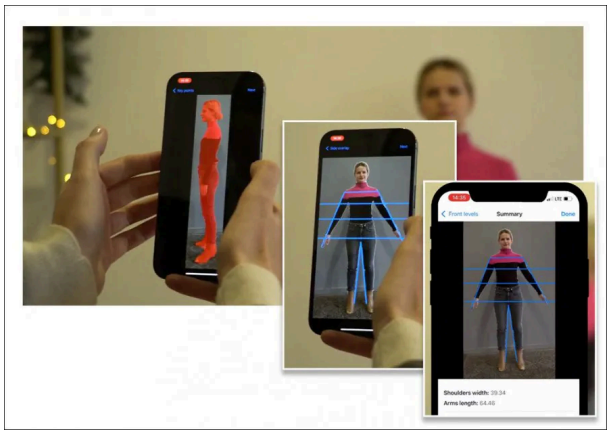
Creating Full Body Deepfakes by Combining Multiple NeRFs

Originally published at https://www.unite.ai/creating-full-body-deepfakes-by-combining-multiple-nerfs/

The image synthesis research sector is thickly littered with new proposals for systems capable of creating full-body video and pictures of young people – mainly young women – in various types of attire. Mostly the generated images are static; occasionally, the representations even move, though not usually very well.

The pace of this particular research strand is glacial in comparison to the current dizzying level of progress in related fields such as latent diffusion models; yet the research groups, the majority in Asia, continue to plug away relentlessly at the problem.



One of dozens, if not hundreds of proposed or semi-launched 'virtual try-on' systems from the last 10-15 years, where bodies are evaluated through machine learning-based object recognition and adapted to the proposed items of clothing. Source: https://www.youtube.com/watch?v=2ZXrgGyhbak

The goal is to create new systems to enable 'virtual try-ons' for the fashion and clothing market – systems that are can adapt both to the customer and to the specific product that's currently available or about to be released, without the clunkiness of real-time superimposition of clothing, or the need to ask customers to send slightly NSFW pictures for ML-based rendering pipelines.

None of the popular synthesis architectures seem easily adaptable to this task: the latent space of Generative Adversarial Networks (GANs) is ill-suited to producing convincing temporal motion (or even for editing in general); though well-capable of generating realistic human movement, Neural Radiance Fields (NeRF) are usually naturally resistant to the kind of editing that would be necessary to 'swap out' people or clothing at will; autoencoders would require burdensome person/clothing-specific training; and latent diffusion models, like GANs, have zero native temporal mechanisms, for video generation.

EVA3D

Nonetheless, the papers and proposals continue. The latest is of unusual interest in an otherwise undistinguished and exclusively business-oriented line of research.

EVA3D, from Singapore's Nanyang Technological University, is the first indication of an approach that has been a long time coming – the use of *multiple* Neural Radiance Field networks, each of which is devoted to a separate part of the body, and which are then composed into an assembled and cohesive visualization.

CLICK TO VIEW ANIMATED GIF

A mobile young woman composited from multiple NeRF networks, for EVA3D.
Source:
https://hongfz16.github.io/projects/EVA3D.html

The results, in terms of movement, are…okay. Though EVA3D's visualization are not out of the uncanny valley, they can at least see the off-ramp from where they're standing.

CLICK TO VIEW ANIMATED GIF

What makes EVA3D outstanding is that the researchers behind it, almost uniquely in the sector of full-body image synthesis, have realized that a single network (GAN, NeRF or otherwise) is not going to be able to handle editable and flexible human full-body generation for some years – partly because of the pace of research, and partly because of hardware and other logistical limitations.

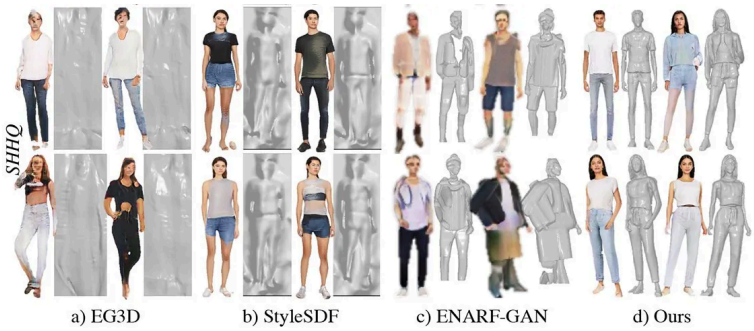
Therefore, the Nanyang team have subdivided the task across 16 networks and multiple technologies – an approach already adopted for neural rendering of urban environments in Block-NeRF and CityNeRF, and which seems likely to become an increasingly interesting and potentially fruitful half-way measure to achieve full-body deepfakes in the next five years, pending new conceptual or hardware developments.

