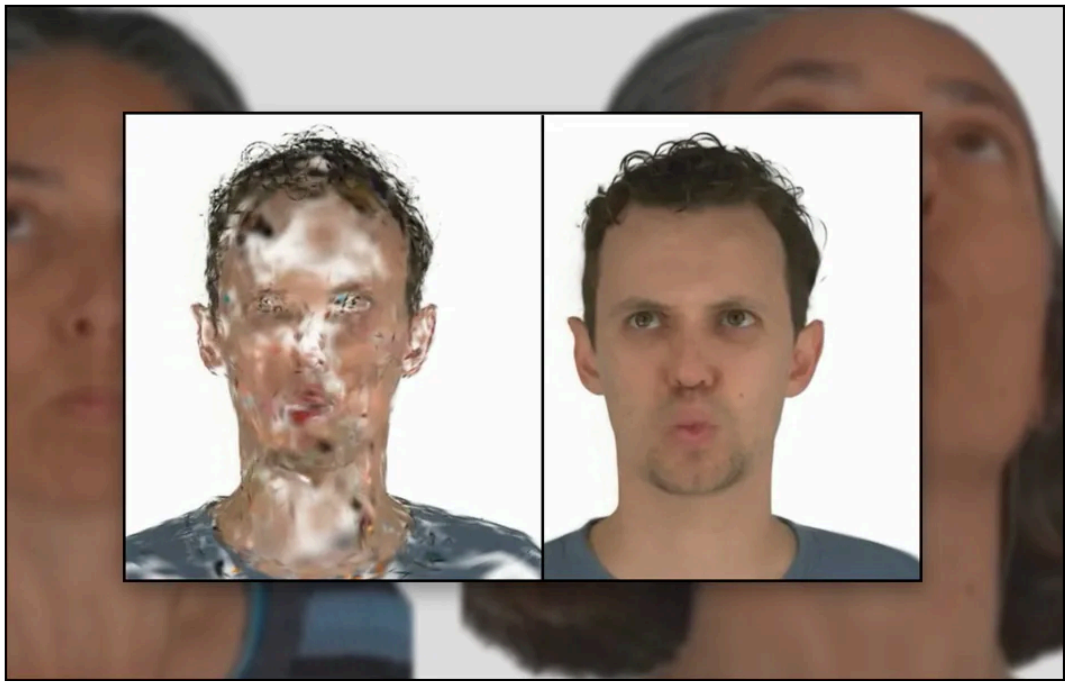
Streaming AI Avatars Like It’s 1999

Originally published at https://www.unite.ai/streaming-ai-avatars-like-its-1999/



New research presents a way to stream lifelike 3D avatars that appear almost instantly and sharpen in real time, instead of forcing users to wait for massive downloads to finish.

In many ways, the enormous resource demands of generative AI and AI-assisted rendering systems have taken consumer-readiness back twenty or more years. Only in 2023, a 64GB RAM allocation in a laptop or desktop PC seemed like overkill; now, with the growing popularity of RAM and/or CPU offloading, 64GB is pretty modest for local AI needs; and these once-banal and affordable elements of PCs continue to rocket in price as corporations struggle to meet demand for AI services.

The scale and greed of AI and its processes and environments typically dwarfs consumer-level hardware, and even running ‘slimmed down’ local-oriented models as GGUF versions will typically strain the average system.

Even text-based AI services such as ChatGPT are subject to significant strain both at the client and server level. Therefore, once AI is tasked with delivering online multimedia experiences in real time, we can reasonably expect some very serious compromises in latency and/or quality – similar to the internet’s early struggles with streaming media, and the much-hated animated ‘buffering’ icons of RealPlayer and QuickTime.

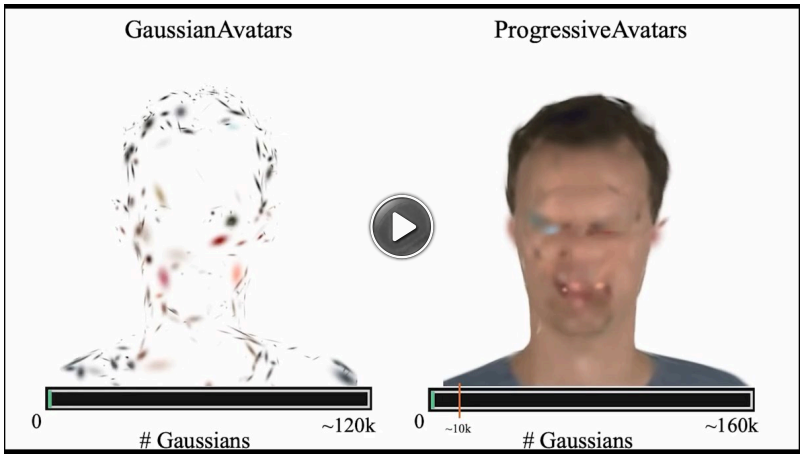
The last time that multimedia and network issues created friction in the user experience, consumer-level hardware was still evolving through Moore’s Law, getting almost exponentially better each year, even as Oses, networks and other supporting infrastructure evolved to meet demand; and for the last ten years, more or less, the capabilities of consumer technology have exceeded multimedia demands (perhaps even to the point where churn needed to be kick-started in order to maintain sales).

