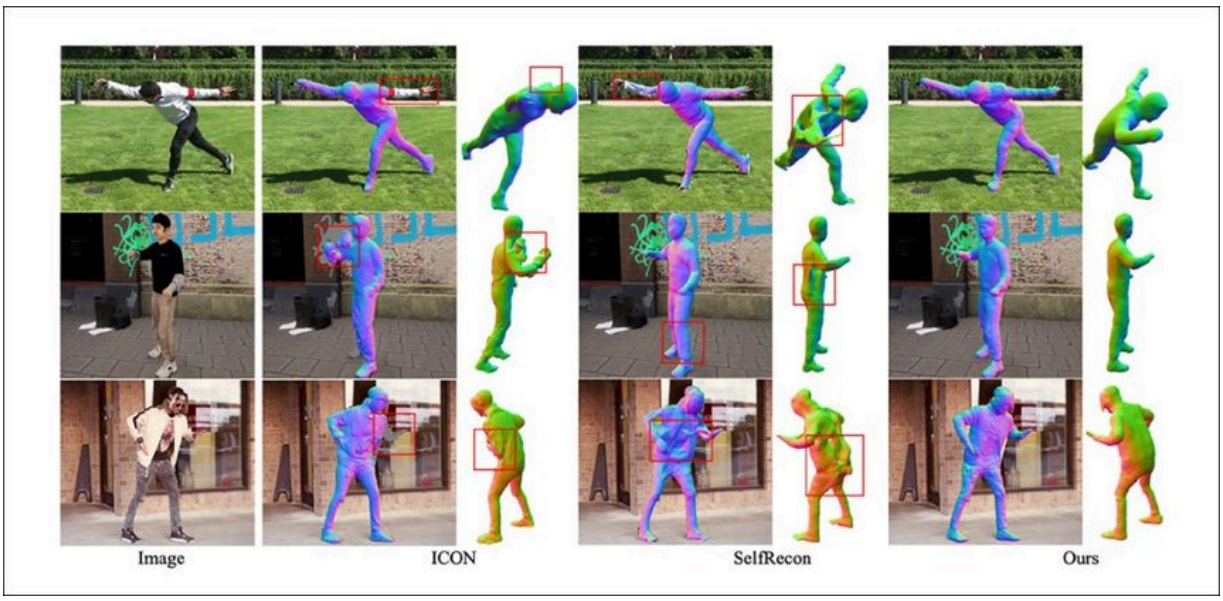
Obtaining Editable Neural Humans From Short Video Clips

Originally published February 24, 2023 at https://blog.metaphysic.ai/obtaining-editable-neural-humans-from-short-video-clips/



Currently, one of the most interesting pursuits in neural image synthesis is the search for a method that can easily extract an accurate neural representation of a person from a short video clip – because a video segment in which a person moves and shows themselves in different positions and attitudes can stand in nicely for the painfully-curated dataset that would otherwise be needed in order to obtain all these views as separate images.

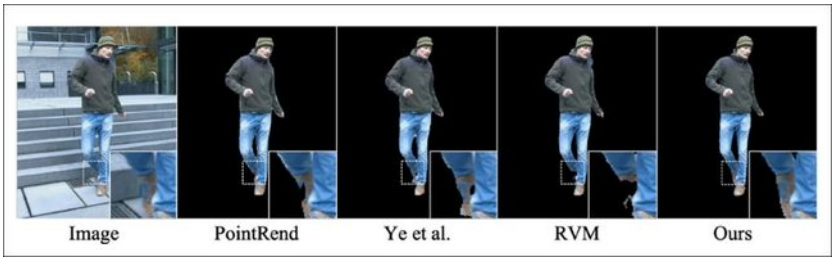
But it's not an easy task, since the objective is to obtain this information from *ad hoc* videos, where the individual is not conveniently standing in front of a green screen or against a non-detailed background. In the real world, the task of disentangling the human from their environment is not a trivial objective.

