BAUHAUS-UNIVERSITÄT WEIMAR

EXPOSÉ

Multi-View 3D Avatar Style Transfer using Differential Rendering

Author:
Lucky CHANDRAUTAMA

Supervisor:
Prof. Dr.-Ing. habil. Volker RODEHORST

An exposé to a thesis submitted in fulfillment of the requirements for the degree of Master of Science

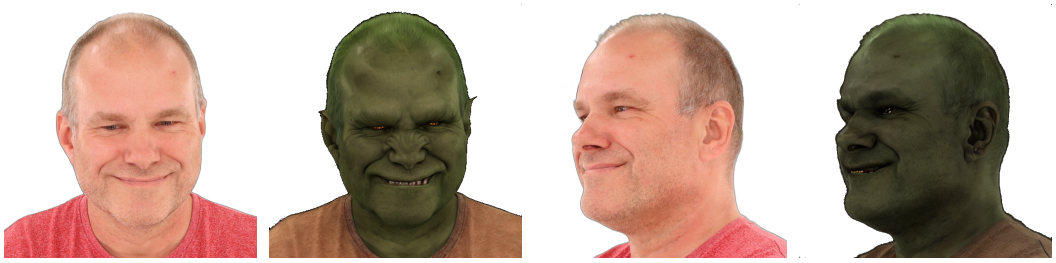
in the

Computer Science Department

September 13, 2024

Motivation

Diffusion models have pushed generative image modeling to an unprecedented level of photorealism and controllability. Adapting diffusion models to style transfer, i.e. the translation of a visual input into a target style domain with strong conditioning signals from the given input, is a widely studied topic with partly impressive results (Zhang et al., 2023; Brooks, Holynski, and Efros, 2022; Haque et al., 2023). Most strikingly, *Instruct-NeRF2NeRF* accomplishes 3D style transfer on human avatars in the setup of neural radiance fields (NeRF) (Haque et al., 2023). The continuous volume of a NeRF serves as the medium to consolidate the different stylization and blend multiview inconsistencies, which appear when diffusion model is used naively (as seen in Figure 1). It does, however, typically not flawlessly transform into a discrete textured mesh representation suitable for virtual reality applications. This work, thus, adapts the ideas from *Instruct-NeRF2NeRF* in a differential rendering setup to run style transfer on the textured mesh of a 3D avatar directly.



(A) Input at view #1. (B) Naive result #1. (C) Input at view #2. (D) Naive result #2.

FIGURE 1: Multi-view inconsistencies in naive diffusion-based style transfer. Given images of two different view points of the same subject, Figure 1a and 1c as inputs, Figure 1b and 1d are produced in sequence. The used text prompt is "Turn him into an orc." Note the differences in the stylization of the ear and teeth as well as in the wrinkles at the eye brows and nose.

Related Work

2D style transfer has experienced major progress in the last years. A groundbreaking approach was proposed by Gatys, Ecker, and Bethge (2016), who separate the style of an image from the content by means of the Gram matrix to enable the recombination with a different style. **Zhu.2017** proposed *CycleGAN*, an approach which leverages the generator-discriminator game in the setup of generative adversarial networks (GAN) yielding visually appealing results in a variety of contexts. Kolkin, Salavon, and Shakhnarovich (2019) introduced *STROSS*, which allows for user-specific control over points or regions to establish visual similarity between the style and the respective image. More recently, the success of the generative stable diffusion model (Rombach et al., 2021) shifted the focus towards diffusion-based style transfer. *Instruct-Pix2Pix* (Brooks, Holynski, and Efros, 2022) incorporates an image conditioning and *GPT-3* text embedding into the denoising pipeline of the diffusion model. The best paper awardee from ICCV'23, *ControlNet* (Zhang et al., 2023), adds more control to the diffusion pipeline allowing i.a. for edge, depth, segmentation, and human pose conditioning by making use of zero-convolutions. 2D style transfer has experienced major progress in the last years. A groundbreaking approach was proposed by Gatys, Ecker, and Bethge (2016), who separate the

style of an image from the content by means of the Gram matrix to enable the recombination with a different style. **Zhu.2017** proposed *CycleGAN*, an approach which leverages the generator-discriminator game in the setup of generative adversarial networks (GAN) yielding visually appealing results in a variety of contexts. Kolkin, Salavon, and Shakhnarovich (2019) introduced *STROSS*, which allows for user-specific control over points or regions to establish visual similarity between the style and the respective image. More recently, the success of the generative stable diffusion model (Rombach et al., 2021) shifted the focus towards diffusion-based style transfer. *Instruct-Pix2Pix* (Brooks, Holynski, and Efros, 2022) incorporates an image conditioning and *GPT-3* text embedding into the denoising pipeline of the diffusion model. The best paper awardee from ICCV'23, *ControlNet* (Zhang et al., 2023), adds more control to the diffusion pipeline allowing i.a. for edge, depth, segmentation, and human pose conditioning by making use of zero-convolutions.

