Dense Intrinsic Appearance Flow for Human Pose Transfer Supplementary Material

Yining Li[1] Chen Huang[2] Chen Change Loy[3]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Robotics Institute, Carnegie Mellon University
[3]School of Computer Science and Engineering, Nanyang Technological University
ly015@ie.cuhk.edu.hk chenh2@andrew.cmu.edu ccloy@ntu.edu.sg

A. Network Architecture

Fig. S1 illustrates the detailed network architecture of our image generator described in Sec.4.1 in the main paper. There are $N = 7$ feature levels in the network, where the number of channels at each level increases linearly from 32 to 128. We apply flow guided feature warping at the first 5 feature levels because higher level features with small spatial resolution are not location-sensitive. At the bottleneck, the encoded pose features and warped appearance features are directly concatenated.

B. More Qualitative Results on DeepFashion Dataset

B.1. 3D Appearance Flow Regression

Fig. S2 shows more examples of the 3D appearance flow map and visibility map. The predicted appearance flow maps are accurate (close to the ground-truth) regardless of diverse pose and viewpoint changes. While the predicted visibility map can accurately identify invisible regions caused by self-occlusion (*e.g.*, the 1st and 3rd rows) or out-of-field of view (*e.g.*, the 2nd row).

B.2. Comparison with Previous Works

Fig. S3 shows more qualitative comparisons between our method and previous works [2, 3]. Results show that our method is able to generate more realistic images and better preserve the key appearance attributes.

B.3. Arbitrary Pose Transfer

We further test our method on transferring a reference image to arbitrary poses, and show the results in Fig. S4. In each row, the leftmost image is the reference image, which is used to synthesize new images in different target poses. It is good to see that our method can effectively generalize to diverse and difficult human poses.

B.4. Failure Case Analysis

Fig. S5 illustrates some failure cases of our method. The 1st and 4th rows show that our method has difficulty synthesizing some complex textures or special clothing layout, *e.g.*, coat wrapped around the person's waist. Larger training data are expected to improve out model's ability to hallucinate rare textures. In the second row, our method fails to correctly infer the backside of a person from her frontal appearance. Although our generated result is also plausible, the shoulder part seems not so compatible with the frontal image. We think more training data can help enrich the expressiveness of our image generator. The third row shows one failure case when our pixel warping module tries to blend an inconsistent reference image region into the generated image, which is again due to the large front-back pose discrepancy.

C. Experiments on Market-1501 dataset

Table S1: Quantitative results on Market-1501 dataset.

Model	SSIM	Masked SSIM	IS	Masked IS
PG2 [2]	0.253	0.792	**3.460**	3.435
DSC [3]	0.290	0.805	3.185	3.502
w/o. dual encoder	0.290	0.868	2.918	3.568
w/o. flow	0.292	0.869	2.905	3.664
w/o. visibility	0.296	0.872	3.193	**3.730**
w/o. pixel warping	0.303	0.873	2.986	3.699
Ours full	**0.308**	**0.874**	3.010	3.700

We further evaluate our model on the Market-1501 dataset [4], which consists of 32,668 surveillance images of 1,501 persons. Images in this dataset have a lower resolution of 128×64 pixels, but contain more diverse poses and complex backgrounds in comparison to images in DeepFashion dataset [1]. We follow the data splits in [3] and select 263,631 pairs for training and 12,800 pairs for testing. We modify the U-Net architecture of our image gen-

