Multistage Adversarial Losses for Pose-Based Human Image Synthesis

CVPR 2018
Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR)
Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),Institute of Automation, Chinese Academy of Sciences (CASIA)
University of Chinese Academy of Sciences(UCAS)

汇报人：梁天保

Motivation

Method Overview

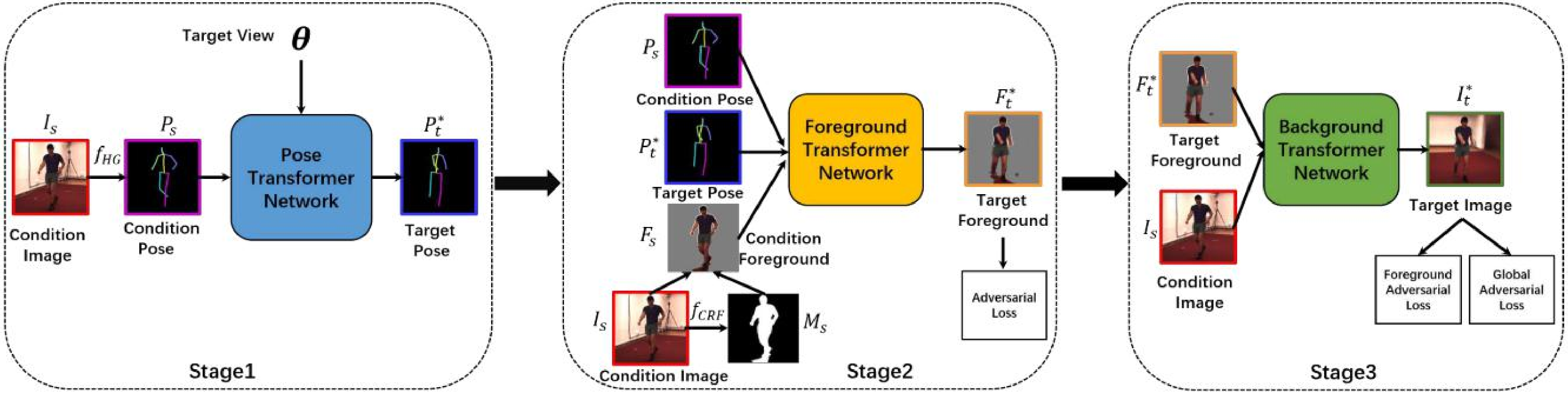


Figure 2. The overall pipeline of our multistage approach which contains three transformer networks for three stages. In the first stage, the pose transformer network synthesizes a novel view 2D pose. Then, the foreground transformer network synthesizes the target foreground image in the second stage. Finally, the background transformer network generates the target image. f_{HG} and f_{CRF} donate the stacked hourglass networks [18] and the CRF-RNN [30] for pose estimation from image and foreground segmentation, respectively.

