Disentangled Person Image Generation Supplementary Material

Liqian Ma[1] Qianru Sun[2*] Stamatios Georgoulis[1]
Luc Van Gool[1,3] Bernt Schiele[2] Mario Fritz[2]

[1]KU-Leuven/PSI, Toyota Motor Europe (TRACE) [3]ETH Zurich
[2]Max Planck Institute for Informatics, Saarland Informatics Campus
{liqian.ma, sgeorgou, luc.vangool}@esat.kuleuven.be
{qsun, schiele, mfritz}@mpi-inf.mpg.de

This supplementary material includes additional details regarding the network architecture (§A) and training (§B), as well as extended results for image manipulation (§C), pose-guided person image generation (§D), inverse interpolation (§E) and image sampling (§F), respectively.

A. Network architecture

In this section, we provide details regarding the network architectures in our two-stage framework used on the Market-1501 dataset. Fig. 2 shows 4 network architectures used at stage-I: 1) FG encoder consists of 5 convolutional residual blocks; 2) BG encoder consists of 5 convolutional residual blocks; 3) FG & BG decoder follows a "U-net"-based architecture; 4) Pose auto-encoder follows a fully-connected auto-encoder architecture. Fig. 1 shows the network architecture of the mapping functions Φ used at stage-II. It contains 4 fully-connected residual modules.

B. Training details

On Market-1501, our method is applied to disentangle the image into three factors: foreground, background and pose. We train the foreground and background models with a mini-batch of size 16 for $\sim 70k$ iterations at stage-I and with a mini-batch of size 32 for $\sim 30k$ iterations at stage-II. The pose models are trained with a mini-batch of size 64 for $\sim 30k$ iterations at stage-I and with a mini-batch of size 32 for $\sim 60k$ iterations at stage-II.

DeepFashion data contain clean background, therefore, our method is applied to disentangle the image into only two factors: appearance (*i.e.* foreground) and pose. We train the appearance model with a minibatch of size 6 for $\sim 100k$ iterations at stage-I and with a minibatch of size 16 for $\sim 60k$ iterations at stage-II. The pose models are trained with a minibatch of size 32 for $\sim 30k$ iterations at stage-I and with a minibatch of size 32 for $\sim 60k$ iterations at stage-II.

On both datasets, we use the Adam optimizer [**?**] with weights $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is set to $2e$-5. For adversarial training, we optimize the discriminator and generator alternatively.

C. Image manipulation results

In Fig. 3 and Fig. 4, we provide results on appearance sampling and pose sampling for the DeepFashion dataset as an extension of Fig. 1 in the main paper. For each factor, we sample the embedding feature from Gaussian noise and fix the other factors by using the embedding feature extracted from the real data as explained in Sec. 4.2 in the main paper.

D. Pose-guided person image generation results

For pose-guided person image generation, we provide more generated results. As an extension of Fig. 5 in the main paper, Fig. 5 shows the generated images of one appearance with various real poses selected randomly from DeepFashion.

E. Inverse interpolation results

In this section, we provide more inverse interpolation results in Fig.6 as an extension of Fig. 6 in the main paper. For two images x_1 and x_2, we find the corresponding Gaussian codes z_1 and z_2 as explained in the Sec. 4.4 of the main paper. As shown in Fig. 6(a)(b), our method successfully generates the intermediate states between two images of the same person. Note that, the inverse interpolation between two images of different persons is more challenging (see Fig. 6(c)) since we need to interpolate both the appearance and pose.

*Corresponding author

F. Image sampling results

We also give more sampling results as extensions of Fig.7 in the main paper. Fig. 7 shows the sampling results (a-e) and real images (f) on Market-1501 dataset. VAE generates blurry images and DCGAN sharp but unrealistic person images. In contrast, our model generates more realistic images (c)(d)(e). By comparing (d) and (c), we observe that our model using body ROI generates more sharp and realistic images whose colors on each body part are more natural. By comparing (e) and (d), we see that when sampling foreground and background but using the real pose keypoints randomly selected from the training data, we generate better results.

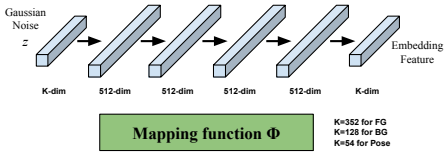

Figure 1: Network architecture of the mapping functions for FG, BG and Pose in stage-II.