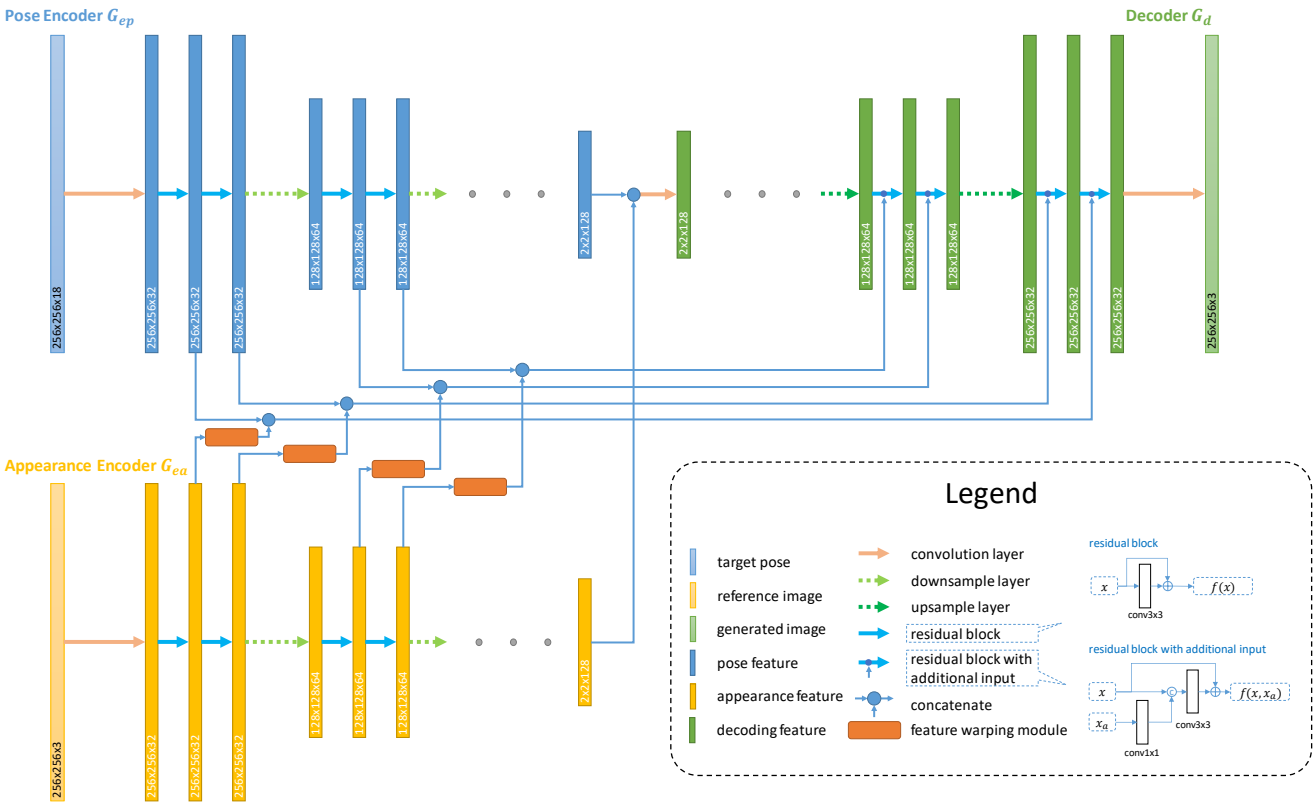

Figure S1: Detailed image generator network architecture.

erator to reduce to $N = 5$ levels due to the lower image resolution.

We show the quantitative results on Market-1501 in Table S1, and visualize some generated results in Fig. S6. Our method achieves pretty strong results when compared to state-of-the-art baselines [2, 3], and is able to generate higher quality details such as the backpack and clothing pattern.

References

[1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1

[2] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 1, 2, 4

[3] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 1, 2, 4

[4] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1

Figure S2: Examples of 3D appearance flow map and visibility map: generated ground-truth (middle) and prediction from our flow regression module (right).

Figure S3: Qualitative comparison between our method and previous works (PG2 [2] and DSC [3]).

