Pose Guided Person Image Generation for Novel 3D View Synthesis

Byeongjoo Ahn, Anqi Yang, Sihan You
Carnegie Mellon University
{bahn,anqiy1,sihany}@andrew.cmu.edu

1 Introduction

We consider the problem of novel 3D view synthesis of person image. The goal of this problem is, given a single view of a person in an arbitrary pose, to synthesize an image of a person after a specified transformation of viewpoint. It has a variety of practical applications in computer vision, graphics, and robotics.

Some recent works have demonstrated good performance on realistic image generation beneficial from advanced models like Variational Autoencoder (VAE) [4] and Generative Adversarial Networks (GANs) [1]. However, generating images conditioned by specific attributes such as its viewpoint is still a challenging problem. This is because we need to infer the appearances of unobserved parts of human while preserving the global shape of input images. Although the input and desired output views may have similar low-level image statistics, enforcing the correspondence directly is difficult. Previous novel view synthesis approach [13, 10, 9] learn the transformation between the input view and target view by relocating pixels, but those methods mainly synthesize rigid objects with simple textures. Zhao et al. [11] proposed novel view synthesis method for human, but their method is limited to a predefined set of view (i.e., front, side, and back) because of the difficulty of acquiring consistent multi-view clothing dataset. In summary, the challenges we need to address is as follows:

- 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.

In order to address the above challenges and generate human images in novel view while preserving the global shape, we will make use of pose information as explicit guidance to the viewpoint change. Given input images, we first extract 3D keypoint for pose estimation [2, 8], 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 [7]. 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. The output image will be able to have not only realistic-looking but also accurate pose in the desired target view by making use of geometric keypoint information. Also, while the training data with different viewpoint is limited in both views and numbers, the training data with different pose is much easier to obtain, which greatly adds the usefulness of pose guidance.

Our main goal in this project is to implement the pose guided person image generation method and integrate it with 3D pose estimation for novel 3d view synthesis. We will synthesize the novel 3D view firstly from multi-view input images (e.g., stereo), and extend it finally to a single-view input image.

1.1 Related Work

Image generation is a heated topic in recent years. There are many methods proposed to take advantage of GAN [1] and VAE [4] in order to generate the image without time-consuming sampling

10-701: Introduction to Machine Learning, Spring 2018, Carnegie Mellon University.

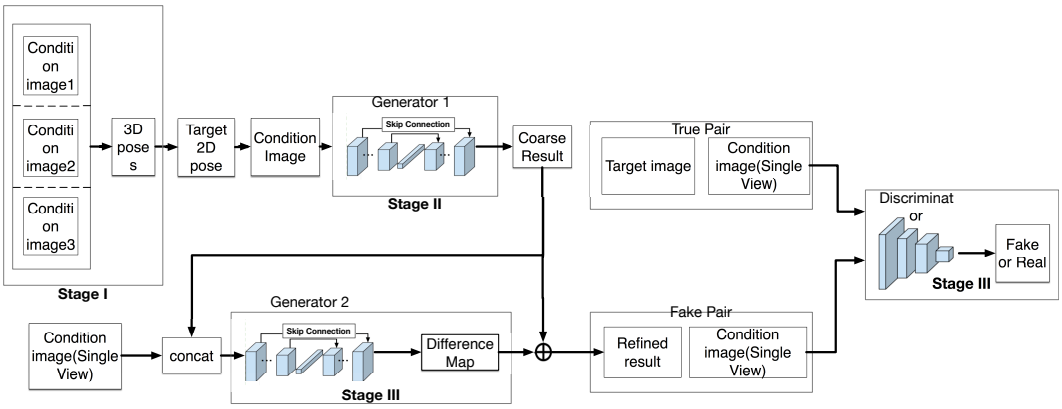

Figure 1: The overall framework of our multi-view image generation network.

process and also with details. Zhao et al. [11] proposed VariGAN to generate the multi-view image from a single view. They first generate a low resolution image in a different view by variational inference and then use GAN to generate the high resolution image by filling the details. Ma et al. [7] proposed a method to synthesize person images by reference image and a target pose. Ma et al. [7]'s model and architecture is similar to Zhao et al. [11], however they make use of pose information in a more explicit way, by generating the pose key points and heat map, compared with Zhao et al. [11].

2 Methodology

2.1 Our Method

Our task is to transfer the appearance of a person to any desired novel view from a set of multi-view images (e.g., stereo). As is shown in Fig. 1, we adopt a three-stage approach to address this task. At the first stage, we reconstruct a 3D body skeleton from the given multi-view images $\{I_A\}_{i=1}^{n}$, which is further used to be projected to the desired view as the target view condition P_B. At the second stage, we integrate the conditioning appearance image $\{I_A\}_0$ and the conditioning target pose P_B to generate a coarse result $\hat{I}_B$. A variant of VAE is adopted to generate the coarse image which captures the global structure of human body. At the final stage, we exploit a variant of conditional GAN to refine the coarse image $\hat{I}_B$ generated in the second stage into the final result I_B. In this stage, we focus on adding high-resolution details and correcting what is wrong or missing in the initial result $\hat{I}_B$.

2.2 Dataset

MVC The Multi-View Clothing (MVC) dataset [5] is one of the landmark datasets for multi-view clothing image retrieval and generation. It is consist of 36,323 clothing items, most of which has 4 views, $i.e.$ front side, left side, right side, and back side.

DeepFashion The DeepFashion (In-shop Clothes Retrieval Benchmark) dataset [6] contains 8,697 clothing items from four different views, $i.e.$ front side, left or right side, back and full body. All images are in high-resolution of 256×256. Note that images in DeepFashion have a high variance in scale, making it a challenging for multi-view image generation.

Market-1501 Market-1501 dataset [12] contains 32,668 images of 1,501 persons captured from six disjoint surveilance cameras, providing 6 different view simultaneously. The dataset also varies in pose, illumination, and background, which makes the image generation task more challenging.

Panoptic Studio The Panoptic Studio [3] is an integrated system that takes 480 synchronized VGA video stream and 30+ HD video stream as inputs, generating a large variety of views at the same time. We evaluate our multi-view image generation algorithm on the Panoptic Studio Pose sub-dataset, which contains 33 video action fragments. Each video action fragment is consist of around 6,000 frames, and is recorded from 480 different views.

3 Goal

3.1 Implementation of pose guided person image generation (75% Goal)

- Training on DeepFashion dataset [6].
- Training on Market-1501 dataset [12].

3.2 Novel 3D view synthesis from multi-view images (100% Goal)

- Implementation of 3D pose keypoint reconstruction from multi-view images [3].
- Implementation of 3D pose keypoint projection that is consistent with 2D keypoint used in pose guided person image generation.
- Training on CMU Panoptic dataset [2].
- Integration of pose estimation module and image generatation module.

3.3 Novel 3D view synthesis from a single-view image (125% Goal)

- Implementation of 3D pose estimation from single-view images [8].
- Implementation of 3D pose projection that is consistent with 2D keypoint used in pose guided person image generation.
- Integration of pose estimation module and image generation module.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[3] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 313–316. ACM, 2016.

[6] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017.

[8] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4): 44, 2017.

[9] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–711. IEEE, 2017.

[10] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.

[11] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017.

[12] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[13] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016.