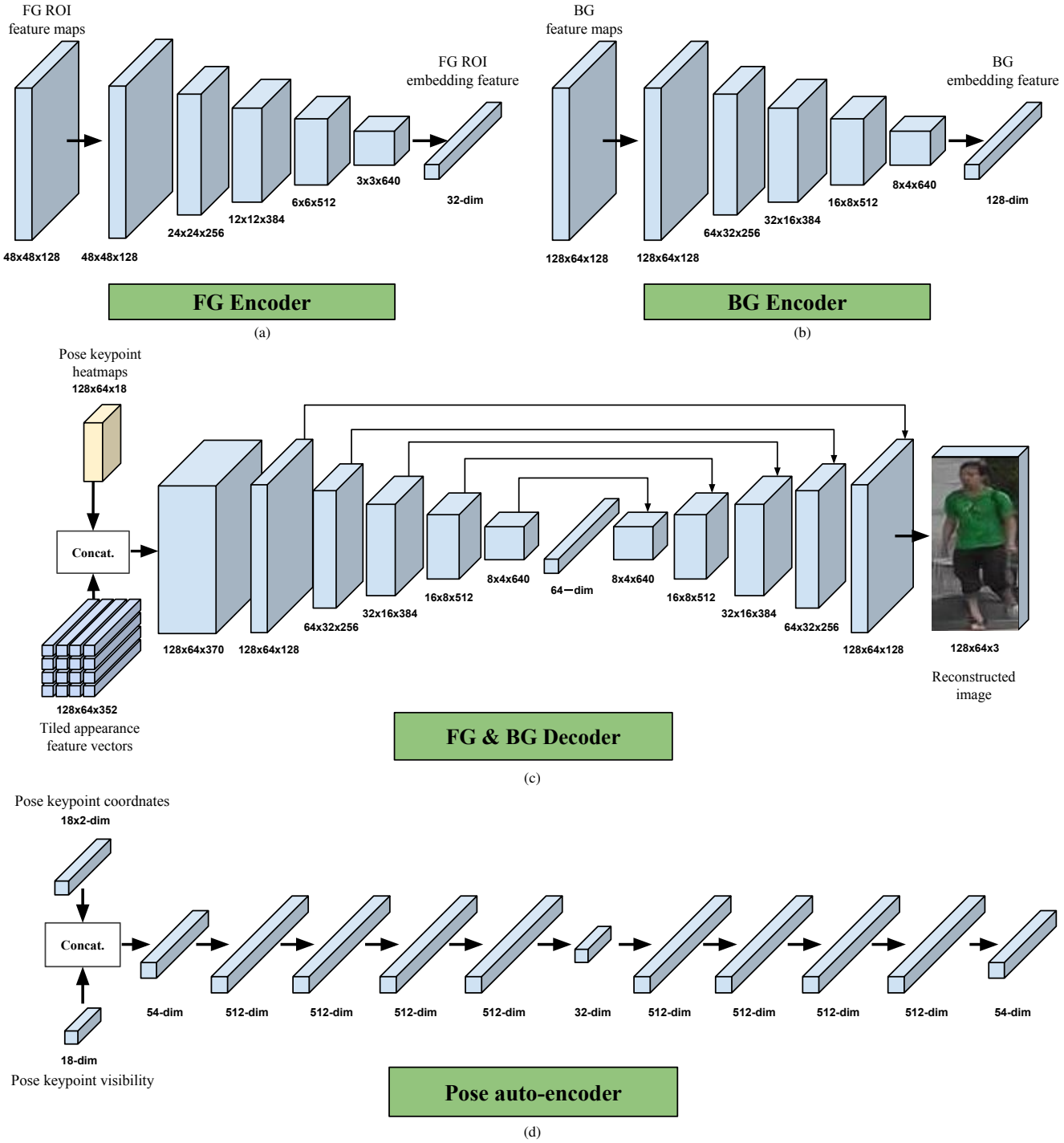

Figure 2: Network architectures of stage-I. (a) FG encoder, fed with the extracted 7 FG body ROI feature maps and outputting 7 FG embedding features of 32-dim after 5 convolutional residual blocks. (b) BG encoder, fed with the BG feature maps and outputting a BG embedding feature of 128-dim after 5 convolutional residual blocks. (c) FG and BG decoder, fed with the concatenated appearance and pose feature maps and outputting the generated image after the "U-net"-based [**?**] architecture. (d) Pose auto-encoder, fed with the concatenated keypoint coordinates and visibility vector and outputting the reconstructed vector after the auto-encoder.

