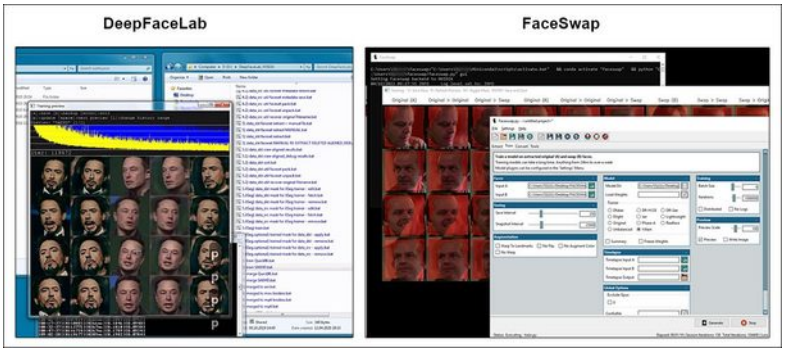
NeRF: An Eventual Successor for Deepfakes?

Originally published June 27, 2022 at https://metaphysic.ai/nerf-successor-deepfakes/

In February of 2022, a study from Lancaster University in the UK found not only that most people cannot distinguish deepfake faces from real faces, but also that we tend to trust synthesized faces more.

While this is excellent media fodder, and while there's little doubt that AI-generated humans have increased in quality since the advent of deepfakes in 2017, it's worth considering the extent to which our discriminatory powers and standards for authenticity are likely to rise in line with improvements in deepfake generation; that there's still a long way to go before machine learning can reproduce humans with complete fidelity; and that the technologies that are currently dazzling us are not necessarily the ones that will achieve the next evolutionary leap in this sphere.

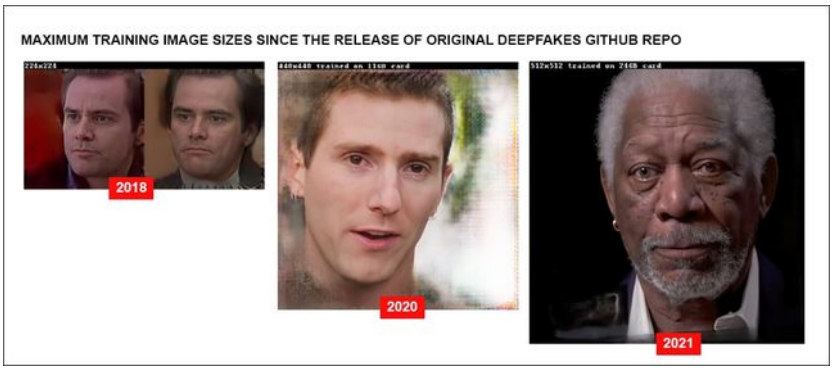
Most of the attention-garnering headlines regarding deepfakes are referring to the use of two open source packages that became popular after deepfakes entered the public arena in 2017: DeepFaceLab (DFL) and FaceSwap.



DeepFaceLab (lower right) and FaceSwap (left), both derived from the original 2017 deepfakes source code, have come to dominate deepfake production workflows.

Though each of these packages boasts a wide user-base and an active community of developers, neither project has significantly departed from the now five year-old GitHub code, which was almost immediately abandoned by its enigmatic creator when controversy began to build over the startling new technology.

The DFL and FaceSwap developers have not been idle, for sure: it's now possible to use larger input images for training deepfake models (see image below), though this requires more expensive video cards; masking out occlusions (such as hands in front of faces) in deepfakes has been semi-automated by innovations such as XSEG training; and the often ungainly command line environment of DFL has been integrated into the more user-friendly ministrations of the Machine Editor graphical host environment.



Between 2018 and 2021, the maximum size of input training data has gone up for DeepFaceLab, though it's not a 'magic bullet' in terms of obtaining more realistic results. Source: https://github.com/iperov/DeepFaceLab

But in truth, the improvements in deepfake quality that so many media outlets have noted over the past three years have been primarily due to end-users gaining time-consuming and hard-won experience in data gathering; in discovering the best ways to train models (where it can take several weeks to run even a single experiment); and generally learning to fully exploit and extend the outermost limits of the original 2017 code.

