Supplementary Material of Paper: Text Guided Person Image Synthesis

Xingran Zhou[1] Siyu Huang[1*] Bin Li[1] Yingming Li[1] Jiachen Li[2] Zhongfei Zhang[1]

[1] Zhejiang University [2] Nanjing University

{xingranzh, siyuhuang, bin_li, yingming, zhongfei}@zju.edu.cn, jiachen_li_nju@163.com

The supplementary material includes additional discussions on the details about pose inference (§A), the implementation details (§B), the network architecture (§C), the model generalization (§D), and the visualization of multi-scale attention (§E), respectively.

We also show more examples of the pose and attribute transferred person image generation, the interactive editing, and the synthesis images (§F). Our project page is https://xingranzh.github.io/text_guided_synthesis/.

A. Details of Pose Inference

Most of the existing work on human pose keypoints is about detection. That is to say, an image is the input to the deep neural nets and the output is heatmaps. In the common setting, the real labels are annotated for the original image; so it is natural to calculate the distance between the predicted heatmaps and the real labels. We observe that the deviations of facial joints are relatively slight compared to the swing amplitudes of arms and legs, such that the coordinate regression method may lead to the distortion of facial joints. In addition, some of the key words in the text, such as walking, are not explicitly explained to be whether the left or right leg; thus the predicted pose may either be the shape of striding on both two legs or staying upright. In short, the regression-only method may cause an unnatural prediction of pose (*e.g.*, Fig. [9] (b) in the paper).

We implement a coordinate regression method which serves as a baseline for our text-guided pose generator. Specifically, after selecting the orientation, the feature of the selected basic poses and the sentence representation vector are concatenated and fed into the fully-connected layers. The objective function is the distance between the regression coordinates and the real ones. We introduce a paired relative offset L2 distance

$$\frac{1}{J}\sum_{i=1}^{J} \min_{\text{left,right}} \sqrt{(x'_i - x_i)^2 + (y'_i - y_i)^2} \qquad (1)$$

where (x_i, y_i) and (x'_i, y'_i) are respectively the offsets relative to one of its own calibration coordinates for the real and the inferred poses. The term $\min_{\text{left,right}}$ denotes the minimum value for symmetrical joints (such as the pair of left and right hands). J is the number of human joints.

Model	Distance
Coordinate regression	6.996
Our T2P model	3.729

Table 1: Paired relative L2 distance (px). The error of our method is about less than 4 pixels on average. Our method provides more competitive pose inference results than the coordinate regression method.

Table 1 shows the quantitative comparison result. The error of our method is less than 4 pixels on average, and the error of coordinate regression is up to 7 pixels on average. For an 128×64 image, the large deviation denotes an obvious visual divergence.

Fig. 3 shows more pose inference results by our method. The predicted poses are vivid and valid when the model encounters specific words involving movements (such as walking is related to the legs, carrying is about the arms). The corresponding joints change to the right position in a larger offset, which makes the generated posture closer to the real one.

B. Implementation Details

Following the previous works, we resize all the images in the dataset into the shape of $128 \times 64 \times 3$. We use OpenPose [1] to get human body poses. The representation of

*Corresponding author

keypoint heatmap $p \in \mathbb{R}^{128\times64\times18}$ is the same as that of PG^2 [3], where 18 is the number of human joints. Since the original descriptions rarely contain the information of person orientation, we manually add some words into the text to describe the orientation. For instance, if the person in an image is facing toward the camera, we add words like "He/She is walking forward". In Stage-II, we use the real target poses for training. The poses inferred by Stage-I are used in both testing and sampling.

The Stage-I and Stage-II are trained separately. For adversarial training, we optimize the discriminator and generator alternatively. All networks are optimized using Adam optimizer with a learning rate of 0.0002, a momentum of 0.5, and a batch size of 32. We train Stage-I for 150,000 steps and we train Stage-II for 200,000 stpes. In our experiment, we adopt hyperparameters $\lambda_1 = \lambda_2 = 1$ for Eq. 3 and $\gamma_1 = 0.5, \gamma_2 = 30$ for Eq. 9 in the paper.