However, a new paper from ETH Zurich and the Max Planck institute, among the leading research groups into 3DMM-style parametric human neural synthesis, offers an improved method for extracting an 'instrumentalized' neural human from a brief and random 'in-the-wild' video clip.

CLICK TO VIEW ANIMATED GIF

Vid2Avatar can extract configurable neural humans from exemplary source clips. Source: https://www.youtube.com/watch?v=EGi47YeIeGQ (also embedded at the end of this article)

The new approach, called *Vid2Avatar*, improves upon previous methods by dedicating a separate Neural Radiance Field (NeRF) network to the human in the video, and another to the background against which the person is performing movement, obtaining more accurate detail, in tests, than prior methods.



Vid2Avatar obtains more realistic cloth detail and more complete humans than any method yet devised to derive editable neural humans from a brief monocular source video. Source: https://arxiv.org/pdf/2302.11566.pdf

The resulting neural avatar is effectively an animatable entity which can be textured either according to the captured data (recreating the appearance of the person in the original video) or, in theory, mapped to an altered appearance. It can also be constrained into new dynamic poses and movement sequences:

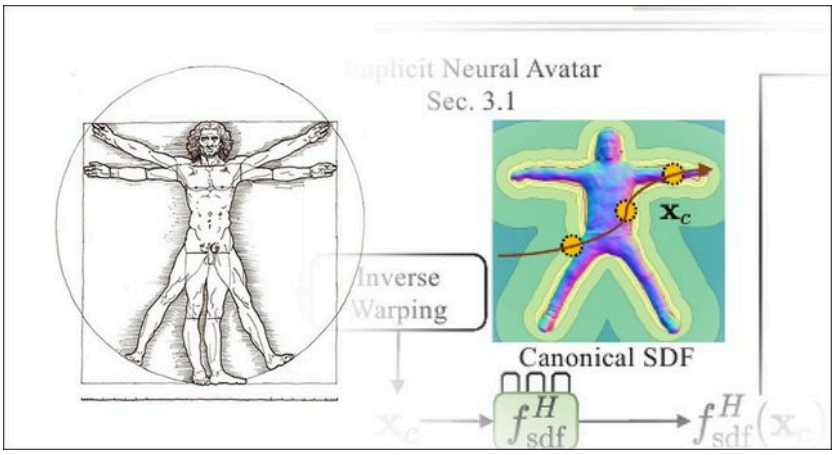
CLICK TO VIEW ANIMATED GIF

In the example materials in the paper's accompanying video (embedded below), the extracted neural humans have the generic appearance of *normal maps* – color-coded gradations of tone that represent 'blank' 3D coordinates into which texture can be applied; these can then be treated as an accessible and editable canvas.

The new paper is titled *Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition*, and comes from five researchers at ETH Zurich (one of whom is affiliated with the Max Planck Institute for Intelligent Systems at Tubingen).

Approach

Central to Vid2Avatar's approach is canonicalization – the extraction, from *ad hoc* video material, of a 'teleological ideal' for the person being represented. Da Vinci's Vitruvian Man prefigured (or, perhaps, informed) this concept in 3D neural reconstruction – the notion of a central 'default' reference point against which derivative deformations can be applied (i.e., *'bending an arm'*, *'bending down'*), all of which proceed from the ideal 'at rest' posture.