The Challenge of 'Scaling Up'

Among a myriad of new innovations in the VFX and ML research communities, some are attempting to break through the 'hard limits' of the popular deepfake packages by extending the architecture so that a machine learning model can train on images as large as 1024×1024 pixels – double the current practical confines of DeepFaceLab or FaceSwap, and nearer to the kind of resolution that would be useful in film and television production:

CLICK TO PLAY VIDEO ON SITE

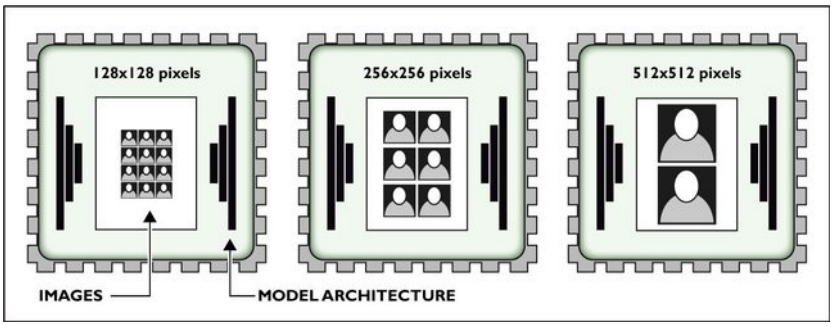
We'll take a deeper look at this proprietary technique when we chat with its creator, in a later article on autoencoder-based deepfakes.

However, results as impressive as these are difficult to obtain with standard open source deepfakes software; require expensive and powerful hardware; and usually entail very long training times to obtain very limited sequences.

Machine learning models are trained and developed within the capacity of the VRAM and tensor cores on a single video card — a prospect that becomes more and more challenging in the age of hyperscale datasets, and which presents some specific obstacles to improving deepfake quality.

Approaches that shunt training cycles to the CPU, or divide the workload up among multiple GPUs via Data Parallelism or Model Parallelism techniques (we'll examine these more closely in a later article) are still in the early stages. For the near future, a single-GPU training setup remains the most common scenario.

Consequently, improvements in deepfake techniques must work around this architectural bottleneck. For instance, in order to train a deepfake model on 5-10,000 source images, those images must be passed through the 'live' area of the training architecture in limited batches. The larger the input image size (512×512 pixels is currently considered quite large), the smaller the batch must be, even on the most expensive and best-specced cards available.



Larger training images cut down batch sizes during model training. Some space on the GPU must also be sacrificed to accommodate the software architecture.

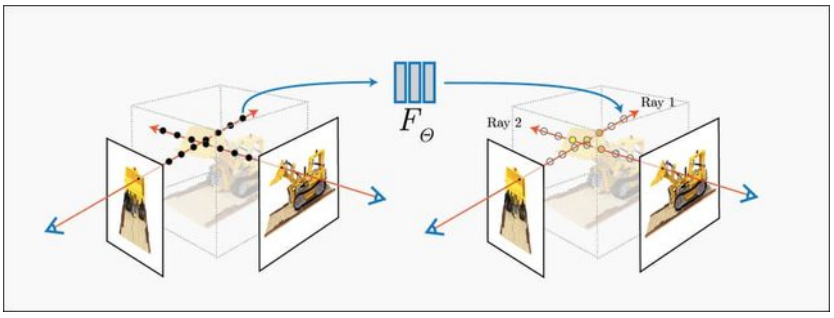
Small batches can hinder the model from 'generalizing' at an optimal level, risking overfitting or under-fitting. In these cases, the final model only operates well on the original data, or else fails to distil the essential features of the data — and in either eventuality, does not obtain a useful and flexible solution.

Neural Radiance Fields (NeRF)

Though popular branches of the 2017 deepfakes code (such as DFL and FaceSwap) may eventually benefit from better multi-GPU deployments, innovations in GPU architecture, and the restoration of chip production after years of GPU famine, their architecture is quite brittle, and not necessarily amenable to some of the most interesting new developments in human image synthesis.

It's therefore possible that new and less constrained methods will eventually provide the next evolutionary leaps in deepfake technologies.

One such approach is Neural Radiance Fields (NeRF), which emerged in 2020 as a method of recreating objects and environments by stitching together multiple viewpoint photos inside a neural network.