C. Network Architecture

Fig. 4 shows four network architectures used in our two-stage framework: (a) Pose generator at Stage-I consists of five transposed convolution layers; (b) both Image encoder and pose encoder at Stage-II consist of three convolutional residual blocks for downsampling; (c) Attentional Upsampling modules at Stage-II. We show the structure in the special case (when $i = 1$) and the normal case (when $i \neq 1$). The architecture can be calculated in the recursive way.

All the convolutional layers consist of the 3×3 convolution, batch normalization, and ReLU activation. In both stages, we use the convolutional residual blocks which are similar to that in [2, 3], where each residual block has two stride-1 convolution layers, followed by a stride-2 convolution layer.

In Stage-I, we adopt $K = 8$ basic poses. The orientation selection net F^{ori} is the fully-connected layers. The pose generator consists of five 4×4 stride-2 transposed convolution layers with batch normalization and ReLU for upsampling.

In Stage-II, we use $m = 3$ attentional upsampling (AU) modules for image generation. Each AU consists of an attention layer F^{attn}, a convolutional layer, and an upsampling layer F^{up}. The attention layer F^{attn} is similar to that in [5]. The upsampling layer F^{up} consists of two convolutional residual blocks followed by nearest-neighbor upsampling.

The pose encoders in Stage-I and Stage-II consist of five and three convolutional residual blocks, respectively. The filter numbers of the convolutional residual blocks increase linearly. For discriminators in both two stages, we use four convolutional layers adapted from AttnGAN [5].

D. Model Generalization

Fig. 1 shows the results of inputting descriptive text without attribute words. The person poses in the generated images are exact to those of the target images while the color attributes are different.

Figure 1: We delete the attribute words in the text. The person poses in the generated images are exact to those of the target images. The person identities are maintained in the generated images.

It demonstrates that

- Our pose inference depends on words about motions and directions in the text.
- In pose and attribute transferred synthesis, our method is able to transfer the appearance of a person according to attribute words.
- The person identities in the source images are maintained, as the person in a generated image looks like the same as that in the source images.

E. Visualization of Multi-scale Attention

To better understand what has been concerned by the attentions of different scales, we visualize some intermediate results.

Fig. 2 shows the regions with attention corresponding to words. We can see that attentions of the mid and small scales relate more to the abstract and holistic concepts. For the attention of large scale which is resized once by ConvNets, the identity information of the reference image is preserved, making the generated images more visually invariant w.r.t. person identity. Intuitively, the attention of a large scale acts on relatively shallow feature maps, which functions similarly to the skip-connection in the U-Net framework [4].

F. More Experimental Results

More experimental results are given in Fig. 5, Fig. 6, and Fig. 7.

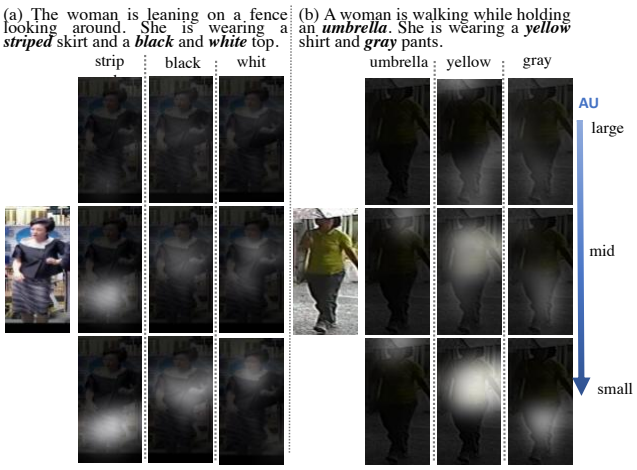


Figure 2: Visualization of our multi-scale attentional upsampling (AU) module. The AUs of small and middle scales lean to the abstract and holistic concepts. The AU of a large scale captures more detailed information, which works similarly to the skip-connection in U-Net.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1