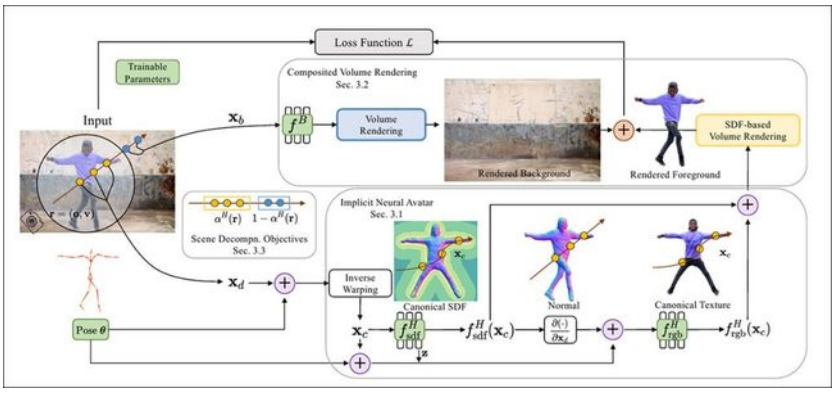


Da Vinci's 'Vitruvian Man', resolved into an equally 'default' canonicalization in Vid2Avatar.

Users of traditional professional and amateur CGI software such as Poser and the Daz line will be familiar with this 'starting position' that's supplied when a new instance of a catalogue parametric model is created. However, it's not easy to obtain this gold-standard reference from the random and goofy movements people might make in a short video clip.

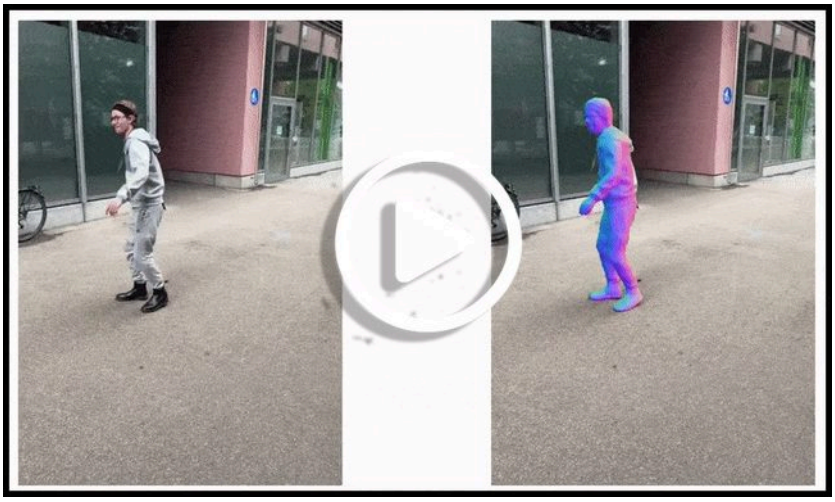
Therefore Vid2Avatar converts the perceived human movement into 3D space via a Signed Distance Field (SDF), frame by frame, until enough views that correspond to points on the 'default' template have been obtained in order to fully populate a canonical representation of that person.

CONTINUED ON NEXT PAGE



Schema for the workflow of Vid2Avatar.

Naturally, if the person's movement is unusually constrained or conservative, the ultimate obtained canonical representation is going to be missing some necessary information, and the neural representation will be less accurate. Therefore video clips that feature extensive coverage of the individual are ideal for extracting comprehensive material for the canonical representation.



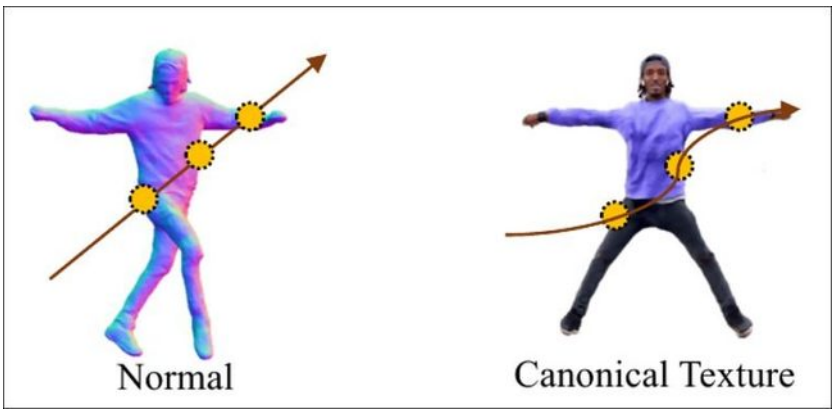
CLICK TO VIEW ANIMATED GIF

The more that the individual moves in the source video, the more complete and comprehensive their canonical representation will be.