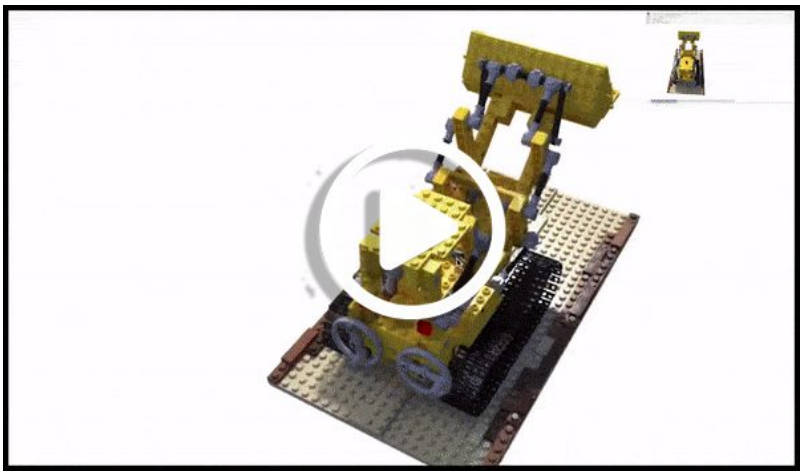
NeRF's photogrammetry: virtual light rays are calculated along the estimated geometry and RGB values of a scene. Source: https://www.matthewtancik.com/nerf

Given a limited number of viewpoints, NeRF calculates the 'missing views' by recognizing shapes, textures, transparency, and lighting values, and estimating and synthesizing the views that aren't present in the source data.

Here we see the origin of the term 'radiance' in Neural Radiance Fields, as the training data's various viewpoints 'radiate out' from the object being assimilated into the neural network. Source: https://www.youtube.com/watch?v=EpH175PY1A0

The NeRF process maps the entire volume of the target scene, which is now similar to a solid Lego construction made from thousands of bricks.

In terms of ‘classic’ CGI techniques, these blocks are equivalent to traditional CGI voxels, in that they are mapped to three-dimensional coordinates in a restricted volume of space, and can then be explored in a 2D volume rendering.



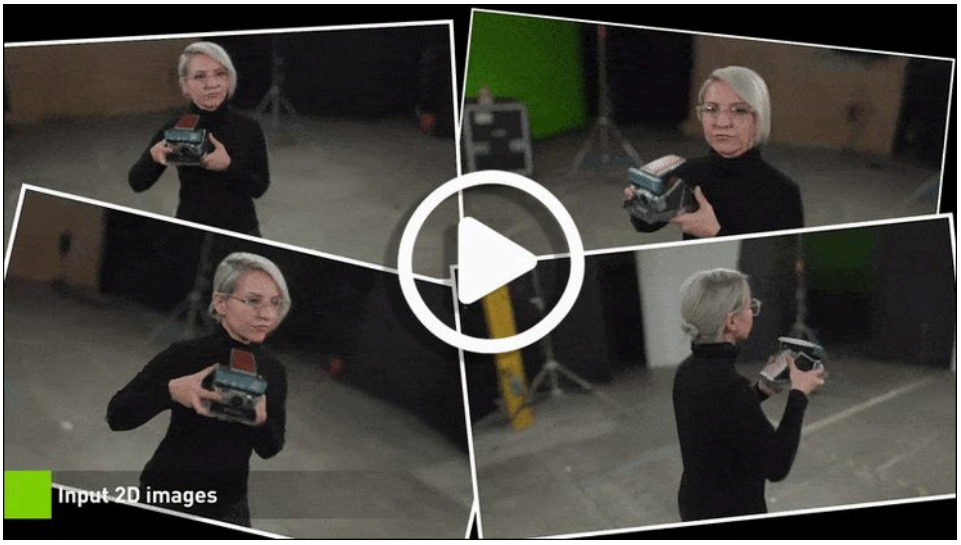
By 'baking' information in a NeRF, this 2021 project from Google Research enabled real-time rendering of a NeRF environment on a laptop, at frame rates above 30fps. Note the controlling cursor. Source: https://phog.github.io/snerg/

Instant NeRF

Together with storage requirements, the most significant obstacle to the deployment and development of NeRF-based VFX pipelines has been the extensive training times required for earlier implementations.

