Reposing Humans by Warping 3D Features (Supplementary Material)

Markus Knoche István Sárándi Bastian Leibe
RWTH Aachen University, Germany
{knoche,sarandi,leibe}@vision.rwth-aachen.de

A. Inception Score's Unsuitability in Reposing

Many related works use the Inception score (IS) [3], as a metric for person reposing. IS was proposed to evaluate unconditioned GANs, *i.e.*, GANs which are supposed to generate a diverse dataset like ImageNet [2] based on random inputs. Two aspects are combined: realism of single generated images and variability of a large set of generated images. Generated images are passed through the Inception network [5], a single realistic image x should be confidently assigned to a single class, so the assigned label distribution $p(y|x)$ has a single high activation. In contrast, multiple generated images should belong to different classes, thus $p(y)$ is rather uniform. IS compares these distributions using the Kullback-Leibler divergence, which means that the score is high, if the distributions are dissimilar. For human reposing only a single output class exists, such that for a perfect generator both $p(y)$ and $p(y|x)$ are the same, because the Inception network always assigns the label "human". This issue invalidates the Inception Score as a metric for person reposing.

B. Evaluation Protocol on iPER

Liu *et al*. [1], who published the iPER dataset, perform evaluation by selecting three frames per person and then generating the full video based on each of these frames. The results are then compared to the original videos using the quantitative metrics.

As the authors have not published their frame selection procedure, replicating their exact evaluation protocol is currently not possible. We therefore use the following selection procedure: we first uniformly sample a random clothing layout from the test set, then randomly select two frames from this person. The network then generates the second frame based from the first one. This process is repeated 10,000 times, the mean scores are reported.

C. Additional Qualitative Results

We compare our model to [4, supplementary Figure 9] in Table 1.

Table 2 shows generated images of our model compared to the ablation models and to the results of [1, Figure 7].

References

[1] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019.

[2] Olga Russakovsky et al. ImageNet large scale visual recognition challenge. *IJCV*, 2015.

[3] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NIPS*, 2016.

[4] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable GANs for pose-based human image generation. In *CVPR*, 2018.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Table 1. Qualitative comparison with a 2D feature warping method. The target image is not used as input, only its pose.

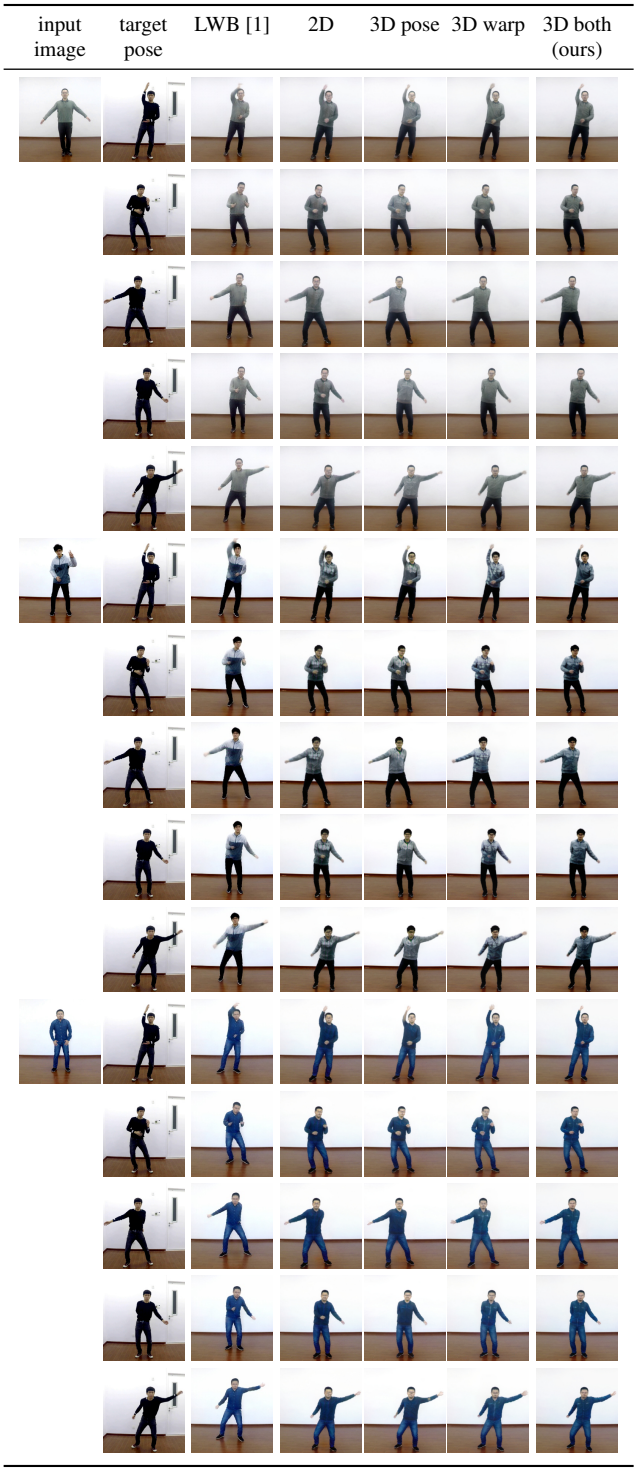

Table 2. Qualitative comparison to a mesh-based approach and to our ablation models.