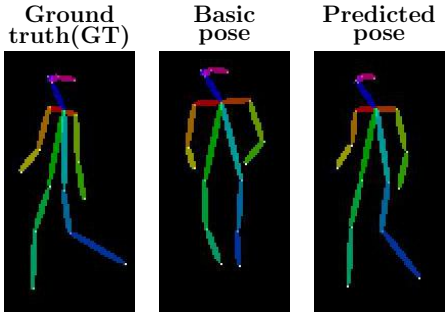
[2] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 2

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

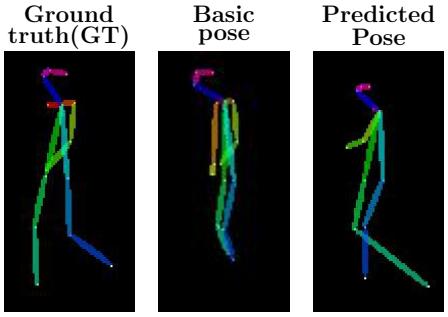
[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2

[5] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2

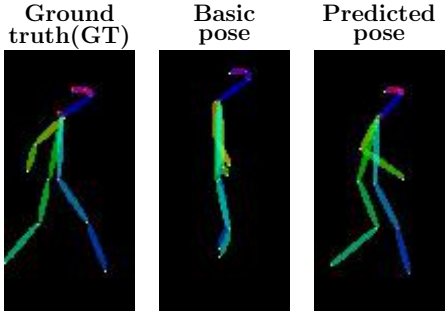
(a) The girl has her hair in a pony tail. She is wearing a white shirt, black jacket, dark jeans and heeled sandals. She is also wearing a backpack and ***has a bag in her left hand***. She ***is walking toward the forward left side***.



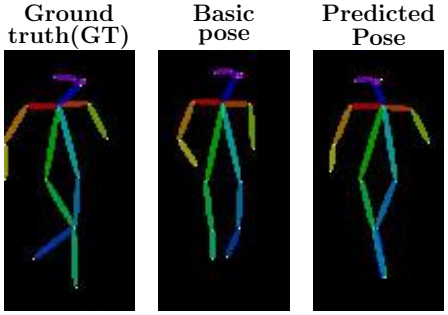
(b) A woman is wearing a green shirt and black and white skirt. She has on sandals and white socks. She ***is walking toward the left***.



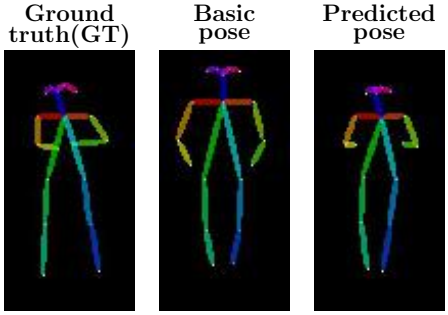
(c) A man with very short hair blue shirt and denim shorts and sandals ***carrying a black backpack***. He ***is walking toward the right***.



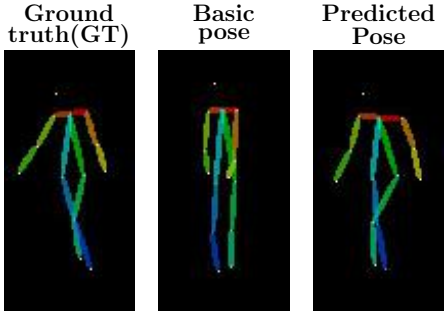
(d) A woman with short black hair is wearing a grey shirt black pants and platform sandals. she ***is walking toward the forward right side***.



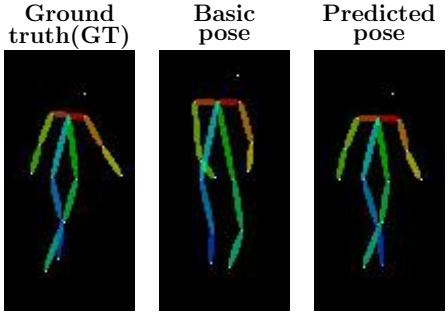
(e) A lady with dark hair wearing a white dress beige shoes and ***carrying a blue purse***. She ***has head inclined and facing toward the camera***.