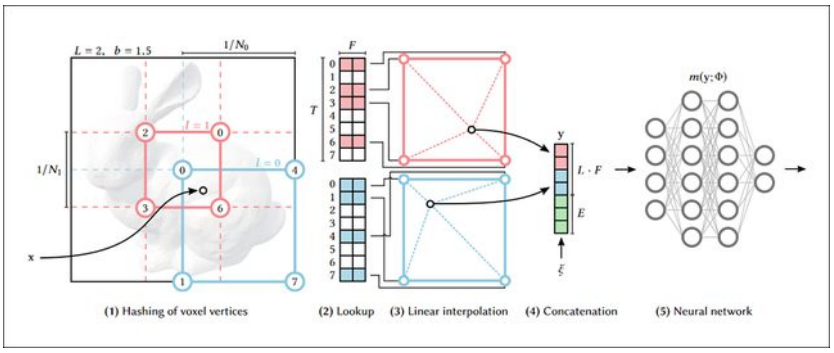
For instance, in the original 2020 paper, the training of a single frame took 30 seconds, while an entire scene took 24-48 hours to train to 300,000 iterations over the 32gb of VRAM in a NVIDIA V100 GPU.

Though later innovations, such as Mip-NeRF, Plenoctrees, KiloNeRF, DietNeRF and Plenoxels would eventually offer notable reductions in training times, the announcement of NVIDIA’s own Instant Neural Graphics Primitives (NGP) framework in January of 2022 represents a potential evolutionary leap for NeRF-based image synthesis, offering usable training times as short as *five seconds*. Instant NeRF is able to infer an extraordinary breadth of simulated frames from only a handful of ‘real’ photos:



NVIDIA's Instant NeRF derives a complex and explorable neural scene from just four 'real' photos, complete with realistic depth of field. Source: https://www.youtube.com/watch?v=DJ2hcC1orc4

This extraordinary new economy is achieved in part by *multiresolution hash encoding*, which stores *representative* markers for the high volume of data in the NeRF’s neural network, helping the system to discard any information that does not directly result in image content.



NVIDIA's Neural Graphics Primitive workflow uses hashes as representative placeholders for voxel vertices, so that it's no longer necessary to traverse or consider the entirety of the voxel coordinates (including non-contributing coordinates) in a local neural network. Source: https://nvlabs.github.io/instant-ngp/assets/mueller2022instant.pdf

Thus the many thousands of non-visible 'Lego bricks' which intervene between the viewer and the depicted objects no longer need to be considered, or initialized, at training time.

A neural representation of a factory scene under NVIDIA's NGP. Source: https://github.com/NVlabs/instant-ngp

The best of the prior optimization methods, such as Plenoctrees, required the development of a slower, feature-rich NeRF (including redundant information that would later be automatically excised during training) in order to produce surfaces to optimize.

By contrast, NVIDIA's method pre-calculates the network's trainable feature vectors in lightweight hash tables. These hash tables are generated at multiple resolutions (hence 'parametric'), which can be explored according to the requirements of the viewpoint at any given time.

An additional benefit of this sparse approach to hashing is that it makes caching far more flexible and capable, leading to a more responsive interface experience.

The optimized NGP approach improves caching, explorability and latency of NeRF environments.

Human Representations in NeRF

NeRF was quickly adapted to facilitate the depiction of motion in a neural environment. Since it proved possible to capture the movement of any object, a large portion of the research community has shifted emphasis to the re-creation of human appearance and movement.

Above is a clip of test footage from ST-NeRF, a 2021 ShanghaiTech University implementation of NeRF that allows individual temporal captures of actors and performers to be arbitrarily resized and edited in neural environments (for high quality footage from the project, see the accompanying video).

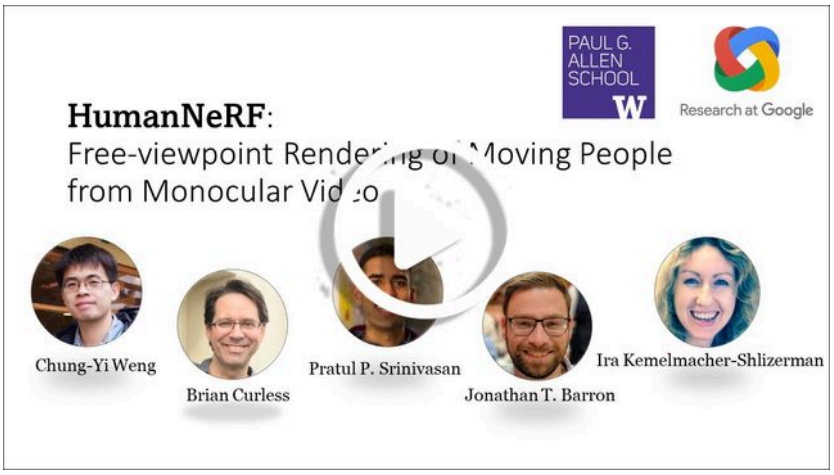
CLICK TO PLAY VIDEO ON SITE

In the above clip from the ST-NeRF project, we see that two separate performances have been merged into a single rendering, but at different playback rates.