Pose Transformer Network

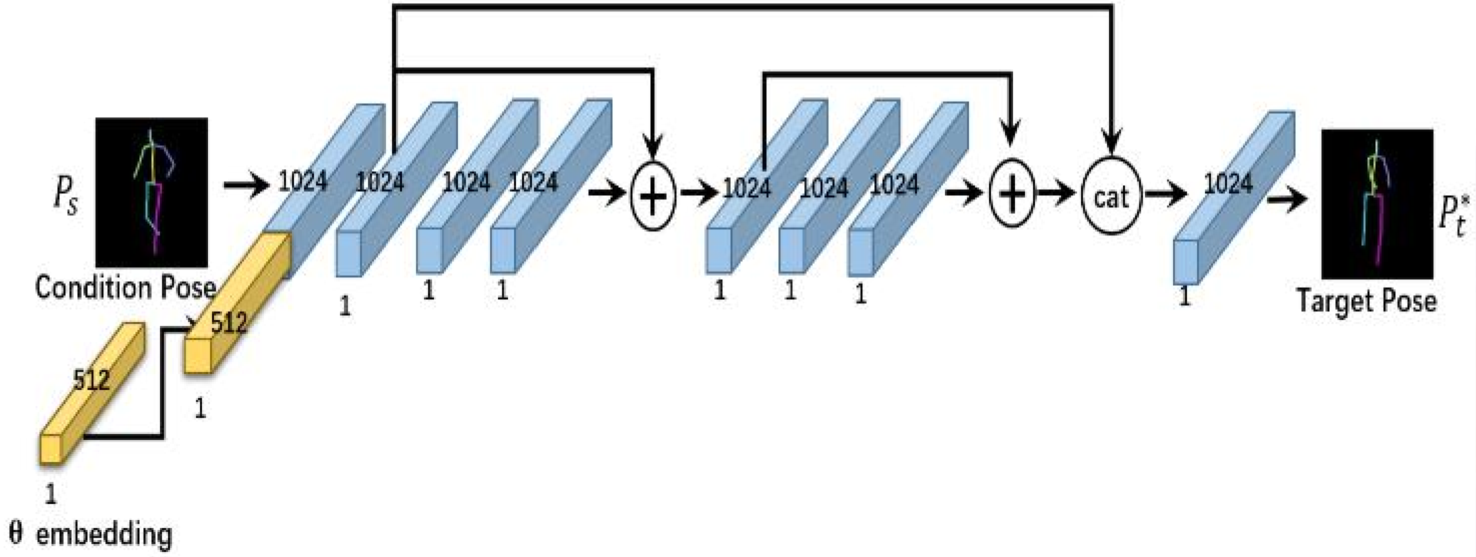


Figure 3. The architecture of the pose transformer network.

