Pose Guided Person Image Generation for Novel 3D View Synthesis


Byeongjoo


Anqi Yang

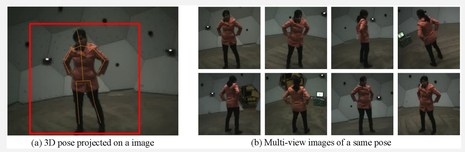

Sihan You

Introduction

- Goal
  - Given a single view of a person in an arbitrary pose, to synthesize an image of a person after a specified transformation of viewpoint
- Challenges
  - Inferring the appearances of unobserved parts in input image.
  - Preserving the global shape between input image and synthesized novel view image.
  - Utilizing existing clothing dataset for novel 3D view synthesis.
- Main idea
  - Utilize pose information as explicit guidance to the viewpoint change.
  - Since the viewpoint change is equivalent to the pose change of body (i.e., rotation), pose includes the essential information for novel 3D view synthesis to preserve the global shape.
  - Given input images, we first extract 3D keypoint for pose estimation, and project the extracted 3D pose into the desired target view.
  - Then, we synthesize novel 3D view images using the variant of VAE and GAN based on the pose guidance.

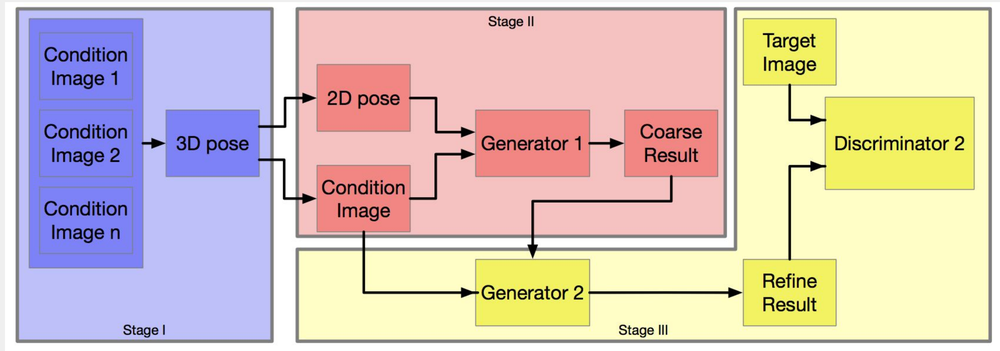
Dataset

- Deep Fashion (256×256)
  - Resolution 256×256
  - 146,680 pairs
- Market dataset (128×64)
  - Resolution 128×64
  - 439,420 pairs
- CMU Panoptic studio dataset (256×256)
  - Resolution 256×256
  - 175,272 pairs


(a) 3D pose projected on a image (b) Multi-view images of a same pose

Images from Panoptic studio dataset

Technical Details



Stage I: Multi-view pose skeleton generation

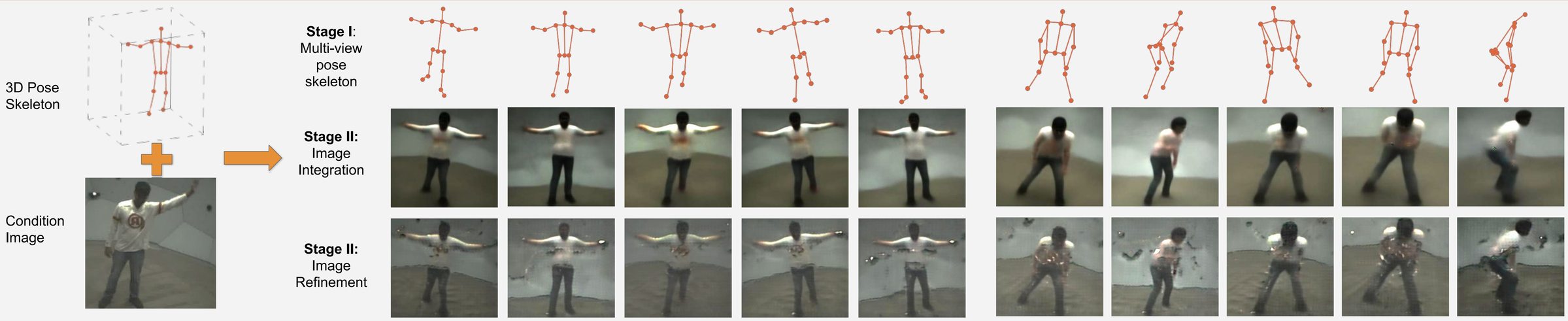
Stage II: Image integration: Pose skeleton + condition image (U-Net)

Stage III: Image refinement (DCGAN)

$$\mathcal{L}_{G1} = \|(G1(I_A, P_B) - I_B) \odot (1 + M_B)\|_1$$

$$\mathcal{L}_{G2} = \mathcal{L}_{adv}^{G} + \lambda\|(G2(I_A, \hat{I}_{B1}) - I_B) \odot (1 + M_B)\|_1$$

Experiments and results



Future work

- Some of details of condition image are lost in our final output, such as shoes and texture of the clothes. As we can see in the result, the text on the shirts becomes white spots in our final result, this is because our model doesn't focus on the different parts of body and compute the loss on the whole image.
- The current model only supports generation of a single image. To make this model more practical and powerful, we need to find the way to generate the image of crowds.