In February of 2022, the same team behind ST-NeRF released a new project which claims to enable real-time rendering of NeRF environments at a processing rate 3000 times faster than the original NeRF implementation, and an order of magnitude faster than the best competing state of the art approaches.

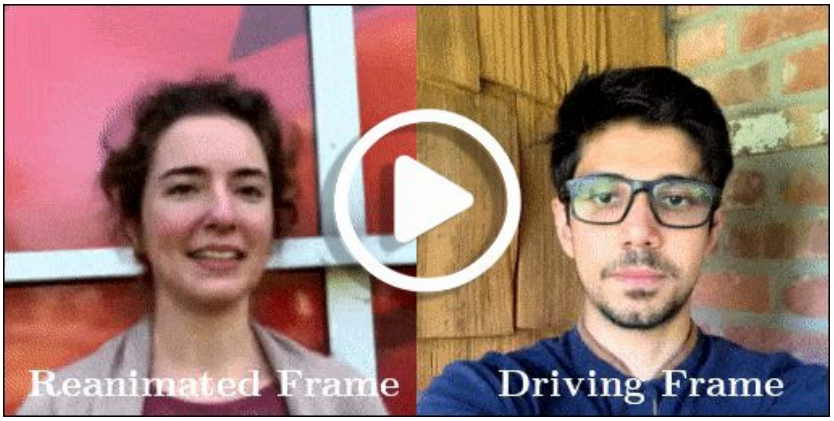
A novel implementation of the Plenoctrees algorithm has allowed researchers in China to create virtual humans in a dynamic interactively-rendered neural network. Project: Fourier PlenOctrees for Dynamic Radiance Field Rendering in Real-time, https://arxiv.org/pdf/2202.08614.pdf

The authors contribute to the ShanghaiTech Digital Human Project, which is actively engaged in advancing the cause of NeRF-based digital humans. Among its projects is the recent HumanNeRF, an initiative that offers free-view exploration and manipulation of neural humans captured from studio settings.



CLICK TO PLAY VIDEO ON SITE

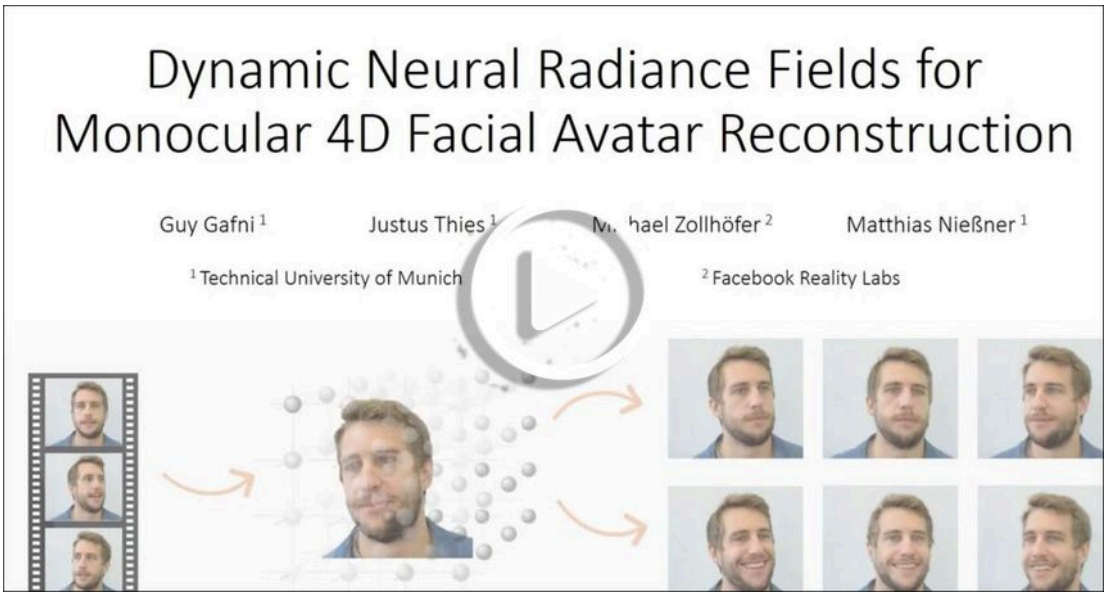
Meanwhile 2021's FLAME-in-NeRF, a collaboration between Adobe and Stony Brook University, uses Neural Radiance Fields to control and puppet expressions, allowing arbitrary expressions and novel views.