$$P_t^* = G_p(P_s, \theta)$$

$$\mathcal{L}^1 = \sum_{i}^{N} \left\| {P_t^*}^i - {P_t}^i \right\|_2^2$$

Foreground Transformer Network

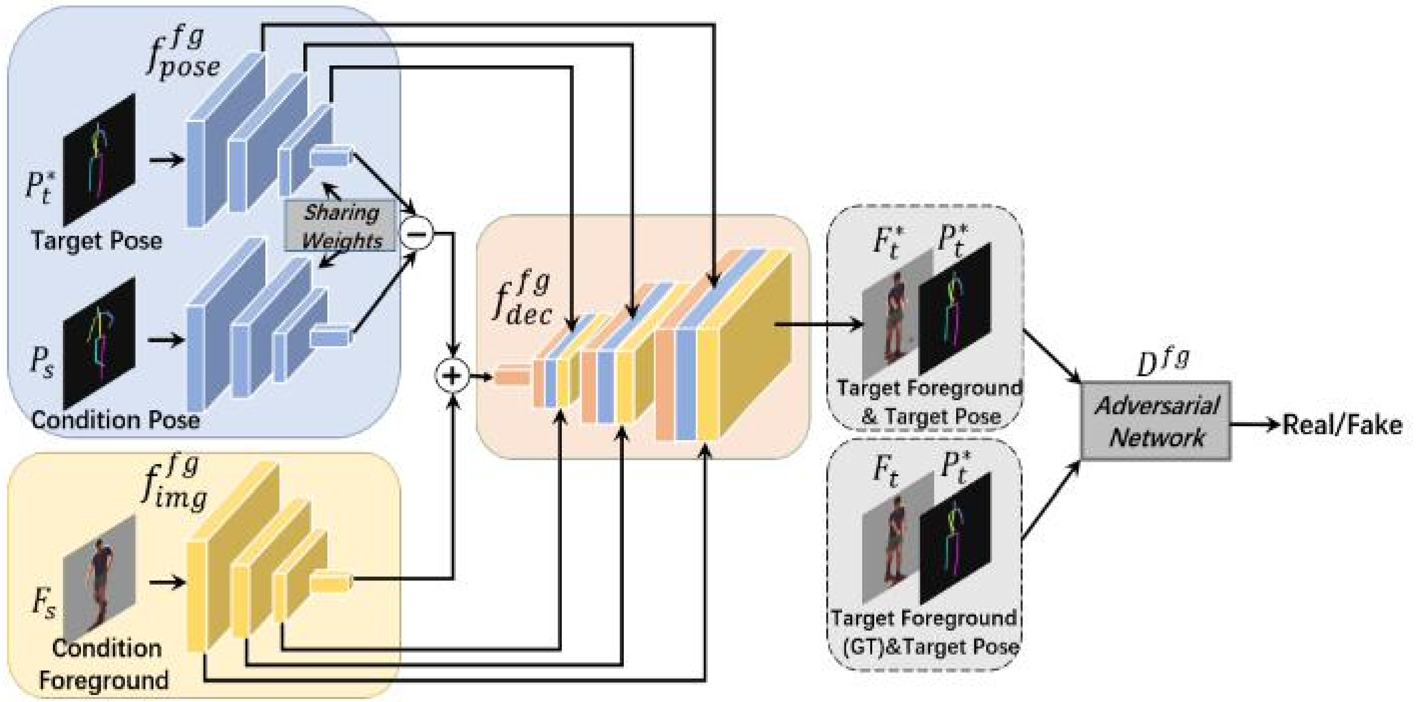


Figure 4. The architecture of the foreground transformer network.

$$F_t^* = f_{dec}^{fg}(f_{pose}^{fg}(P_t^*) - f_{pose}^{fg}(P_s) + f_{img}^{fg}(F_s))$$

Foreground Transformer Networks

$$\mathcal{L}^2 = \alpha_f \mathcal{L}^2_{fg} + \beta_f \mathcal{L}^2_{bg} + \mathcal{L}^2_{gen}$$

$$\begin{aligned}\mathcal{L}^2_{fg} &= \|F_t \odot M_t - F^*_t \odot M_t\|_1 \\ &= \frac{1}{\sum\limits_{M_t^{i,j}=1} M_t^{i,j}} \sum_{i,j} \left|({F_t}^{i,j} - {F^*_t}^{i,j}) \times {M_t}^{i,j}\right|\end{aligned}$$

$$\mathcal{L}^2_{gen} = -\log(D^{fg}([F^*_t, P^*_t]))$$

$$\begin{aligned}\mathcal{L}^2_{bg} &= \|F_t \odot (1 - M_t) - F^*_t \odot (1 - M_t)\|_1 \\ &= \frac{1}{\sum\limits_{M_t^{i,j}=0} (1 - M_t^{i,j})} \sum_{i,j} \left|({F_t}^{i,j} - {F^*_t}^{i,j}) \times (1 - {M_t}^{i,j})\right|\end{aligned} \quad (6)$$

$$\begin{aligned}\mathcal{L}^2_D = &-\log(D^{fg}([F_t, P^*_t])) \\ &-\log(1 - D^{fg}([F^*_t, P^*_t]))\end{aligned}$$