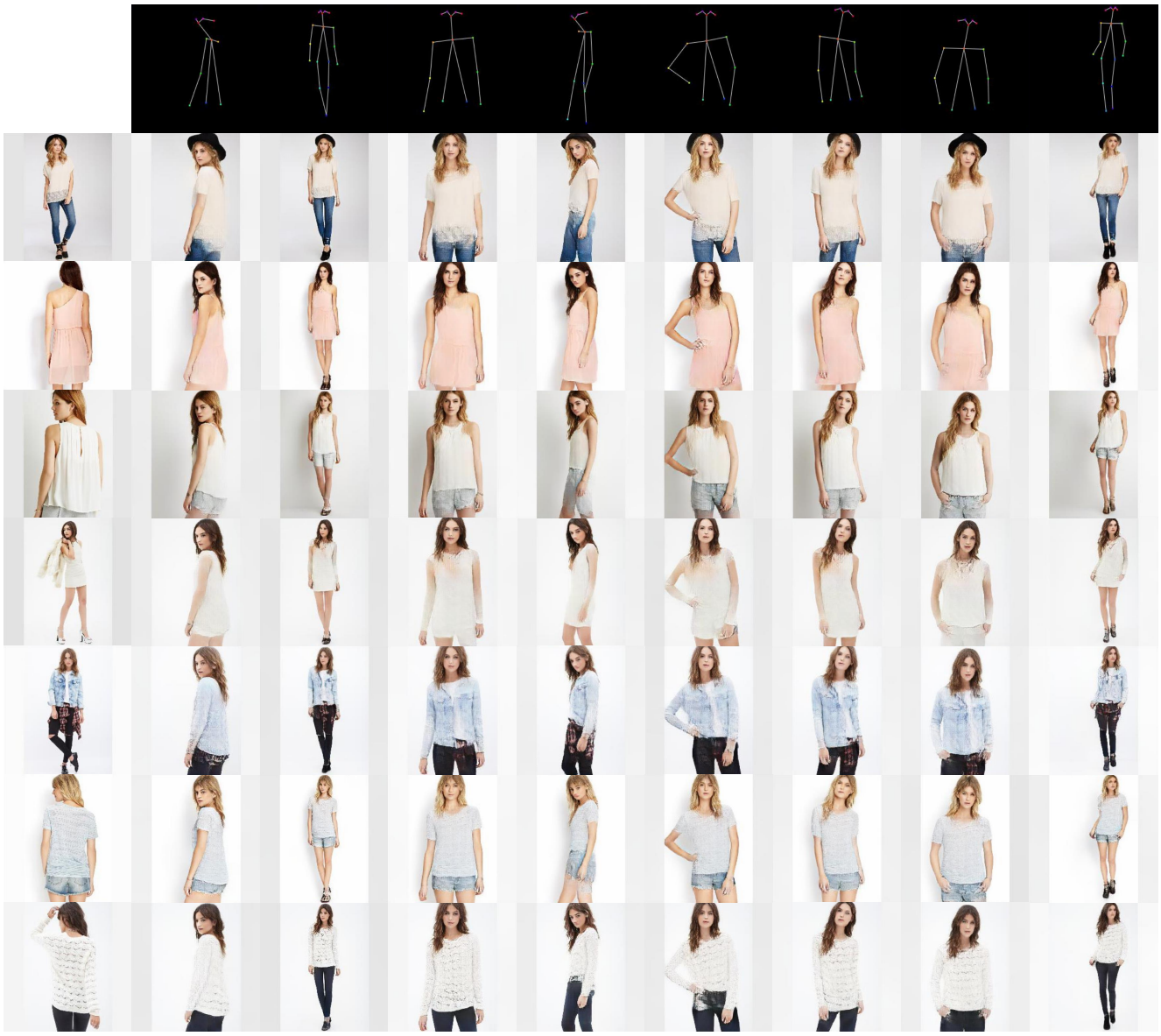

Figure S4: Arbitrary pose transfer results. Each image is synthesized using the leftmost reference image and the corresponding target pose.

Figure S5: Example failure cases

Figure S6: Qualitative results on Market-1501 dataset.