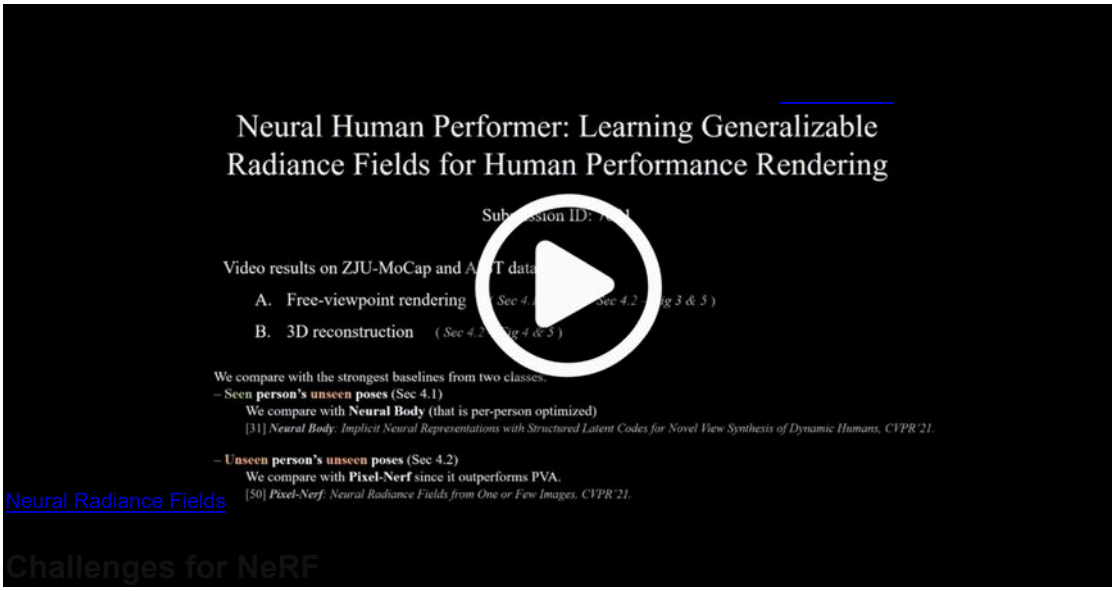
Flame-in-NeRF produces free viewpoint synthesis of face subjects that can be 'driven' by a source video (on the right). Source: https://arxiv.org/pdf/2108.04913.pdf

This kind of deepfake puppetry is a popular strand in NeRF research, with some of the strongest new initiatives emerging from Asian research communities. AD-NeRF, a collaboration between four Chinese universities, addresses the popular challenge of generating video from audio speech, using NeRF techniques to create synthetic face reconstructions that are driven by interpreted audio recordings:

A 2020 collaboration between Facebook Reality Labs and the Technical University of Munich offers Dynamic Neural Radiance Fields, an AR/VR-centric application of NeRF that uses a low-resolution 3D model to provide the 'driving' movement model, removing the need for expensive and elaborate facial capture setups.



Neural Human Performer, a joint research project from Adobe Research and two universities in the US and Korea, is a seminal initiative towards the fabrication of controllable parametric humans derived from NeRF workflows. NHP is capable of generating 'unseen' poses (i.e. abstract and arbitrary poses that were not originally captured in source footage):



NeRF offers a path forward to pure neural rendering of humans, authentically capturing not only facial characteristics, but correct proportions and behavior for the *entire body* of the target character or personage. This is almost certainly unattainable functionality for the current crop of popular deepfake distributions, since their source architecture is not very extensible, and they remain rooted in the strictures and limitations of the original 2017 release.