Not all the challenges present in creating this kind of 'virtual try-on' are technical or logistical, and the paper outlines some of the data issues, particularly in regard to unsupervised learning:

> *'[Fashion] datasets mostly have very limited human poses (most are similar standing poses), and highly imbalanced viewing angles (most are front views). This imbalanced 2D data distribution could hinder unsupervised learning of 3D GANs, leading to difficulties in novel view/ pose synthesis. Therefore, a proper training strategy is in need to alleviate the issue.'*

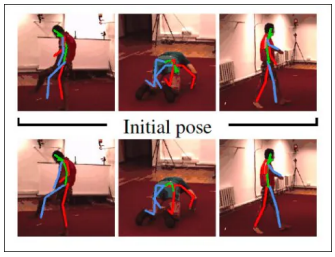
The EVA3D workflow segments the human body into 16 distinct parts, each of which is generated through its own NeRF network. Obviously, this creates enough 'unfrozen' sections to be able to galvanize the figure through motion capture or other types of motion data. Besides this advantage, however, it also allows the system to assign maximum resources to the parts of the body that 'sell' the overall impression.

For instance, human feet have a very limited range of articulation, whilst the authenticity of the face and head, besides the quality of the entire body motion in general, is likely to be the focal token of authenticity for the rendering.



A qualitative comparison between EVA3D and prior methods. The authors claim SOTA results in this respect.

The approach differs radically from the NeRF-centric project to which it is conceptually related – 2021's A-NeRF, from the University of British Columbia and Reality Labs Research, which sought to add an internal controlling skeleton to an otherwise conventionally 'one piece' NeRF representation, making it more difficult to allocate processing resources to different parts of the body on the basis of need.



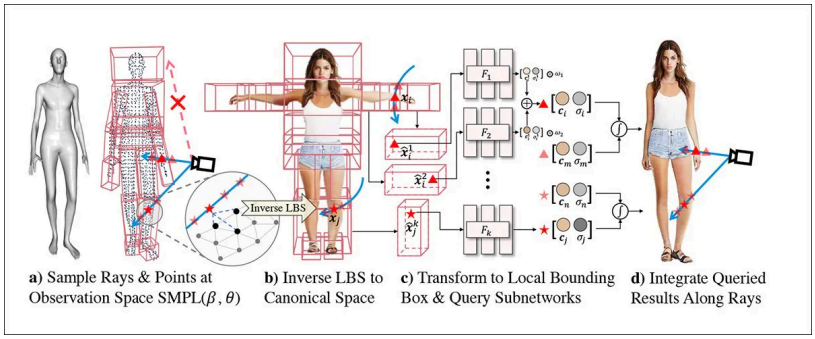
Prior motions – A-NeRF outfits a 'baked' NeRF with the same kind of ductile and articulated central rigging that the VFX industry has long been using to animate CGI characters. Source: https://lemonatsu.github.io/anerf/

In common with most similar human-centric projects that seek to leverage the latent space of the various popular approaches, EVA3D uses a Skinned Multi-Person Linear Model (SMPL), a 'traditional' CGI-based method for adding instrumentality to the general abstraction of current synthesis methods. Earlier this year, another paper, this time from Zhejiang University in Hangzhou, and the School of Creative Media at the City University of Hong Kong, used such methods to perform neural body reshaping.

EVA3D's qualitative results on DeepFashion.

Method

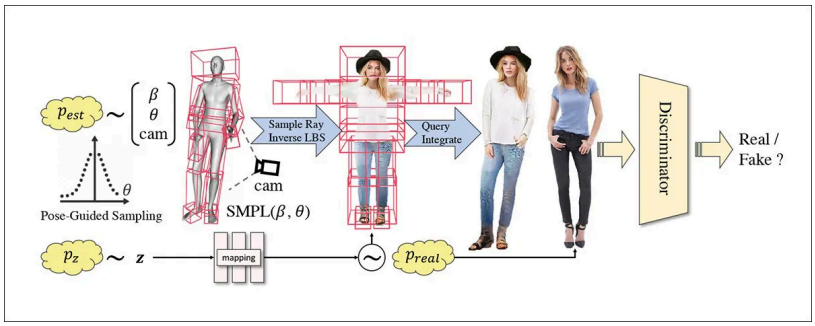
The SMPL model used in the process is tuned to the human 'prior' – the person who is, essentially, being voluntary deepfaked by EVA3D, and its skinning weights negotiate the differences between the canonical space (i.e. the 'at rest', or 'neutral' pose of an SMPL model) and the way that the final appearance is rendered.



The conceptual workflow for EVA3D. Source: https://arxiv.org/pdf/2210.04888.pdf

As seen in the illustration above, the bounding boxes of SMPL are used as the boundary definitions for the 16 networks that will eventually compose the body. Inverse Linear Blend Skinning (LBS) algorithm of SMPL is then used to transfer visible sampled rays to the canonical (passive pose) space. Then the 16 sub-networks are queried, based on these configurations, and ultimately conformed into a final render.