But that surfeit of local capability may be coming to an end soon, as local hardware becomes lower-specced and more expensive, and as AI-based services demand higher server-side and local resources.

Getting a Head

Back in the pre-broadband era, even before the earliest usable streaming video, web users were accustomed to images slowly coming into focus, as progressive JPEGs allowed the bandwidth-starved user to watch the downloading image forming, sometimes painfully slowly, as more of the image data was loaded locally.

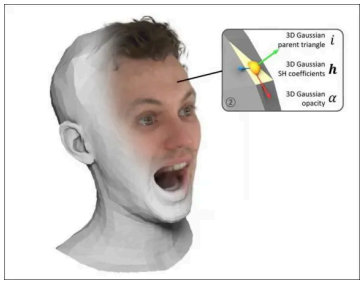
Now, it seems, we could be in for a similar experience with AI-aided Gaussian Splat avatars:



CLICK TO PLAY VIDEO ON SITE

Click to play. *From the new ProgressiveAvatars project, a comparison of streaming Gaussian avatars. On the left, the older GaussianAvatars project slowly gets new data but looks terrible as the data builds up; on the right, the Progressive Avatars version also builds detail slowly, but does so in an intelligent way that gives a basic human semblance right from the start.* Source

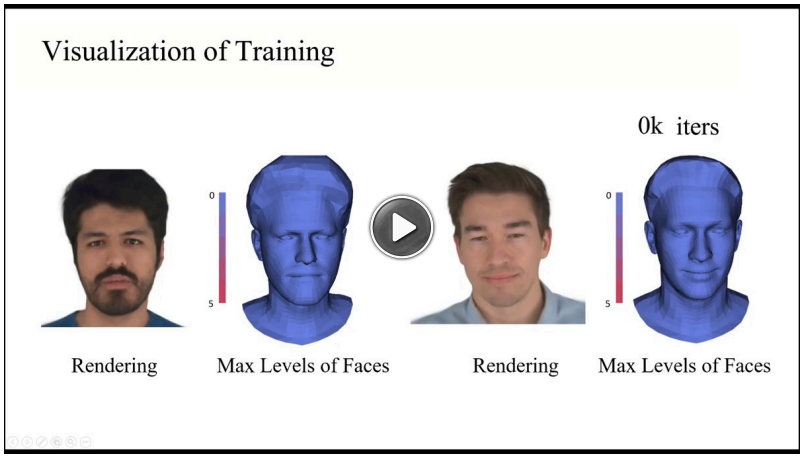
Above we see two versions of a Gaussian Splat-based (GSplat) Avatar – a human representation enabled partially by a non-AI rendering technique that hails back to the early 1990s, and also by more modern methods, such as the FLAME parametric human model, and AI-based training approaches:



Gaussian Splatting uses a Gaussian representation of color and 3D information in place of a pixel or voxel, and maps this ultra-realistic texture onto a more traditional kind of CGI mesh, which in itself is facilitated by a 'parametric human', a CGI face and/or body, in systems such as FLAME and STAR. Source

On the left in the video above we can see that a traditional implementation of a Gaussian splat avatar looks fairly horrific as we wait for the data to load. On the right, a new implementation from China, dubbed *ProgressiveAvatars*, is able to resolve far more elegantly as the data loads, presenting a non-alarming human image right from the start.