In Vid2Avatar, the pose parameters are initially obtained from the source video via the older Max Planck technology *Skinned Multi-Person Linear Model* (SMPL), which makes use of a traditional CGI model, which can be mapped to neural coordinates, and acts as an interface for the neural representation.

After this point, any deviation from the obtained canonical representation of the person is considered a 'deformation', much as it is in traditional CGI modeling.

The texture data has its own dedicated network, which utilizes surface normals – mapped relationships which can be depended on to stay where they are (relative to any deformations that may be applied), which allow for consistent texture-mapping.



A canonical texture is also created for the canonical neural representation. On the left we can see traditional-looking surface normals in the established representation – the kind of process and tools that CGI artists were working with as far back as the late 1970s (see https://www.microsoft.com/en-us/research/wp-content/uploads/1978/01/p286-blinn.pdf).

Vid2Avatar uses a system of 'composited volume rendering' to bring all the disparate incoming elements together. The principle components here are two Neural Radiance Fields – one 'uber-NeRF' for the background representation, and one 'inner NeRF', where the human data is processed.

The schema for this workflow is derived from NeRF++, a 2020 collaboration between Cornell Tech and Intel Labs.

The effective result of concatenating the various networks involved in the process comes in the form of two renderable neural assets – the background, which has been disentangled from the human content through differentiable analytical processes, and the extracted human, now a discrete neural entity.

We can see on the left of the image above that the person in the source video has been 'painted out' of the data. This is because the background has, like the person, been estimated and assembled as a separate object, based on movement in the video. Just as with the human, if the background is not adequately exposed in the video, there are going to be gaps in this representation.

Tests

In extensive testing rounds, the researchers compared Vid2Avatar with comparable frameworks, using a variety of metrics, including F1 score and Mask IoU, for human segmentation evaluation; volumetric IoU, Chamfer distance and normal consistency, for surface reconstruction; and Structural Similarity Index (SSIM) and Peak signal-to-noise ratio (PSNR), for rendering quality.

For the 2D segmentation comparison, which evaluates how well the system can mask out human from background content, Vid2Avatar was pitted against SMPL tracking, PointRend, Deformable Sprites ('*Ye et al.*', in the results below), and Robust High-Resolution Video Matting with Temporal Guidance (RVM), using the MonoPerfCap dataset.

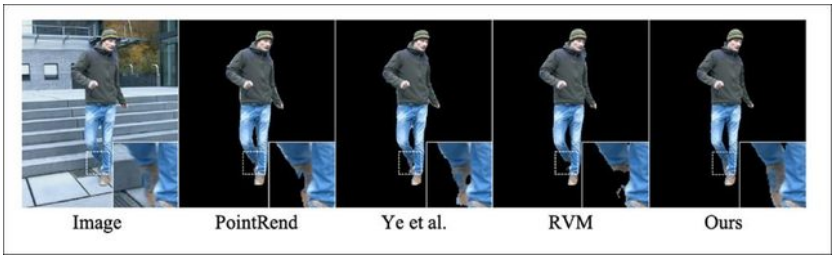
For this, PointRend and RVM were trained on 'large datasets' that contained human-annotated masks, while Deformable Sprites instead relies on optical flow, which effectively generates its own data dynamically.

Method	Precision ↑	F1 ↑	IoU ↑
SMPL Tracking	0.829	0.781	0.659
PointRend [26]	0.957	0.960	0.915
Ye et al. [62]	0.945	0.947	0.890
RVM [30]	0.975	0.977	0.950
w/o Scene Dec. Loss	0.979	0.974	0.942
Ours	**0.983**	**0.983**	**0.961**

Results from tests on segmentation performance.