On the other hand, NeRF faces many of the same technological bottlenecks as DeepFaceLab and its stablemates, most notably in the form of practical limitations for input size of training images (see above). Additionally, most of the current crop of NeRF accelerator initiatives sacrifice other useful features (such as flexibility and/or explorability) in exchange for low latency, more interactive environments, and savings on training time and storage.

Furthermore, current trends in neural network training make parallel computation or any kind of real scalability as challenging for NeRF as it is for 'traditional' deepfakes.

Finally, editing methodologies must be devised to make NeRF a truly controllable and flexible environment. This shortcoming is currently being addressed by early projects such as Control-NeRF, which can remove objects from NeRF scenes, and MoFaNeRF, [EDIT: defunct] which can perform free view synthesis, face editing and even face rigging.

Controllable human synthesis is becoming a crowded branch of NeRF research: other facial and full-body NeRF research projects include MirrorNeRF, A-NeRF, Animatable Neural Radiance Fields, Neural Actor, DFA-NeRF, Portrait NeRF, DD-NeRF, H-NeRF, and Surface-Aligned Neural Radiance Fields.

Challenges for NeRF

NeRF offers a path forward to pure neural rendering of humans, authentically capturing not only facial characteristics, but correct proportions and behavior for the *entire body* of the target character or personage. This is almost certainly unattainable functionality for the current crop of popular deepfake distributions, since their source architecture is not very extensible, and they remain rooted in the strictures and limitations of the original 2017 release.

On the other hand, NeRF faces many of the same technological bottlenecks as DeepFaceLab and its stablemates, most notably in the form of practical limitations for input size of training images (see above). Additionally, most of the current crop of NeRF accelerator initiatives sacrifice other useful features (such as flexibility and/or explorability) in exchange for low latency, more interactive environments, and savings on training time and storage.

Furthermore, current trends in neural network training make parallel computation or any kind of real scalability as challenging for NeRF as it is for 'traditional' deepfakes.

Finally, editing methodologies must be devised to make NeRF a truly controllable and flexible environment. This shortcoming is currently being addressed by early projects such as Control-NeRF, which can remove objects from NeRF scenes, and MoFaNeRF, which can perform free view synthesis, face editing and even face rigging:

Examples of facial transformations with MoFaNeRF. Source: https://neverstopzyy.github.io/mofanerf/

The Parallel Advantage

What NeRF *can* offer, however, is a different *kind* of scalability: the possibility of assembling complex and high resolution environments and objects from multiple NeRFs of lower resolution and complexity. This is the concept behind Block-NeRF, a new paper from UC Berkeley, Google Research, and autonomous driving research company Waymo.

Block-NeRF essentially orchestrates an array of NeRFs into a cohesive environment, collectively creating an 'uber-NeRF' that has greater breadth, resolution, scalability, and disentanglement than any single NeRF output could offer.

A similar approach was taken by the Chinese-led CityNeRF project from late 2021. CityNeRF provides a kind of 'NeRF on demand' facility, similar to the way that videogame assets are loaded when the player gets so near to them that they are likely to be needed, or that online maps will pre-load resources that are adjacent to the data that's currently being used:

CityNeRF assets are loaded dynamically as needed. Source: https://city-super.github.io/citynerf/

This is a core advantage for NeRF: deepfake architectures based on the 2017 code produce a *process* that cannot easily be abstracted into a multi-instance framework, whereas NeRF training produces a discrete *object* that can be used as a component in more complex objects.

