Research on text-based 3D human body model generation

Submission for Assignment 5 CS7GV3

Long Pan
Student ID: 21332147
Trinity College Dublin

Abstract

The report describes an experiment that combines the Text2human and Eva3D methods to generate a 3D human model corresponding to a descriptive text input. The Text2human method generates a 2D human model based on text descriptions, while Eva3D is an unconditional 3D human generative model. The results show that the model can generate realistic human images, but there are limitations, such as potential errors in word embeddings and inaccuracies in SMPL estimation. Overall, the experiment demonstrates the potential of combining these methods for generating 3D human models from natural language descriptions.

Background

In recent years, image generation has made significant progress with the emergence of Generative Adversarial Networks (GANs) [Goodfellow et al. 2014]. This technology has made it possible to easily generate high-fidelity images of diverse faces using pretrained models like StyleGAN [Karras et al. 2020], which has led to various downstream applications, including facial attribute editing [Abdal et al. 2021; Jiang et al. 2021; Patashnik et al. 2021] and face stylization [Pinkney and Adler 2020; Song et al. 2021; Yang et al. 2022].

At the same time, inverse graphics has also witnessed rapid development. Inverse graphics studies the inverse-engineering of projection physics, with the aim of recovering the 3D world from 2D observations. This has numerous applications in VR/AR and VFX. Recently, 3D-aware generative models (Chan et al., 2021; Or-El et al., 2022; Chan et al., 2022; Deng et al., 2022) have demonstrated great potential in inverse graphics by generating 3D rigid objects, such as human and animal faces, from 2D image collections.

Motivated by these advancements, our research aims to generate 2D images of human bodies from the given human pose and user-specified texts describing the clothes shapes based on text2human paper, and subsequently utilize 2D human images to generate 3D models based on Eva3D paper. This presents a significant challenge, as human bodies are articulated objects with complex articulations and diverse appearances. Therefore, we seek to develop 3D human generative models that can synthesize animatable 3D humans with high-fidelity textures and vivid geometric details, which would have important applications in various domains such as virtual clothing try-on and 3D animation.

Implementation details

This research is mainly divided into two parts: Text2human and Eva3D.

Text2human

This part is mainly responsible for generating a 2D human body model corresponding to the descriptive text and the selected pose. The detailed steps are mainly divided into the following two stages：

Stage I translates the given human pose to the human parsing according to the text describing the clothes shapes. The text for clothes shapes is first transformed to one-hot shape attributes and embedded to a vector $fshape$. The shape vector $fshape$ is then fed into the pose-to-parsing module to spatially modulate the pose features.

Stage II generates the human image from the synthesized human parsing by sampling multi-level indices from our learned hierarchical texture-aware codebooks. To sample coarse-level indices, we employ a sampler with mixture-of-experts, where features are routed to different expert heads to predict the indices based on the required textures. At the fine level, I propose a feed-forward network to efficiently predict fine-level indices to refine the generated human image.

Eva3D

EVA3D is an unconditional high-quality 3D human generative model from sparse 2D human image collections only. To facilitate that, I imitate the author to first encode the skeleton with SMPL, and then propose compositional human NeRF representation to improve model efficiency.

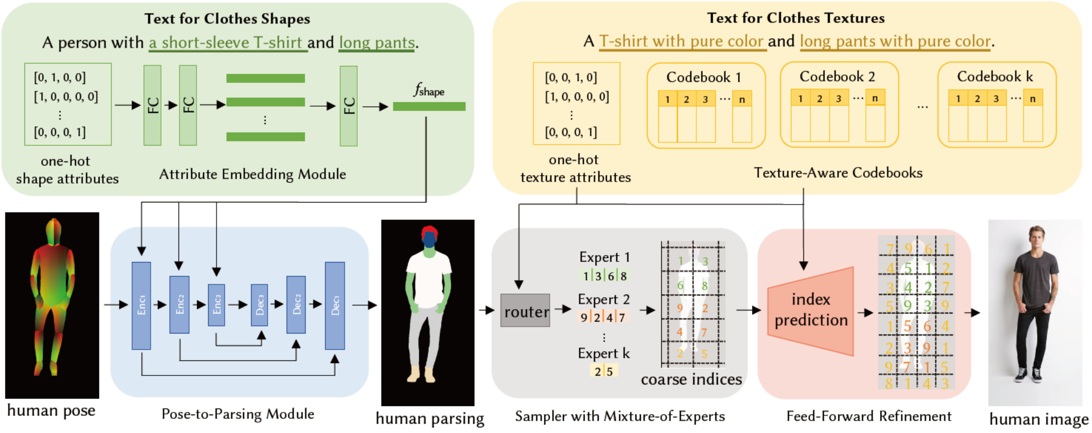

Figure 1: Overview of Text2human

SMPL is a linear algebra-based framework used to generate realistic human poses and shapes. It captures the body's skeletal structure and muscle form using anatomical knowledge and encodes this information into a mathematical model. NeRF is a method used to build realistic 3D scene models. It estimates the color and transparency of the scene at each pixel point based on ray tracing, predicts the color and transparency of each point in the scene generates a highly realistic 3D scene model.