The authors claim that their method is the first to truly 'stream' a Gaussian avatar, and certainly the first to do so in a progressive manner, where the image builds up elegantly, and the most important areas – such as eyes and lips – can be prioritized, so that the avatar can become conversational even when only partially-loaded:



CLICK TO PLAY VIDEO ON SITE

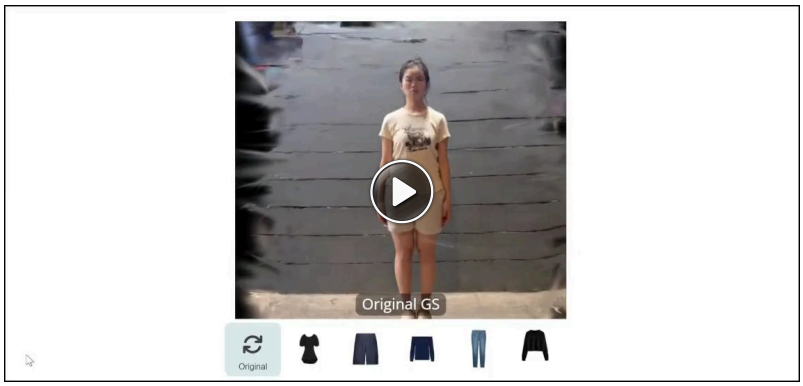
Click to play. *From the ProgressiveAvatars project site, an illustration of attention-aware loading.*

Prior to this, a 'level of detail' (LOD) approach has been used in previous attempts to slim down 'GSplat' avatars, similar to video-game optimizations, where successively more detailed versions of a person are loaded according to whether they occupy enough of the viewport or viewer attention to be worth the effort.

Of course, this entails a severe amount of redundant 'spare' avatars, and the authors frame their approach as a more rational system. By implication, a method of this kind also allows changes to be made in a GSplat figure (i.e., customization) without needing to propagate such changes through a chain of various LOD 'twins'.

An Emergent Domain

If this seems like a niche problem, well, so did streaming video, back in the days when getting the earliest plugins to work was out-tasked to the nearest available nerd. Further, the potential for AI-based streaming representations goes beyond human avatars, extending to city generation, games, and 3D-based* versions of practically any online domain – such as Virtual Try-On, for clothes shopping:



CLICK TO PLAY VIDEO ON SITE

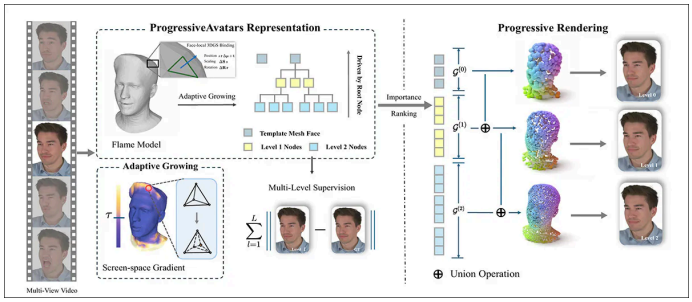
Click to play. *From a 2024 project, a rough look at the future of online 'try-on'. Other projects seek to add motion and interactivity – demanding facets to stream and manage.* Source

Just as LOD-based approaches have until now been mainly leveraged by video-games, many other considerations that were once the sole province of game development are likely to knock on into splat-based representations. For example, most of these early GSplat outings depict a *single human* gurning and mugging, or perhaps talking; but many situations will be needed that feature multiple humans, as well as environmental features and ambience – a scenario where highly performant 'triage' systems will determine where the streaming data needs to be prioritized, in order to keep the viewer in the moment.

The new paper is titled *ProgressiveAvatars: Progressive Animatable 3D Gaussian Avatars*, and comes from three researchers at the University of Science and Technology of China in Hefei.

Method

The approach initially leverages video of a person's head. For each frame, a standard FLAME parametric face model is fitted, so that the shape and expression change over time, while the underlying mesh structure stays fixed. Because the base topology does not change, a stable FLAME template can be reused and refined instead of rebuilt from scratch at every moment, as happens in similar prior works:



Head video is first fitted with a tracked FLAME mesh, after which 3D Gaussians are attached to each face and grown hierarchically where screen-space gradients indicate missing detail. During training, this adaptive subdivision builds a multi-level representation under multi-view supervision, and at inference, per-face importance scores determine which Gaussians are streamed first, allowing the avatar to appear quickly and refine progressively as higher-detail levels are added.

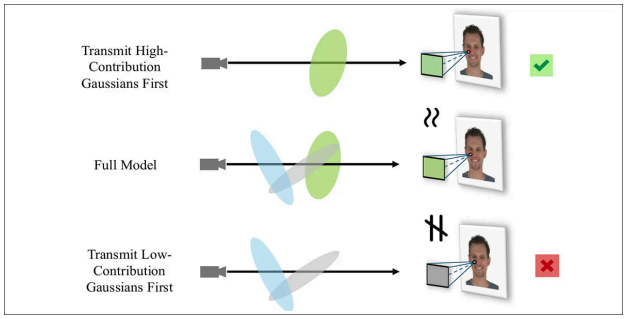
Over this base structure, detail is added in layers; the surface is implicitly subdivided into a hierarchy, and small three dimensional Gaussians are attached to the faces at each level of detail.