Background transformer network

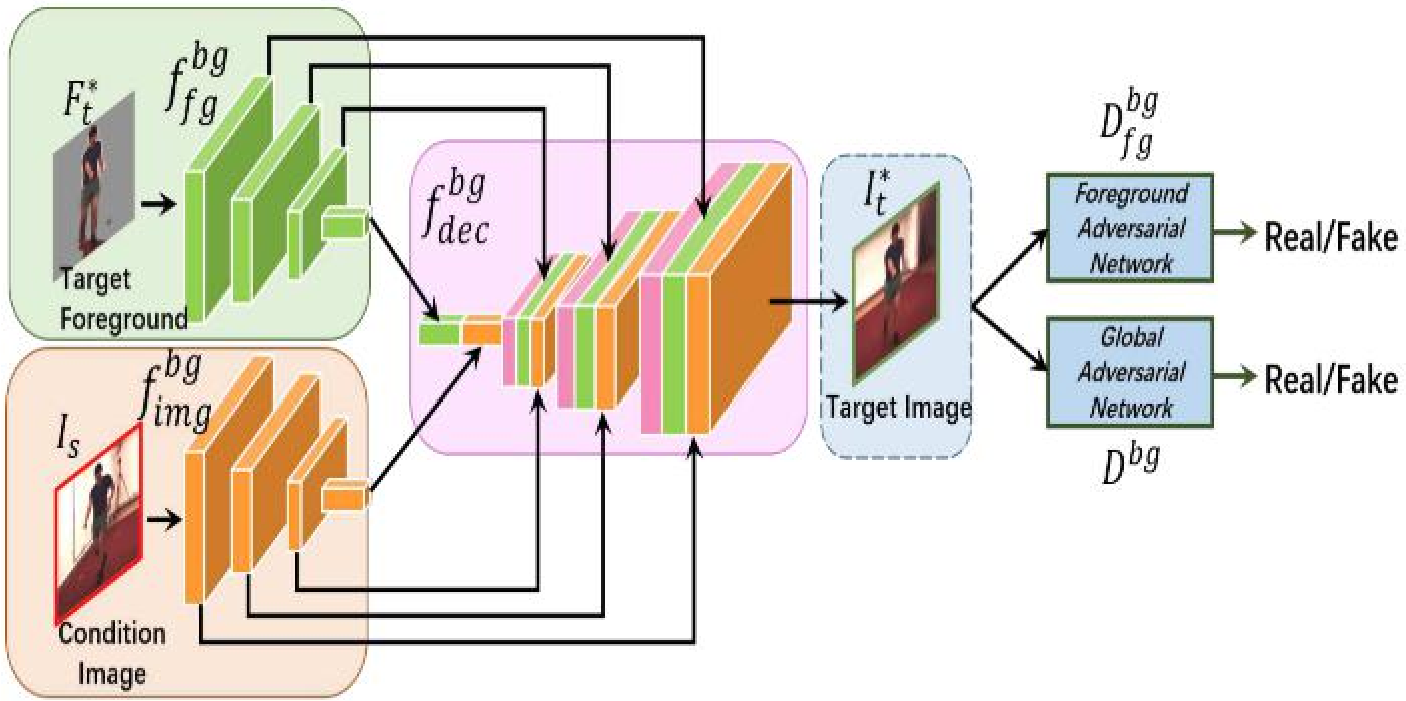


Figure 5. The architecture of the background transformer network.

Background Transformer Networks

$$\mathcal{L}^3 = \alpha_b \mathcal{L}^3_{fg} + \beta_b \mathcal{L}^3_{bg} + \mathcal{L}^3_{gen_{fg}} + \mathcal{L}^3_{gen}$$

$$\mathcal{L}^3_{gen_{fg}} = -\log(D^{bg}_{fg}([I^*_t \odot M_t, P^*_t]))$$
$$\mathcal{L}^3_{gen} = -\log(D^{bg}(I^*_t))$$

$$\begin{aligned}\mathcal{L}^3_{D_{fg}} = &-\log(D^{bg}_{fg}([I_t \odot M_t, P^*_t])) \\ &-\log(1 - D^{bg}_{fg}([I^*_t \odot M_t, P^*_t]))\end{aligned}$$
$$\mathcal{L}^3_{D} = -\log(D^{bg}(I_t)) - \log(1 - D^{bg}(I^*_t))$$