Based on the progress of 2D style transfer, style transfer in the 3D domain is gradually yielding better results. Han et al. (2021) propose a pipeline to generate a stylized 3D human face model with an exaggerated geometry and texture transferred from a 2D cartoon image. The authors suggest a disentanglement between geometric domain and texture domain into two main stages. In the first stage, the method generates a coarse 3D face stylized geometry from a real facial photo of a person and a 2D caricature image as the style reference. Style transfer from the 2D caricature image to texture of the mesh is carried out using the *STROSS* approach by Kolkin, Salavon, and Shakhnarovich (2019). However, the authors acknowledge that the proposed pipeline is limited to only a single style image and a single real photo. Explorative studies indicate that the pipeline works only in very specific settings and does not allow for streamlining the procedure end-to-end.

Haque et al. (2023) propose *Instruct-NeRF2NeRF*, a novel technique of editing a NeRF scene. The main innovation of *Instruct-NeRF2NeRF* is an *iterative dataset update* (iterative DU): the NeRF scene is rendered from multiple viewpoints, which are stored as a dataset. This dataset is then fed into a modified *Instruct-Pix2Pix* to perform style transform on the singular views conditioned by a text prompt and the original image. *Instruct-Pix2Pix* translates every view in isolated fashion, i.e. without knowledge of the results of neighboring views typically resulting in multi-view inconsistencies. The stylized views become the new dataset, which is iteratively updated in the same manner. Upon convergence, the iterative DU has mediated between the views resulting in a view-consistent style transfer. A drawback of NeRF-based style transfer is that the continuous volume needs to be transformed into a textured mesh, which regularly reduces the visual quality of the output.

Recently, *3DAvatarGAN* was proposed by Abdal et al. (2023), which fine-tunes an existing 3D-GAN approach on 2D dataset in a way such that the geometry and texture quality is preserved while adapting to the styles that are defined in the 2D dataset. Due to unpublished code, details of the implementation remain inaccessible.

Objectives

The goal of the thesis is to perform 3D style transfer on a human avatar in a way that the avatar is usable in a virtual reality setup. This encompasses the 3D reconstruction of a human into a textured mesh representation and the stylization of this mesh, both in terms of geometry and texture. The 3D reconstruction is accomplished in the

ecosystem of the photodome of the computer vision chair and the stylization process is guided by text prompting.

Proposed Steps

In order to reach the described objective the following steps are proposed:

1. Implement a processing pipeline for automatically performing a photogrammetric 3D reconstruction on the images captured by the photodome.
2. Design a differential rendering pipeline for joint *textural* and *geometric* style transfer on the captured avatar.
3. Resolve multi-view inconsistencies by adapting the *iterative DU* proposed by Instruct-NeRF2NeRF, which relies on Instruct-Pix2Pix for view stylization.
4. Implement the designed stylization pipeline and deploy it on the photodome ecosystem.
5. Perform qualitative analysis in a bounded user study of at least three participants.

Expected Results

It is assumed, that the qualitative results in terms of texture style transfer closely resemble the achievements by *Instruct-NeRF2NeRF*. Due to the different modalities (mesh vs. volume), there is major uncertainty, how accurately the geometric style transfer can be accomplished in the designed differential rendering setup. This especially refers to considerable geometric modifications such as adding a hat to the avatar.

Bibliography

Abdal, Rameen et al. (2023). "3DAvatarGAN: Bridging Domains for Personalized Editable Avatars". In: URL: https://arxiv.org/pdf/2301.02700.

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros (2022). "InstructPix2Pix: Learning to Follow Image Editing Instructions". In: URL: http://arxiv.org/pdf/2211.09800.pdf.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge (2016). "Image Style Transfer Using Convolutional Neural Networks". In: *29th IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, pp. 2414–2423. ISBN: 978-1-4673-8851-1. URL: https://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html.

Han, Fangzhou et al. (2021). "Exemplar-Based 3D Portrait Stylization". In: *IEEE Transactions on Visualization and Computer Graphics* PP.2, pp. 1371–1383. ISSN: 1941-0506. DOI: 10.1109/TVCG.2021.3114308.

Haque, Ayaan et al. (2023). "Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions". In: URL: https://arxiv.org/pdf/2303.12789.pdf.

Kolkin, Nicholas, Jason Salavon, and Greg Shakhnarovich (2019). "Style Transfer by Relaxed Optimal Transport and Self-Similarity". In: URL: https://arxiv.org/pdf/1904.12785.pdf.

Rombach, Robin et al. (2021). "High-Resolution Image Synthesis with Latent Diffusion Models". In: URL: https://arxiv.org/pdf/2112.10752.pdf.

Zhang, Hao et al. (2023). *Text-Guided Generation and Editing of Compositional 3D Avatars*. URL: http://arxiv.org/pdf/2309.07125.