Group 6 papers

Group members:

Long Pan, 21332147

Nitin Frederick, 21330025

Weiwei Wan, 22301337

Shuo Jia, 22301057

**Group Project Topic:** Text2Human

**Link:** <https://yumingj.github.io/projects/Text2Human.html>

**Description:**

Generating high-quality and diverse human images is an important yet challenging task in vision and graphics. The Text2Human framework in this paper generated more diverse and realistic human images compared to state-of-the-art methods. They synthesize full-body human images starting from a given human pose with two dedicated steps. 1) With some texts describing the shapes of clothes, the given human pose is first translated to a human parsing map. 2) The final human image is then generated by providing the system with more attributes about the textures of clothes. They build a hierarchical texture-aware codebook that stores multi-scale neural representations for each type of texture to model the diversity of clothing textures. Firstly, a diffusion-based transformer sampler with mixture-of-experts was used to sample indices from the coarsest level (structural representations of textures) of the codebook, which then is used to predict the indices of the codebook at finer levels (the details of textures). The predicted indices at different levels are translated to human images by the decoder learned accompanied with hierarchical codebooks.

[**Slides**](https://docs.google.com/presentation/d/1KW80IDErK_XA7LlCX2cbhK6p4aMIxdDOK4iAoSstpYE/edit?usp=sharing) **link**

Description Text2Human: Long Pan

Eg: importance and history: Weiwei Wan

Eg: industry or societal significance Nitin

5 min

10min

Paper1: Nitin 2.5 min

Paper2: Long Pan

Paper3: Weiwei

Paper4: jia shuo

1 min

Conclusion: jia shuo

# Related papers:

**Title：*Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing.***

**Link:** [**https://arxiv.org/pdf/2104.07021v3.pdf**](https://arxiv.org/pdf/2104.07021v3.pdf)

**Description:**

This paper proposed a flexible person generation framework, which supports 2D pose transfer, virtual try-on, and several fashion editing tasks. The key of this system is using a novel recurrent generation pipeline to sequentially put garments on a person and produce dressing effects not achievable by existing work. In addition to ensuring the detail and coherence of the garments, the system also ensures that the shape and texture of each garment are individually editable. At the end of the article, it also evaluates the effect of the system through experiments, and the results show that it is significantly better than other methods in output quality.

**Title:** ***HumanGAN: A Generative Model of Human Images***

**Link**:<https://arxiv.org/pdf/2103.06902.pdf>

**Description:**

This paper presented a VAE-based human image generation framework for full-body images of clothed humans called HumanGAN, which enables control of body pose, as well as independent control and sampling of appearance and clothing style on a body part level. On the body pose and cloth style generating part, Text2Human and HumanGAN could be compared and HumanGAN is a state-of-the-art method in this area. Text2Human framework allows for controllable human generation by giving texts describing the desired attributes and it’s easier to change cloth texture. However, HumanGAN didn’t do that. In the Text2Human paper, they also compared the two frameworks.

**Title: Vector Quantized Diffusion Model for Text-to-Image Synthesis**

**Link:** <https://arxiv.org/pdf/2111.14822.pdf>

**Description:**

This paper is based on a vector quantized variational autoencoder (VQ-VAE) whose latent space is modeled by a diffusion model. The VQ-Diffusion model produces significantly better text-to-image generation results when compared with conventional autoregressive (AR) models with similar numbers of parameters. Compared with previous GAN-based text-to-image methods, VQ-Diffusion can handle more complex scenes and improve the synthesized image quality by a large margin. With traditional AR methods, the text-to-image generation time increases linearly with the output image resolution and hence is quite time consuming even for normal size images. But a diffusion model allows us to achieve a better trade-off between quality and speed.

Text2Human adopts a diffusion model based transformer based on the performance improvements observed in this paper. Some of the advantages include much faster sampling speed because the indices are predicted in parallel and more coherent sampled images.

**Title: ClothFlow: A Flow-Based Model for Clothed Person Generation**

**Link:**<https://openaccess.thecvf.com/content_ICCV_2019/papers/Han_ClothFlow_A_Flow-Based_Model_for_Clothed_Person_Generation_ICCV_2019_paper.pdf>

**Description:**

A generative model based on appearance flow, ClothFlow, which can synthesize clothed people for image generation and virtual try-on of localization-guided people. ClothFlow effectively simulates geometric changes by estimating dense flow between source and target clothing regions, naturally transferring appearance, and synthesizing new images.

This paper achieves this through a three-stage framework: 1) Conditioned on the target pose, they first estimate a person's semantic layout to provide richer guidance for the generation process. 2) The cascaded traffic estimation network is built on the basis of two feature pyramid networks and accurately estimates the appearance matching between corresponding clothing regions. The resulting dense flow distorts the source image, interpreting the deformation flexibly. 3) Finally, the generative network takes the distorted clothing region as input and presents the target view. The method can better handle the problem of clothing deformation to better synthesize pictures of people with different clothes.