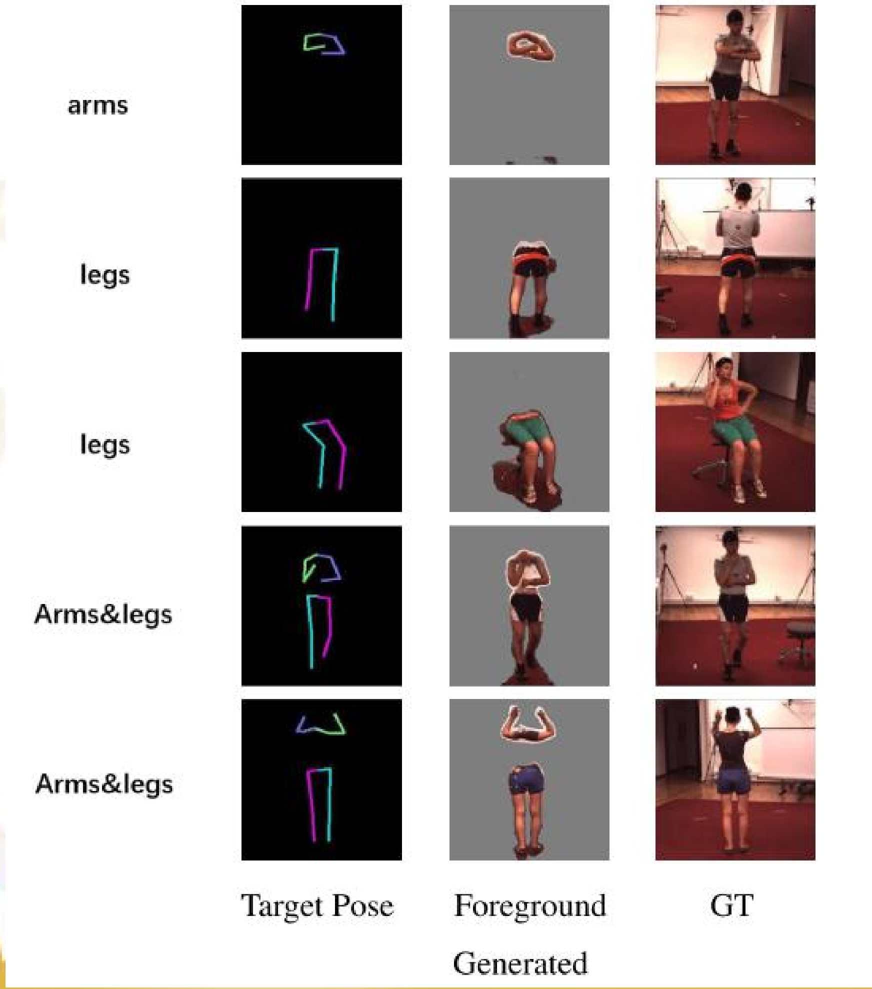
Experiment

Figure 6. Visualization of the synthesized images from three state-of-the-art methods, two baselines and our model. Our method achieves the best results with clear foreground and background.

Experiment

Figure 7. Visualization of the synthesized images from four foreground baselines and our foreground transformer network.