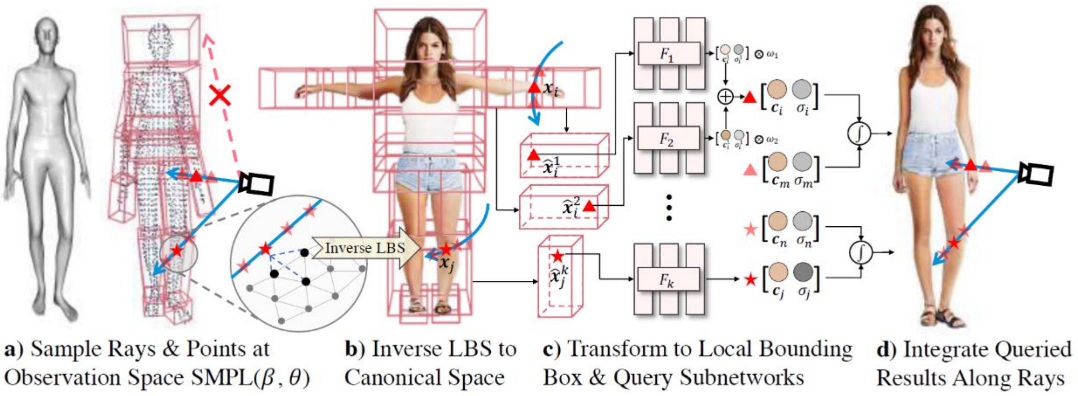
Inverse LBS
x_i
x_j
$\hat{x}_i^1$
$\hat{x}_i^2$
$\hat{x}_j^k$
F_1
F_2
F_k
ω_1
ω_2
c_i σ_i
c_m σ_m
c_n σ_n
c_j σ_j
a) Sample Rays & Points at Observation Space SMPL(β, θ)
b) Inverse LBS to Canonical Space
c) Transform to Local Bounding Box & Query Subnetworks
d) Integrate Queried Results Along Rays

Figure 2: Rendering Process of the Compositional Human NeRF Representation

Results

First, Text2Human generates human images based on text descriptions by using a framework that can generate photo-realistic human images from natural language descriptions. Users can upload a human pose map and then type a text describing the clothing shapes. A human parsing map will be generated accordingly. Then users provide another text describing the clothing textures, and Text2Human generates the corresponding final human image. The texture text used here is *Lady wears a short-sleeve T-shirt with pure color pattern, and a short and denim skirt*. As for the posture, I chose a normal standing posture with the hands hanging down naturally. Then we get a set of pictures like this:

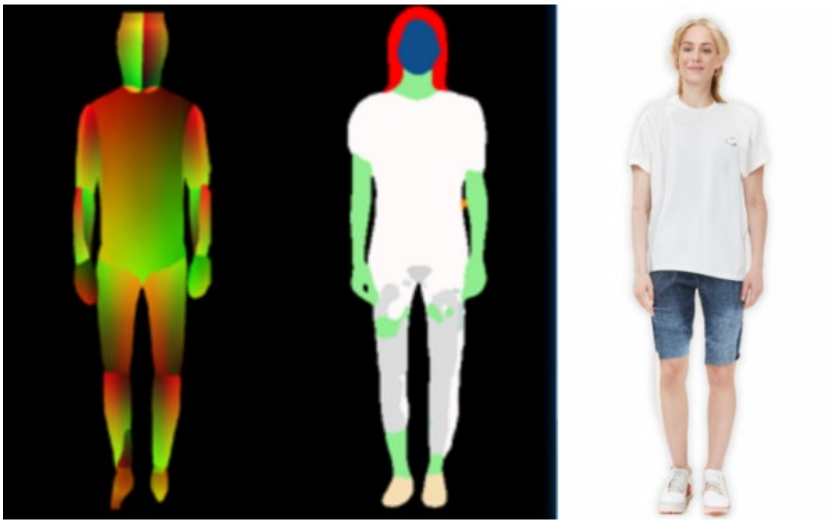

Figure 3: Result of text2human

According to the obtained pictures, we use SMPL to encode and mark each part, and then import the obtained results into Eva3D, EVA3D can support explicit control over human body shapes and poses, adaptively allocate computation resources, and achieve efficient rendering and high-resolution generation.