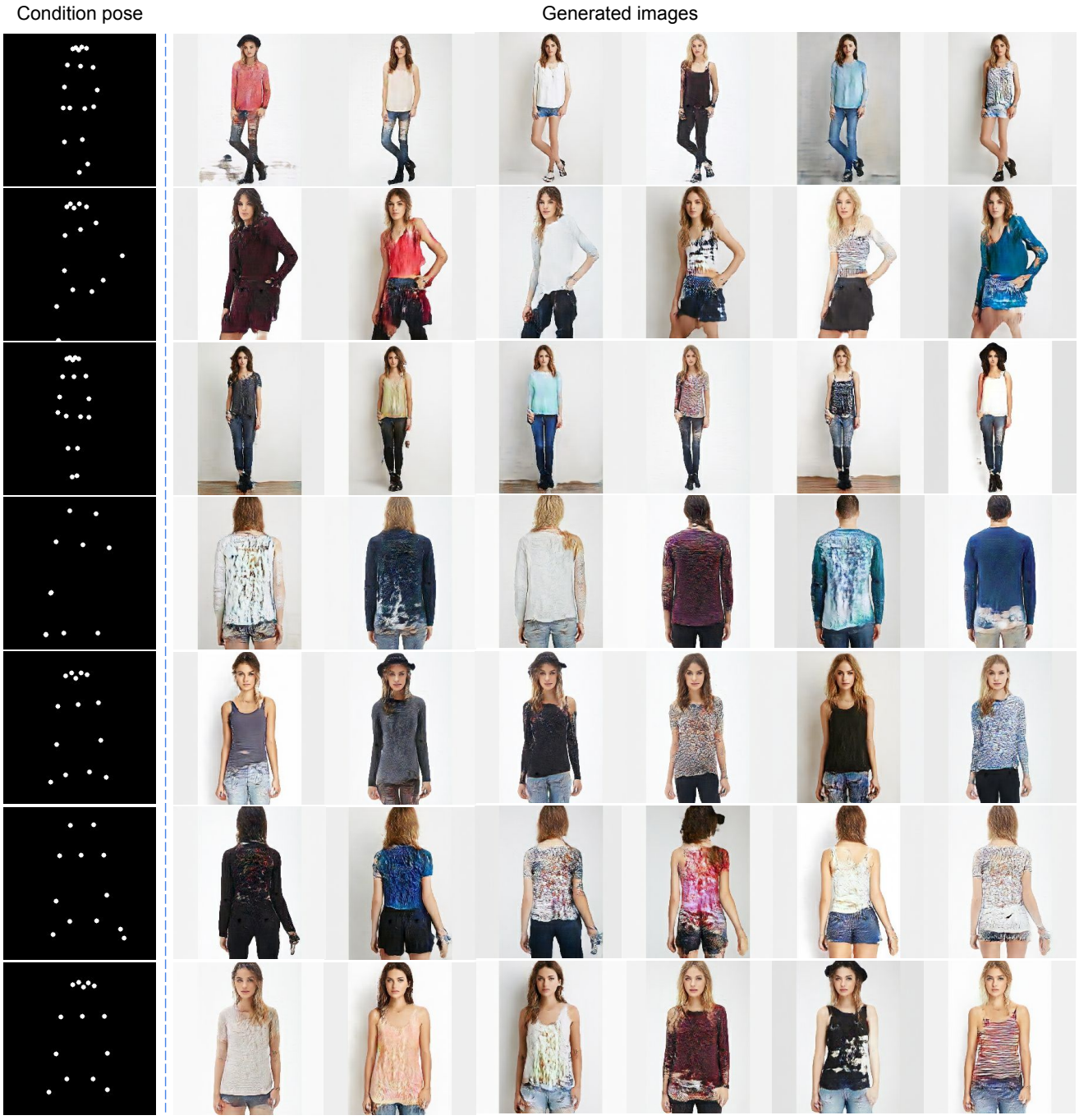

Figure 3: Appearance sampling (fixed Pose) results on the DeepFashion dataset. In each row, 6 different appearance factors are sampled from Gaussian noise and the pose factor is fixed to a real one.