Experiment


arms
legs
legs
Arms&legs
Arms&legs
Target Pose
Foreground Generated
GT

Experiment

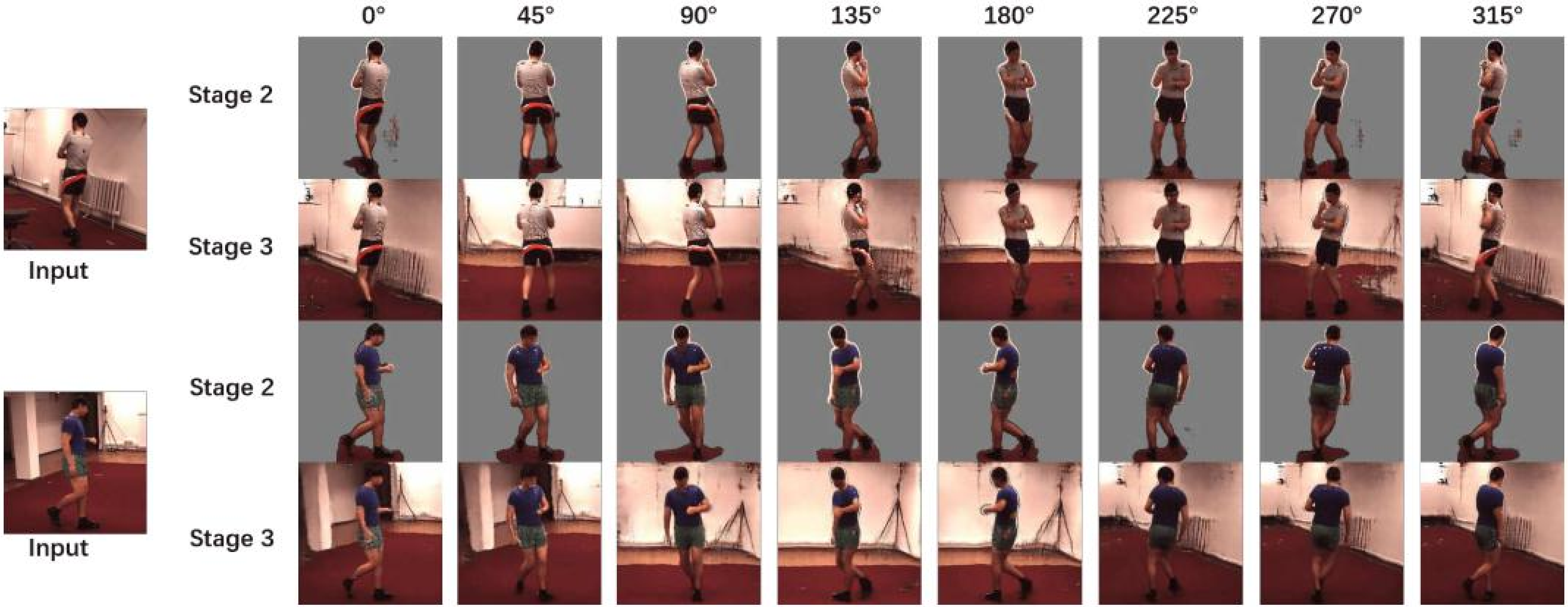


Figure 9. Multiview human images generated by our model.

Experiment

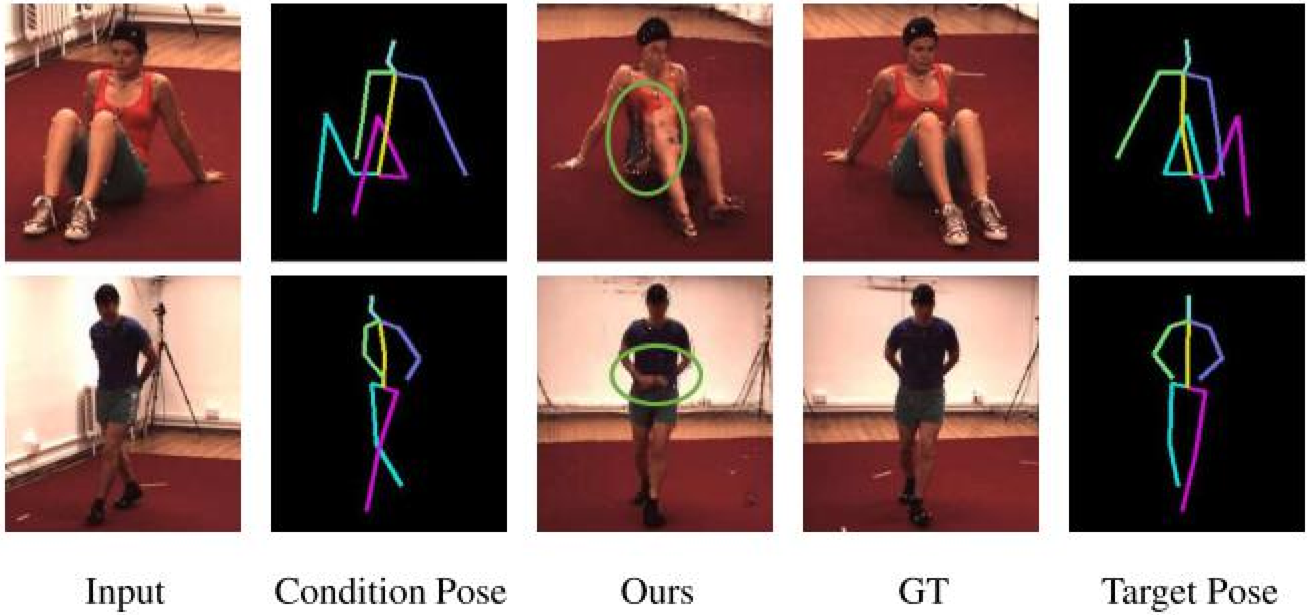


Figure 10. Two failure cases of our model.

THANKS