Of these results, the authors state:

'Our method consistently outperforms other baseline methods on all metrics. [The results show] that other baselines struggle on the feet since there is no enough photometric contrast between the part of the shoes and the stairs. In contrast, our method is able to generate plausible human segmentation via decomposition from a 3D perspective.'



Qualitative observations on the segmentation results. Vid2Avatar is able to disentangle feet better from the background than rival networks.

For the view synthesis comparison round, which evaluates how well the system can generate novel views from the extracted avatars, Vid2Avatar was tested against HumanNeRF and NeuMan. The frameworks were tried on the NeuMan dataset, which features a collection of casual videos shot on a mobile phone.

Method	SSIM ↑	PSNR ↑
NeuMan [24]	0.958	23.9
HumanNeRF [52]	0.963	24.8
Ours	**0.964**	**25.1**

Results from the view synthesis round.

Here, the authors comment:

'Overall, we achieve comparable or even better performance [quantitatively]. [NeuMan] and HumanNeRF have obvious artifacts around feet and arms. This is because, a) off-the-shelf tools struggle to produce consistent masks and b) NeRF-based methods are known to have "hazy" floaters in the space leading to visually unpleasant results. Our method produces more plausible renderings of the human with a clean separation from the background.'

To compare the abilities of the various architectures in terms of surface reconstruction, Vid2Avatar was challenged by Max Planck's *Implicit Clothed humans Obtained from Normals* (ICON) and the 2022 SelfRecon project, from China. The dataset used was *SynWild* – a new collection from the authors themselves. For this novel synthetic dataset, the researchers reconstructed detailed geometries and textures using High-Quality Streamable Free-Viewpoint Video. The obtained scans were then placed into photorealistic HDRI panoramas, and monocular videos rendered from virtual cameras via the Unreal game engine.

3DPW

Method	IoU ↑	C − ℓ_2 ↓	NC ↑
ICON [55]	0.718	3.32	0.731
SelfRecon [23]	0.648	3.31	0.675
w/o Joint Opt.	0.810	3.00	0.737
Ours	**0.818**	**2.66**	**0.753**

SynWild

Method	IoU ↑	C − ℓ_2 ↓	NC ↑
ICON [55]	0.764	2.91	0.766
SelfRecon [23]	0.805	2.50	0.776
Ours	**0.813**	**2.35**	**0.796**

Results from the reconstruction round.

Regarding the results of the reconstruction round, the authors state:

'Our method outperforms [the other frameworks] by a substantial margin on all [metrics]. The difference is more visible in qualitative comparison… where they tend to produce physically incorrect body reconstructions (e.g., missing arms and sunken backs). 'In contrast, our method generates complete human bodies and recovers more surface details (e.g., cloth wrinkles and facial features). We attribute this to the better decoupling of humans from the background by our proposed modeling and learning schemes.'

Results from qualitative reconstruction tests. See the source paper for better resolution and additional results, and check out the embedded video below for further details.

Conclusion

Avatar extraction from monocular video represents, effectively, the automated modeling, rigging and texture-surfacing of a human being from very little information.

Though much of the research currently being done – including this latest paper – is pushing the state-of-the-art in direct *video>neural* evaluation and synthesis, and though this will ultimately benefit the VFX industry, the clear aim here is to create automated pipelines for AR, VR, gaming and lightweight mobile environments.

This could ultimately mean that the 'character creation' stage of video-games will involve webcams or user-uploaded clips, and that similar procedures may be entailed in future metaverse environments.

Inevitably, there is some risk that the ability to upload videos of other people and end up with neural representations that can be controlled by someone else will necessitate a cautious attitude in the use of such technologies. For the time being, the resulting instrumentality is still firmly in the lab, and likely to be headed for the walled gardens of hyper-commercialized and locked-down use contexts.

CLICK TO PLAY VIDEO ON SITE