Figure 4: Pose sampling (fixed Appearance) results on the DeepFashion dataset. In each row, 6 different pose factors are sampled from Gaussian noises and the appearance factor is fixed to a real one.

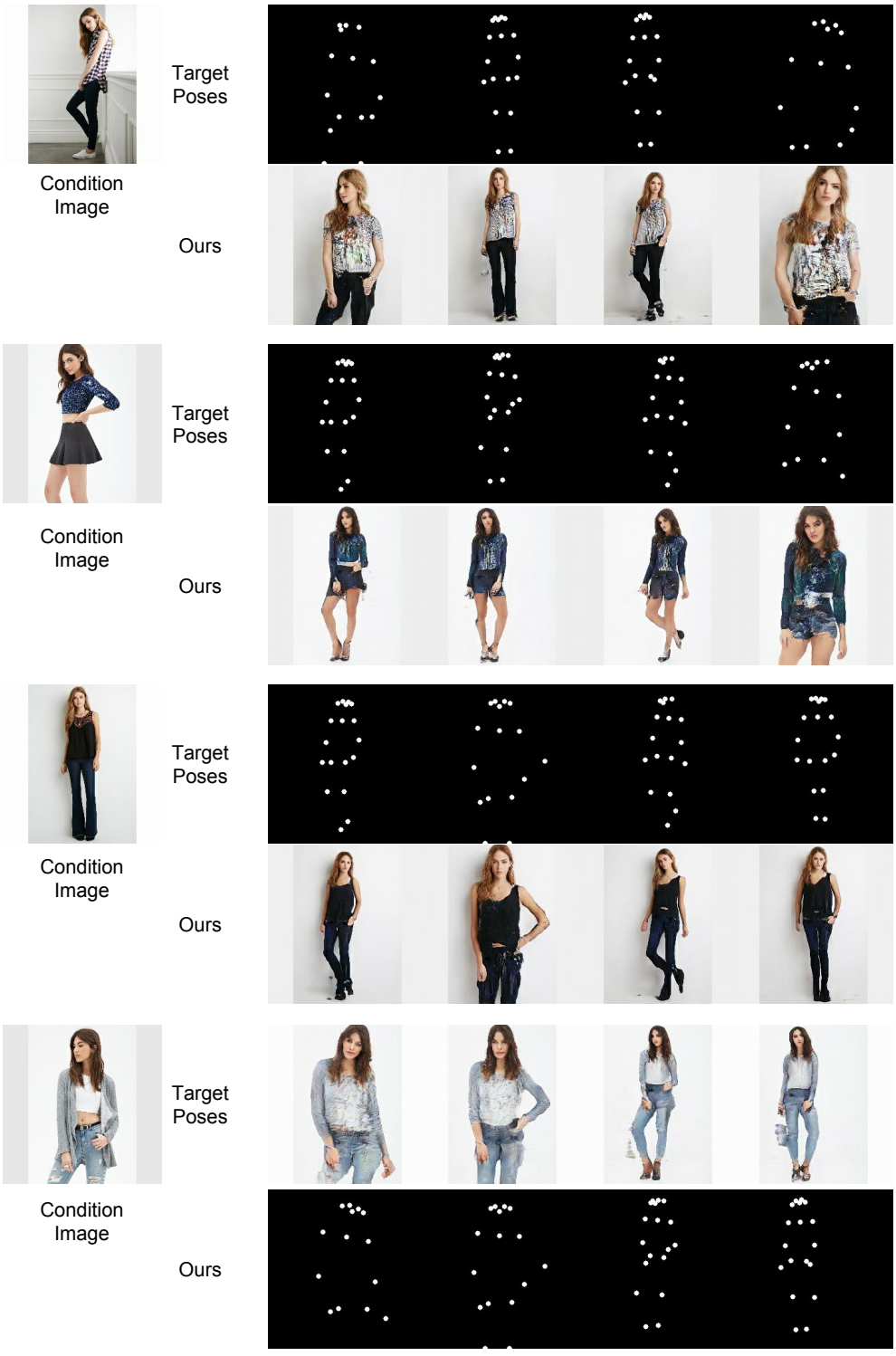

Figure 5: Generated results for one appearance with various poses on the DeepFashion dataset.