(f) She has short black hair she is also wearing bright blue denim pants and is ***carrying a small white bag***. She ***is walking to the forward left side***.



(g) A woman wearing a black shirt with gold spots a pair of blue jeans and a pair of red shoes. She ***is walking to the backward right***.



(h) The man is wearing a brown fleece like collared jacket with black pants and black shoes. He ***is walking backward***.

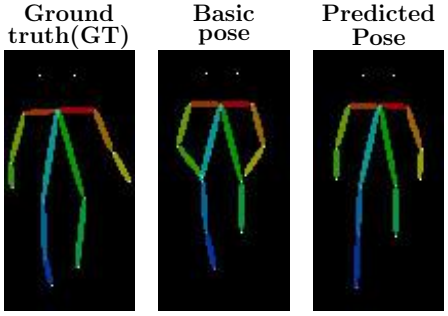


Figure 3: Examples of pose inference (supplement to Fig. 9 in the paper).

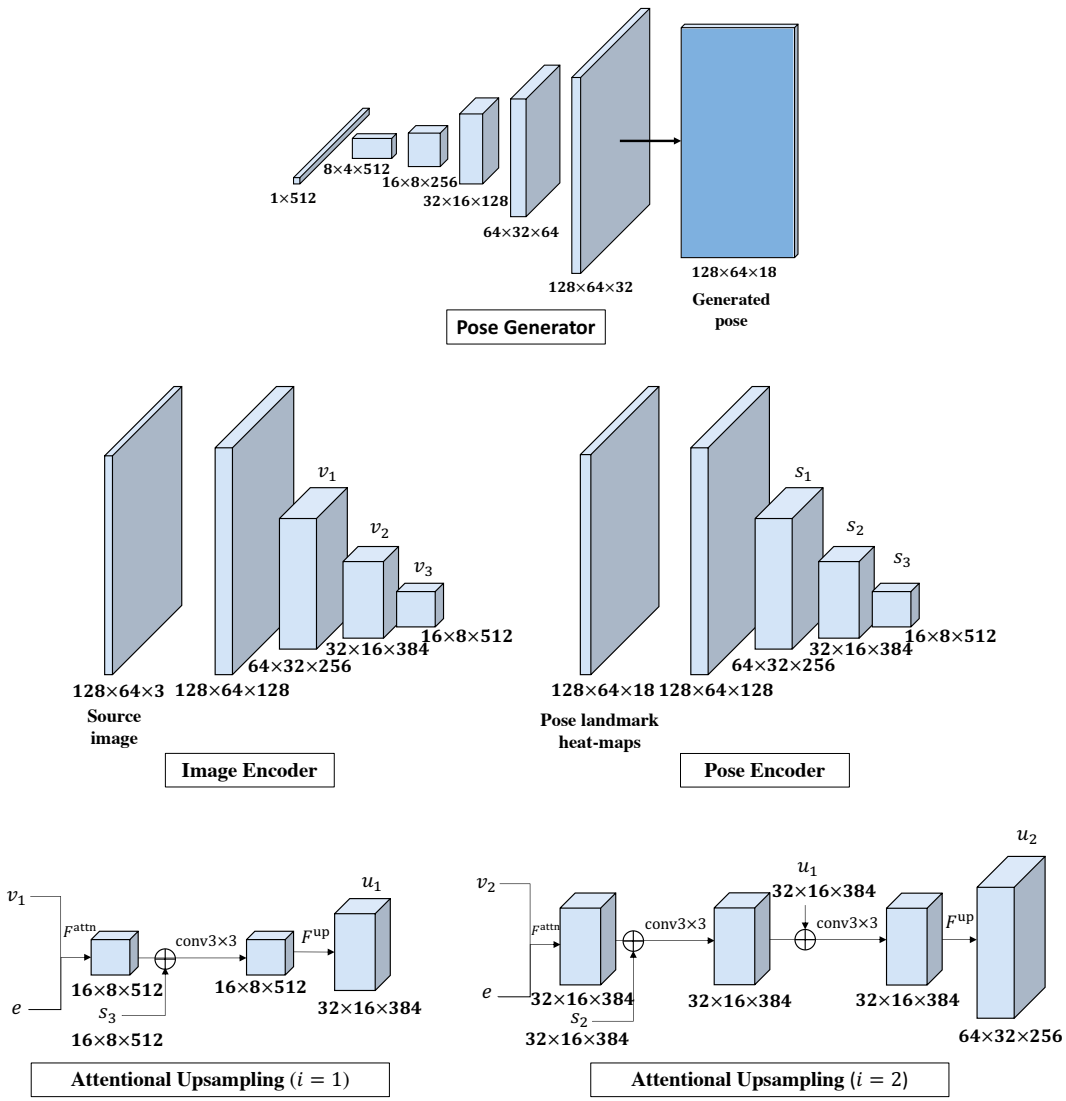


Figure 4: Network architectures of our model.

(a) He is wearing [*a* ***white*** *sweater*] with a collared shirt underneath. He has casual pants on and is carrying a white paper in his hand. *He is walking toward backward.*

(b) A man wearing [*a* ***yellow*** *shirt*] a pair of dark colored shorts and a pair of black shoes. *He is walking forward to the camera.*

(c) The man is wearing [*an* ***orange*** *short sleeved shirt*] and light colored long shorts that fall to mid-calf length. *He is walking toward backward.*

(d) The man has on [*a* ***blue*** *shirt*] and long baggy tan pants with white shoes. *He is facing to the forward right side.*

