blocks, (b) 6 residual blocks, (c) 3 up-sampling blocks and finally (d) a convolution layer with *Tanh* activation that gives the final output.

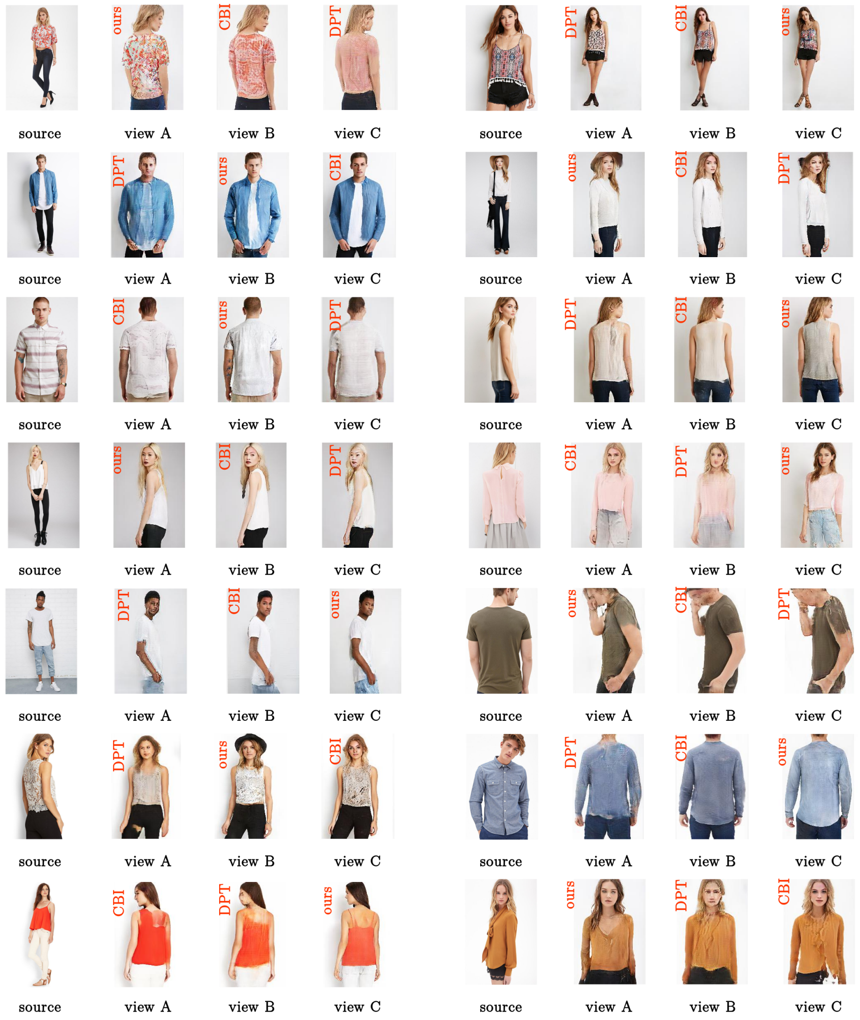

Fig. 3. The first 14 samples from the used study (out of 26). We show the source image and three views generated by CBI [1], DPT [2] and our method, in a randomised order. The keys – which were not exposed during the user study — are shown in orange. See Fig. 4 for the remaining samples.

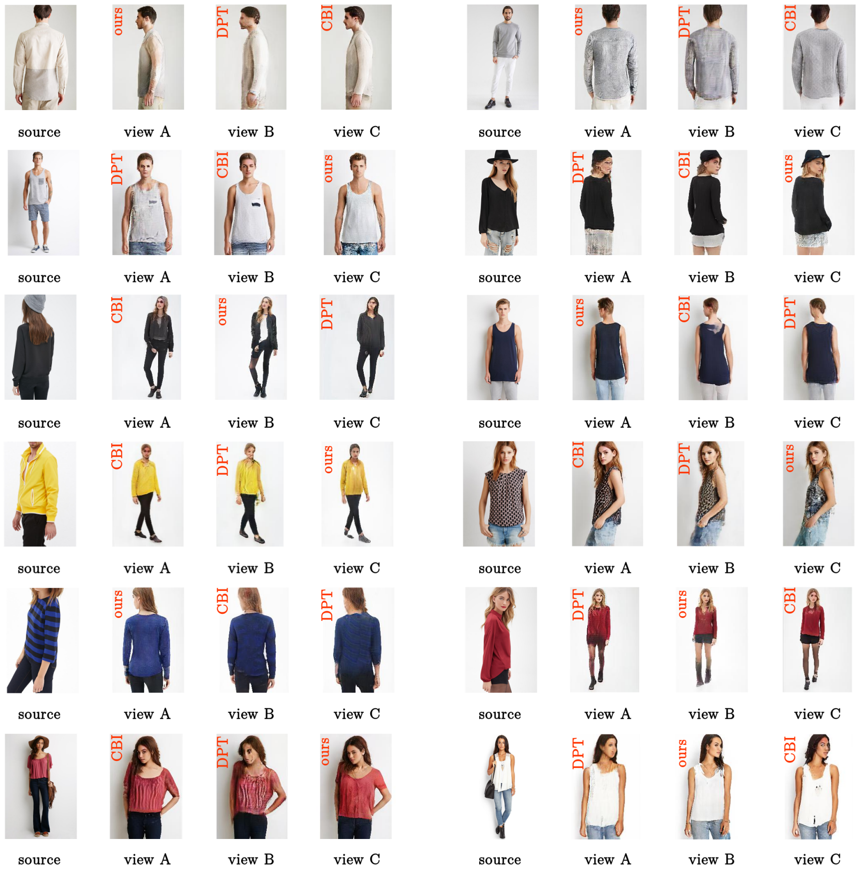

Fig. 4. The further 12 samples from the used study (out of 26).

References

1. Grigor'ev, A.K., Sevastopolsky, A., Vakhitov, A., Lempitsky, V.S.: Coordinate-based texture inpainting for pose-guided human image generation. Computer Vision and Pattern Recognition (CVPR) pp. 12127–12136 (2019)
2. Neverova, N., Güler, R.A., Kokkinos, I.: Dense pose transfer. European Conference on Computer Vision (ECCV) (2018)