The entire NeRF composite is then used to construct a 3D human GAN framework.



The renderings of the second-stage GAN framework will ultimately be trained against genuine 2D image collections of humans/fashion.

Each sub-network representing part of the human body is composed of stacked Multi-Layer Perceptrons (MLPs) with SIREN (Sinusoidal Representation Networks) activation. Though SIREN solves a lot of problems in a workflow like this, and in similar projects, it tends to overfit rather than generalize, and the researchers suggest that alternative libraries could be used in the future (see end of article).

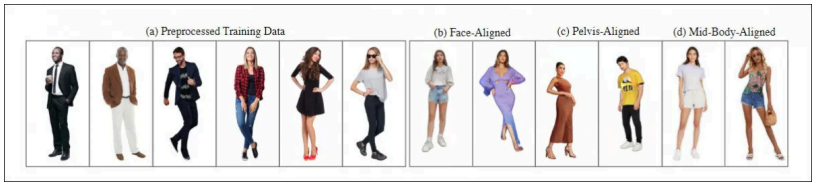
Data, Training, and Tests

EVA3D is faced with unusual data problems, due to the limitations and templated style of the poses that are available in fashion-based datasets, which tend to lack alternative or novel views, and are, perhaps intentionally, repetitive, in order to focus attention on the clothes rather than the human wearing them.

Due to this imbalanced pose distribution, EVA3D uses human priors (see above) based off the SMPL template geometry, and then predicts a Signed Distance Field (SDF) offset of this pose, rather than a straightforward target pose.

For the supporting experiments, the researchers utilized four datasets: DeepFashion; SHHQ; UBCFashion; and the AIST Dance Video Database (AIST Dance DB).

The latter two contain more varied poses than the first two, but represent the same individuals repetitively, which cancels out this otherwise useful diversity; in short, the data is more than challenging, given the task.



Examples from SSHQ. Source: https://arxiv.org/pdf/2204.11823.pdf

The baselines used were ENARF-GAN, the first project to render NeRF visuals from 2D image datasets; Stanford and NVIDIA's EG3D; and StyleSDF, a collaboration between the University of Washington, Adobe Research, and Stanford University – all methods requiring super-resolution libraries in order to scale up from native to high resolution.

Metrics adopted were the controversial Frechet Inception Distance (FID) and Kernel Inception Distance (KID), along with Percentage of Correct Keypoints (PCKh@0.5).

In quantitative evaluations, EVA3D led on all metrics in four datasets:

Methods, Resolution	*DeepFashion* FID↓	KID↓	PCK↑	Depth↓	*SHHQ* FID↓	KID↓	PCK↑	Depth↓
EG3D, 512^2	26.38	0.014	-	0.0779	32.96	0.033	-	0.0296
StyleSDF, 512^2	92.40	0.136	-	0.0359	14.12	0.010	-	0.0300
ENARF-GAN*, 128^2	77.03	0.114	43.74	0.1151	80.54	0.102	40.17	0.1241
Ours, 512^2	**15.91**	**0.011**	**87.50**	**0.0272**	**11.99**	**0.009**	**88.95**	**0.0177**
Methods, Resolution	*UBCFashion* FID↓	KID↓	PCK↑	Depth↓	*AIST* FID↓	KID↓	PCK↑	Depth↓
EG3D, 512^2	23.95	**0.009**	-	0.1163	34.76	0.022	-	0.1165
StyleSDF, 512^2	18.52	0.011	-	0.0311	199.5	0.225	-	0.0236
ENARF-GAN*, 128^2	-	-	-	-	73.07	0.075	42.85	0.1128
Ours, 512^2	**12.61**	0.010	**99.17**	**0.0090**	**19.40**	**0.010**	**83.15**	**0.0126**

Quantitative results.

The researchers note that EVA3D achieves the lowest error rate for geometry rendering, a critical factor in a project of this type. They also observe that their system can control generated pose and achieve higher PCKh@0.5 scores, in contrast to EG3D, the only competing method that scored higher, in one category.

EVA3D operates natively at the by-now standard 512x512px resolution, though it could be easily and effectively upscaled into HD resolution by piling on upscale layers, as Google has recently done with its 1024 resolution text-to-video offering Imagen Video.

The method is not without limits. The paper notes that the SIREN activation can cause circular artifacts, which could be remedied in future versions by use of an alternative base representation, such as EG3D, in combination with a 2D decoder. Additionally, it is difficult to fit SMPL accurately to the fashion data sources.

Finally, the system cannot easily accommodate larger and more fluid items of clothing, such as large dresses; garments of this type exhibit the same kind of fluid dynamics that make the creation of neurally-rendered hair such a challenge. Presumably, an apposite solution could help to address both issues.

Demo Video for EVA3D: Compositional 3D Human Generation from 2D Image Collections

First published 12th October 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More