(e) The woman has long, straight dark hair past her shoulders. She is wearing [*a* ***red*** *jacket*], black pants and white sneakers with black stripe. *She is walking toward the left.*

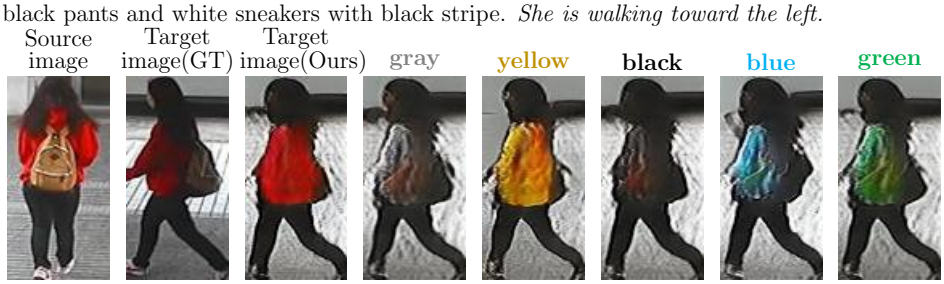


Figure 5: Examples of our text guided person image synthesis (supplement to Fig. 6 in the paper).

(f) He is wearing [*a **white** tea shirt*] with dark colored neck and armhole trim. His jeans and sneakers are black. *He has head inclined toward the left backward side.*

(g) The individual is wearing [*a **pink** shirt*]. The individual is wearing black Capri's and sandals. Individual has dark hair. *It is walking toward the left backward side.*

(h) The man is wearing [*a **purple** jacket*] with grey colored pants with grey shoes with black trim. *He is walking toward forward.*

(i) A man wears jeans, [*a **white** zip-up hoodie*] , and white shoes. He has black hair and glasses. *He is walking toward the right forward side.*

(j) A man in [*a **blue** shirt*], a pair of black shorts and a pair of black and grey shoes. *He is walking toward the left backward side.*

Figure 5: Examples of our text guided person image synthesis (supplement to Fig. 6 in the paper).