Though the initial coarser layers capture the overall head shape and motion, the subsequent finer layers provide wrinkles, subtle deformations, and high frequency texture. Images are then rendered from these Gaussians using a differentiable Gaussian rasterizer and trained against multi-view ground truth footage, so that the avatar learns to reproduce the real person's appearance.

During training, this hierarchy grows automatically: regions that need more detail are subdivided further, guided by screen space signals, so that computational effort concentrates where the viewer's eye is most likely to notice errors.

During inference, this same hierarchy enables *progressive streaming*, wherein a rough version of an avatar can be displayed first, and, as additional layers are loaded, new Gaussians can be added without altering what is already shown, enabling an animatable head avatar that appears quickly, and becomes sharper and more detailed as more data arrives.

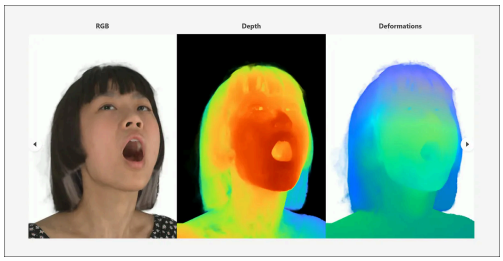
The authors observe that the entire system hinges on the prioritization of incoming data:



When all Gaussians at a given level are available, the full model is rendered with maximum fidelity; but during streaming, sending the highest-contribution Gaussians first allows early partial results to closely match the final image, whereas transmitting low-contribution Gaussians first distorts color balance and emphasizes minor components.

Data and Tests

For tests, the new method was evaluated on the NeRSemble dataset, which consists of multi-view videos for each subject covered, with calibrated parameters across all views:



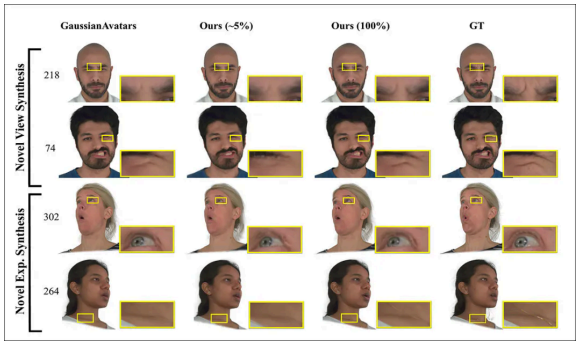
Examples of diverse interpretations of subjects included in the NeRSemble dataset used in tests for ProgressiveAvatars. Source

In line with the original GaussianAvatars methodology, images were downsampled to 802x550px a foreground mask generated, and the original project's training/test split adopted.

The Adam optimizer was used for parameter updates, with a learning rate of 1×10^{-2} on all barycentric coordinates. Training ran for 60,000 iterations, with the hierarchy automatically expanded every 2,000 iterations.

Initially, the authors tested for *reconstruction and animation* – the task of converting flat video into a 3D-aware (x/y/x) system, using FLAME's canonical CGI representation as the anchoring mesh. For this, all the baselines were trained from scratch, and the rival frameworks tested were the aforementioned GaussianAvatars, and PointAvatar.

For these tests, metrics used were Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS):



Qualitative comparison on novel-view and novel-expression synthesis. The baseline GaussianAvatars struggles with fine detail around eyes, wrinkles, and skin texture, while the proposed method already preserves key facial structure at roughly five percent of transmitted data and converges toward ground truth as more Gaussians are streamed, closely matching the full model and reference images (ground truth).

Regarding these results, the authors assert:

> *'[Our] method reconstructs sharper details in several regions, particularly around the neck, shoulders, and clothing. These areas are relatively coarsely tessellated in the FLAME template compared with high-saliency facial zones (e.g., the periocular region).*
>
> *'Consequently, prior methods often allocate too few 3D Gaussians to these regions to faithfully capture their fine-scale detail. In contrast, our adaptive growing strategy increases the number of Gaussians and refines the hierarchy only where needed, making allocation insensitive to FLAME's non-uniform tessellation.'*

The authors further note that their approach is on a par with state-of-the-art methods, yielding a workable avatar with a trivial 5% bandwidth allowance:

Method	NVS			NES		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
PointAvatar	25.8	0.893	0.097	23.4	0.884	0.102
GaussianAvatars	31.1	0.937	0.064	25.8	0.911	0.076
Ours (5%)	27.9	0.851	0.186	25.1	0.804	0.176
Ours (100%)	31.5	0.929	0.068	25.9	0.908	0.080

Quantitative comparison on novel view synthesis and novel expression synthesis using PSNR, SSIM, and LPIPS. At full transmission, the proposed method achieves the highest PSNR on both tasks and remains competitive with GaussianAvatars on perceptual metrics, while the 5% setting illustrates the quality trade-off under extreme bandwidth constraints.