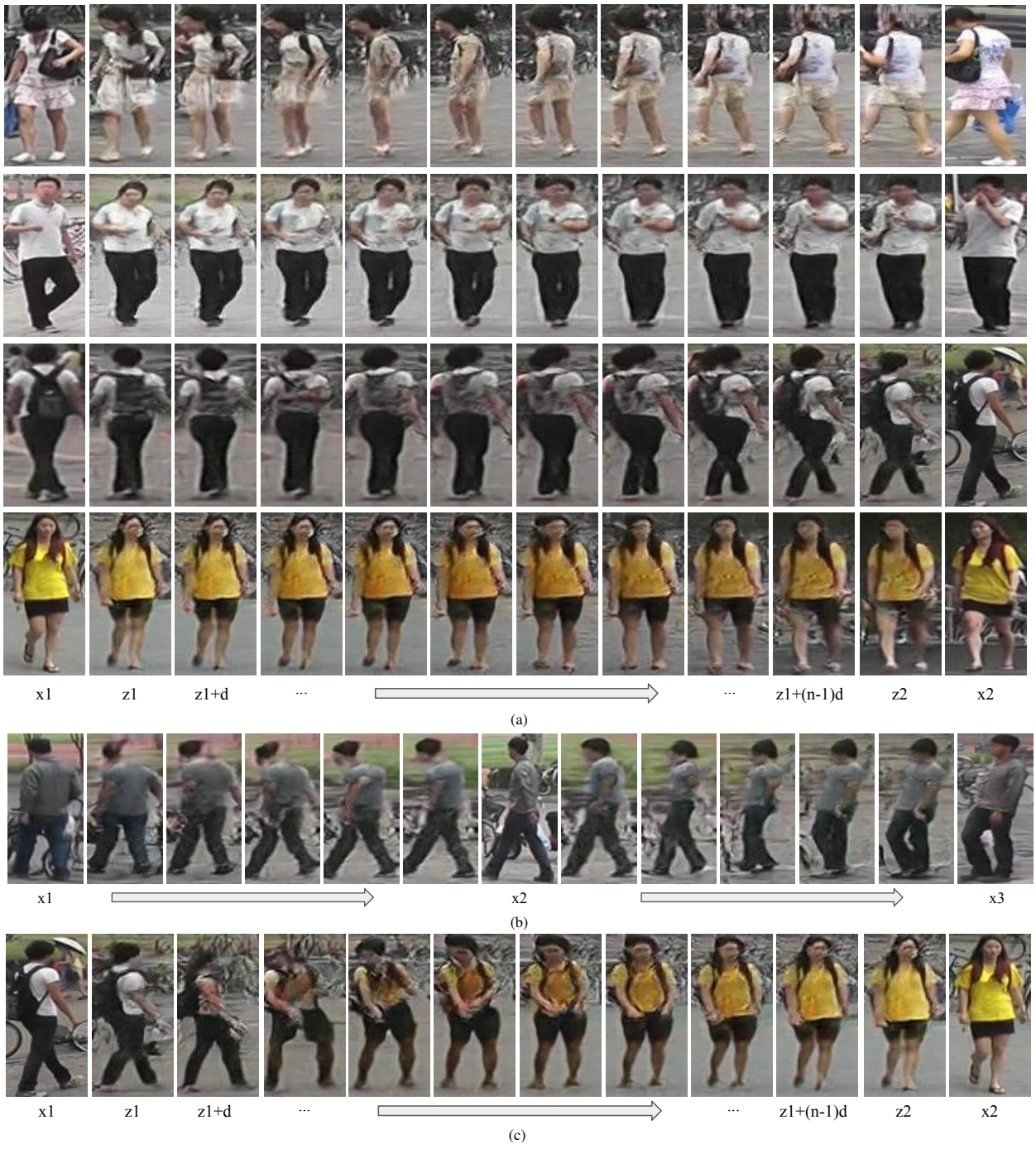

Figure 6: Inverse interpolation results on Market-1501. (a) Interpolation between two images of the same person. (b) Interpolation between three images of the same person. (c) Interpolation between two images of different persons.

(a) Vanilla VAE (b) Vanilla DCGAN (c) Ours - Whole Body

(d) Ours - BodyROI7 (e) Ours - BodyROI7 with real pose from training set (f) Real data

Figure 7: Sampling results. (a) Vanilla VAE; (b) Vanilla DCGAN; (c) Ours - Whole Body; (d) Ours - BodyROI7; (e) Ours - BodyROI7 pose with real pose from training set; (f) Real data.