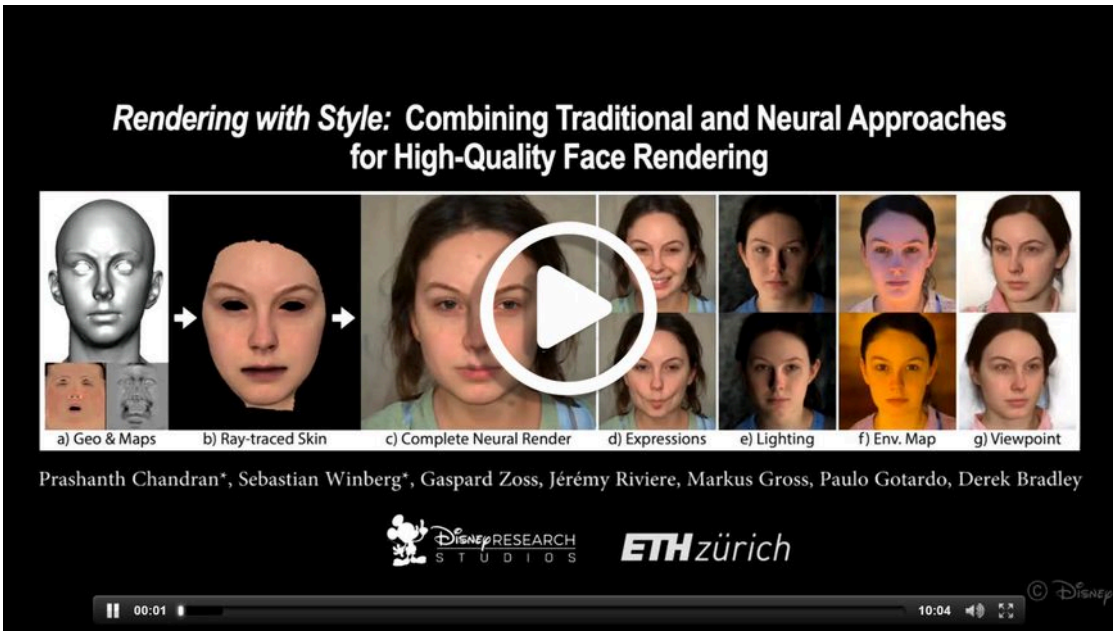
In terms of replicating human appearance, composable NeRFs are a clear possibility, with separate instances at least for the head and body, and granular sub-objects for hair, eyes, and any other asset of facial or personal appearance that could benefit from dedicated training and curation inside an orchestrated 'master' host.

GAN Vs. NeRF

NeRF is not the sole challenger to 2017-era deepfakes repositories. In the course of time we'll take a look at the intense research activity around the new crop of deepfake generation techniques that leverage Generative Adversarial Networks (GANs*), and other emerging and innovative routes to deepfake creation.

Besides the struggle to non-destructively lower its resource usage and training time, what stands against NeRF as a deepfake successor is that it's a new technology that's far behind GAN-based facial simulation research in terms of resolution, detail, and instrumentality.

On the plus side, besides being composable, NeRF draws *consistent* data unambiguously from the real world, and quantifies it in geometrical terms without getting lost in the mysteries of the GAN latent space — and without surrendering ground to hybrid, CGI-based approaches, such as Disney Research's offering from late 2021 (which superimposes 'traditional' texture maps into the latent space of a StyleGAN2 network):



Perhaps inevitably, the two technologies may end up compensating each other's shortfalls: a paper released in February proposes Pix2NeRF, an architecture that feeds GAN-generated imagery into a NeRF pipeline:

Single-shot novel view synthesis with Pix2NeRF, which imposes the output of Generative Adversarial Networks into a NeRF workflow . Source: https://arxiv.org/pdf/2202.13162.pdf

Another crossover GAN/NeRF project is the US/China collaboration FENeRF, which leverages 'Semantic Radiance Fields' to generate consistent novel views of GAN-generated faces.

Since NeRF is fueled by real world imagery, such crossover projects could offer a way to create stable and explorable fictitious neural environments, including GAN-generated people, without the need to map a GAN's latent space through relatively clumsy tools such as GradCAM heatmaps.

Naturally, there is nothing that prevents the use of traditionally rendered CGI content as source material for NeRF. CGI practitioners concerned about the advance of such neural rendering techniques are likely to have some years yet to catch up to the AI VFX scene, since physics simulation (water, fire, kinetics, inverse kinematics) is a relatively laggard area of neural rendering research at the moment, while textural stability and temporal coherency remains a problem for GAN approaches, and the aforementioned hardware limitations currently affect both technologies.

Conclusion

It may be that the 2017 deepfakes release will be remembered as the last time that an 'all-in-one' solution proves capable of delivering state-of-the-art facial and identity simulation without adjunct technologies, proprietary code, or the need for computing resources that are inaccessible to hobbyists and consumers.

Nonetheless, NeRF's late entry into the race for Deepfakes 2.0 represents the tantalizing possibility of an end-to-end neural workflow for deepfake creation that could yet challenge the state of the art while remaining in the open source arena.

***** *GAN is used only as a refinement tool in DeepFaceLab, and not at all in FaceSwap. The functionality of these projects is based on an encoder-decoder architecture, not a GAN architecture.*