Next, the researchers tested the progressive rendering itself. This was undertaken on a NVIDIA RTX 4090, with 24Gb of VRAM, at 550x802px resolution. In this scenario, the authors point out, a 25% budget would use up all the 'level 1' Gaussians, as well as a subset of level 2 Gaussians, which gives a rough overview of the way that the Gaussian groupings accrue detail in the higher number groups, and that the lower number groups essentially build the base canvas:

	NVS			NES			#Gaussians	Transmit Data	FPS
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓			
GaussianAvatars	31.10	0.937	0.064	25.80	0.911	0.076	163,829	41.90 MB	271
5% (Base)	27.89	0.851	0.186	25.13	0.804	0.176	10,144	2.60 MB	291
25%	29.14	0.892	0.080	25.58	0.846	0.124	37,302	9.56 MB	278
50%	30.03	0.904	0.073	25.71	0.884	0.105	84,132	21.56 MB	258
100%	31.47	0.929	0.068	25.89	0.908	0.080	169,438	43.42 MB	260

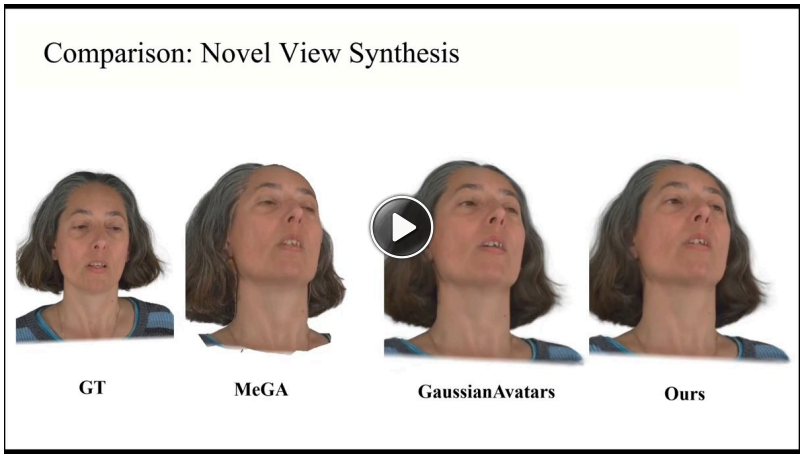
Performance under different transmission budgets for novel view and novel expression synthesis, showing that quality steadily approaches or exceeds GaussianAvatars as more Gaussians and data are streamed, while real time speeds are maintained, on an RTX 4090.

The authors comment:

> *'With only 2.60 MB transmitted (5% budget), the avatar already attains reasonable quality. As higher-level Gaussians are streamed, fine structures such as shirt buttons, teeth, and hair gradually sharpen while temporal stability are maintained.*
>
> *'At 100% transmission, our approach achieves rendering quality comparable to SOTA methods. Notably, the frame rates do not drop significantly, likely because the 3DGS workload has not yet saturated the GPU.'*

However, the authors point out that in multi-user VR scenarios, the number of 3D Gaussians would quickly grow to the point where GPU rasterization becomes a bottleneck. In those heavier scenarios, the proposed approach offers an advantage by allowing the system to trade off the number of primitives against visual quality, easing the load without collapsing the render.

Though the paper does not detail it, the project site features additional test comparisons, also featuring the MeGA Hybrid mesh-Gaussian avatar project:



CLICK TO PLAY VIDEO ON SITE

Click to play. *One of a series of supplementary videos from the paper's accompanying project site, this one comparing the new approach in terms of novel view synthesis.*

Conclusion

Gaussian Splatting may or may not endure, or even be remembered much more than RealPlayer now is, in regard to the dawn of interactive streaming: AI-driven or AI-aided 3D-aware representative experiences, including video chat, virtual shopping, route navigation, and diverse entertainment applications. It could be that alternate technologies or approaches win out, or that GSplat proves the most reliable AI-video representation.

If nothing else, this interesting new paper heralds a little of the scope of this new domain, while reminding us, perhaps nostalgically, of the bandwidth-starved internet of yore.

** By '3D', I do not mean the kind of experience that requires special glasses, but rather experiences where the multimedia content has some kind of understanding of X/Y/Z coordinates.*

First published Wednesday, March 18, 2026

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