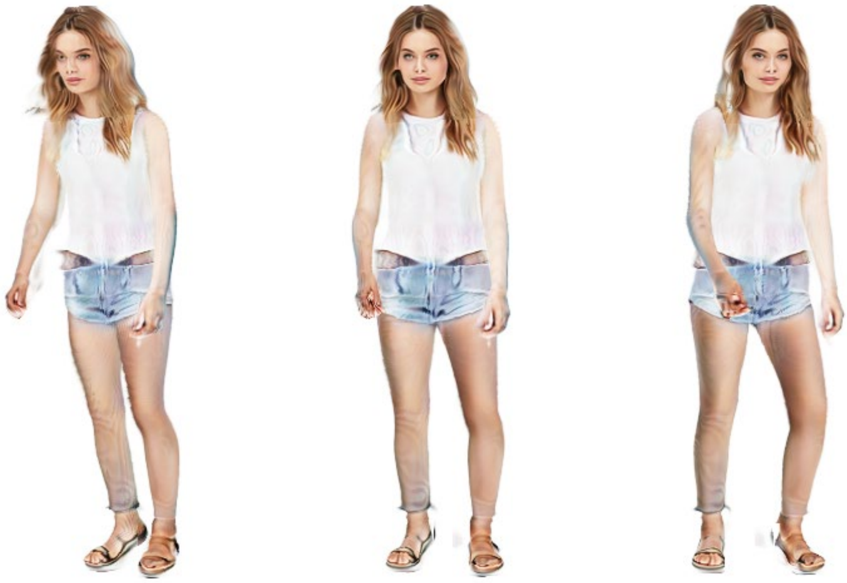

Figure 4: Results of Eva3D

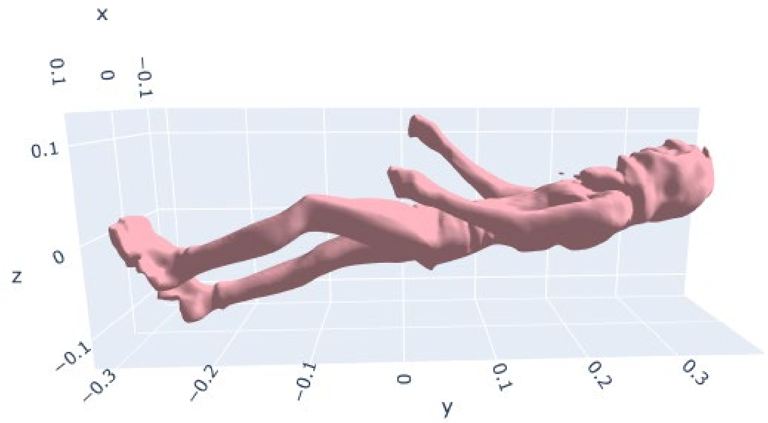

Figure 5: Result 3d model of Eva3D

Limitations

In general, this experiment combined the methods in Text2human and Eva3D to realize the human body 3D model corresponding to the descriptive text input. Through this experiment, although the model results can be roughly obtained, many limitations are still found in the experiment process, mainly as follows:

- Uncommon poses: The performance would degrade with human poses. For example, when the legs are crossed, artifacts will appear in the crossed area, and the generated model cannot handle this situation well.
- Potential error in word embeddings: Translating text descriptions to one-hot embeddings inevitably introduces errors. For example, for the length of sleeves, we only define four classes, i.e., sleeveless, short sleeves, medium sleeves, and long sleeves. If the user wants to generate a sweater with sleeves covering the elbow but not reaching the wrist, the synthesized human parsing cannot be perfectly aligned with the text inputs as the predefined texts cannot handle sleeves with arbitrary lengths.
- SMPL estimation: The estimation of SMPL parameters from 2D image collections is not accurate, which leads to a distribution shift from the real pose distribution and possibly compromises generation results.

Citation

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems 27 (2014).

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8110–8119.

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021. Styleflow: Attribute conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (TOG) 40, 3 (2021), 1–21.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. MaskGIT: Masked Generative Image Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2085–2094.

Justin NM Pinkney and Doron Adler. 2020. Resolution Dependent GAN Interpolation for Controllable Image Synthesis Between Domains. arXiv preprint arXiv:2010.05334 (2020).

Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, nd Tat-Jen Cham. 2021. AgileGAN: stylizing portraits by inversion-consistent ransfer learning. ACM Transactions on Graphics (TOG) 40, 4 (2021), 1–13.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Pastiche Master: xemplar-Based High-Resolution Portrait Style Transfer. In Proceedings of the EEE/CVF Conference on Computer Vision and Pattern Recognition. 1–10

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic mplicit generative adversarial networks for 3d-aware image synthesis. In Proceedings of the EEE/CVF conference on computer vision and pattern recognition, pp. 5799–5809, 2021.