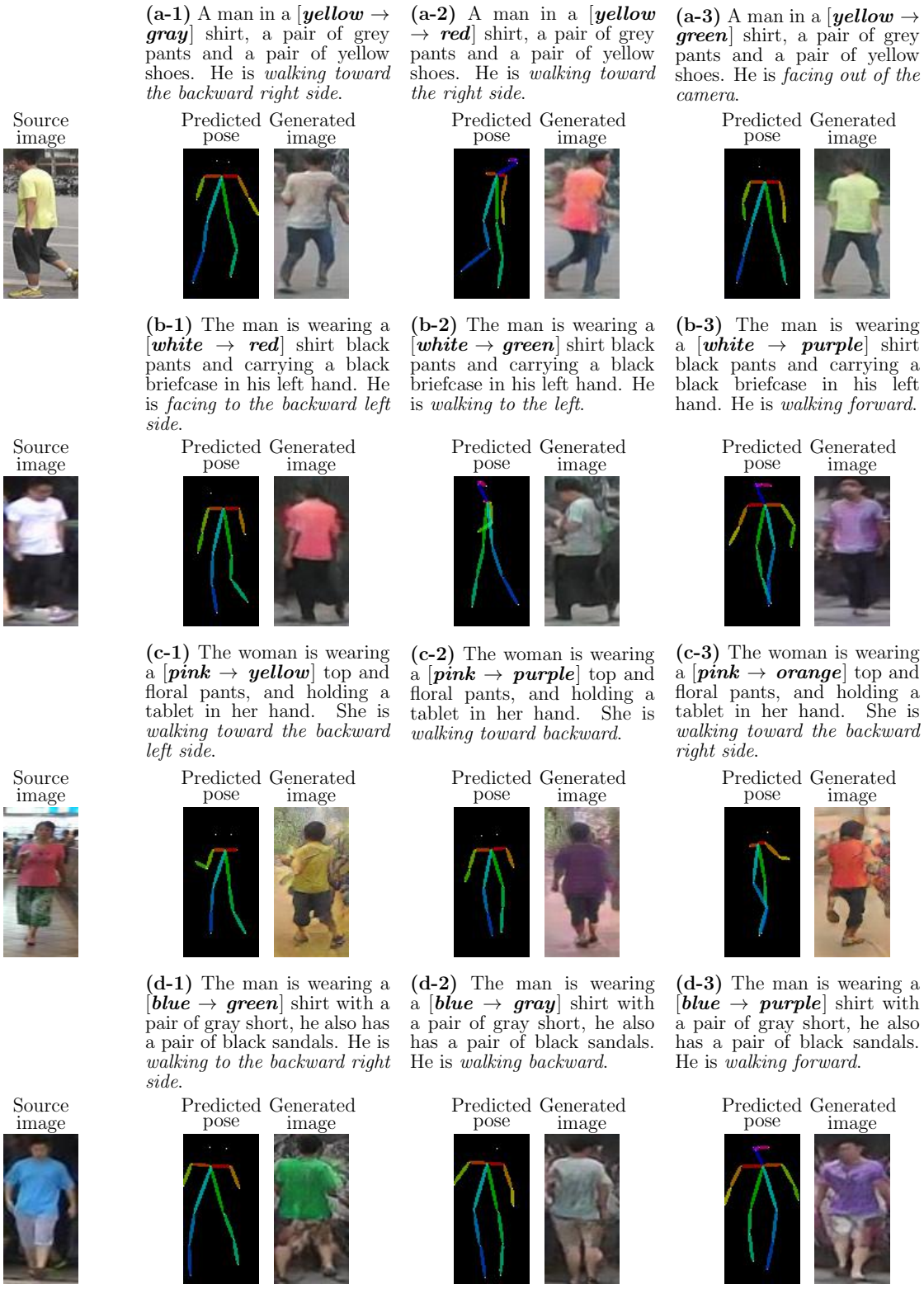


Figure 6: Examples of interactive editing (supplement to Fig. 7 in the paper).

(a) mSIS

(b) mAttnGAN

Figure 7: Synthesis images by different methods. Each column contains two pairs of generated images. In each pair, the upper image is pose transfer (PT) synthesis while the bottom image is pose and attribute (P&AT) transfer synthesis.

(c) mPG^2

(d) SAU

Figure 7: Synthesis images by different methods. Each column contains two pairs of generated images. In each pair, the upper image is pose transfer (PT) synthesis while the bottom image is pose and attribute (P&AT) transfer synthesis.

(e) Ours

Figure 7: Synthesis images by different methods. Each column contains two pairs of generated images. In each pair, the upper image is pose transfer (PT) synthesis while the bottom image is pose and attribute (P&AT) transfer synthesis.