HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images

What is the problem?

Existing methods for avatar reconstruction often rely on expensive setups like multi-view RGB video or large datasets, which are impractical for casual users. The reconstruction of human avatars from such data sources is challenging because of limited data amount and dynamic articulated poses. The paper explores the possibility of creating accurate human avatars using only a small number of unconstrained images (as few as two), taken from various angles and poses. The reconstruction needs to support not only static models but also animated avatars with free-pose articulation and realistic rendering. The paper introduces a solution using a drivable tetrahedral representation, which combines Deep Marching Tetrahedra (DMTet) and a skinning mechanism based on SMPLX. This allows for the adaptation of mesh topologies to handle dynamic human poses with fewer images. they employ a two-phase optimization method—few-shot reference for aligning avatar identity and few-shot guidance using Score Distillation Sampling (SDS) to generate plausible appearances for unseen regions, ensuring realistic rendering and animation even with sparse data.

What has been done earlier?

Prior to the solution proposed in the HAVE-FUN paper, human avatar reconstruction relied heavily on data-intensive methods. Techniques typically used multi-view RGB videos, textured scan videos, or large multi-view image sets, all of which required costly data acquisition setups. Some efforts utilized monocular RGB video for dynamic human reconstruction, while others adopted dynamic neural radiance fields (NeRF) to model human articulation and poses. These methods generally either treated the human body as a static entity or required large-scale data, making them unsuitable for handling dynamic poses and few-shot scenarios.

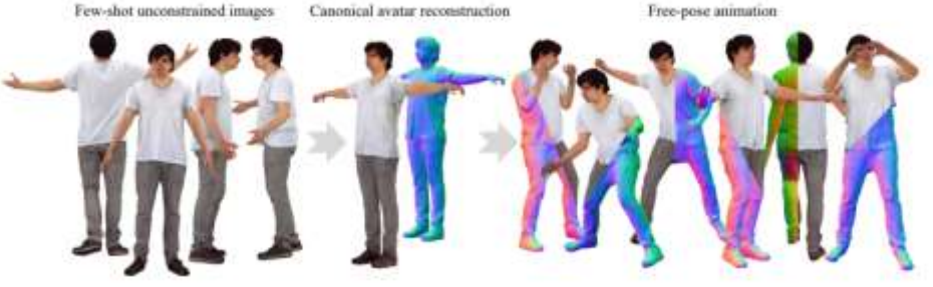
Few-shot unconstrained images
Canonical avatar reconstruction
Free-pose animation

Figure 1. Given a few images with various viewpoints and articulated poses, our approach can reconstruct an animatable human avatar.

Pulkit Sinha, B321059

What are the remaining challenges? What novel solution proposed by the authors to solve the problem?

Although the proposed method introduces a drivable tetrahedral representation, accurately capturing highly intricate or extreme dynamic poses, such as detailed facial expressions or nuanced hand movements, can still be challenging. While Score Distillation Sampling (SDS) helps generate plausible textures for unseen regions, maintaining consistent texture quality across diverse lighting conditions, clothing types, and fine details like hair remains difficult. In addition to these, the current benchmarks like FS-XHumans and FS-DART provide controlled datasets, but real-world images often involve occlusions, poor lighting, and cluttered backgrounds, which may degrade the performance of the avatar reconstruction. When generating avatars from images with varying expressions or body poses, it can be hard to consistently capture and replicate the exact identity and emotional nuances of the individual. Variability in facial expressions across few-shot images may lead to less precise facial geometry reconstruction, as seen in certain multi-shot cases where multiple expressions introduce reconstruction noise.

To tackle these problems, the authors proposed certain novel solutions like refining the blendshape models, such as SMPLX, which are currently limited in accurately representing highly detailed expressions and extreme body articulations. They suggest improving (Score Distillation Sampling) SDS-based optimization by incorporating more advanced neural networks that can infer richer texture details and geometries from extremely sparse input data. The authors plan to enhance the drivable tetrahedral representation by incorporating more flexible, adaptive grids that can better capture human body nuances. They also suggest utilizing synthetic data generation techniques to augment the few-shot datasets by training on a wider variety of synthetic poses, textures, and lighting conditions.

Pulkit Sinha, B321059