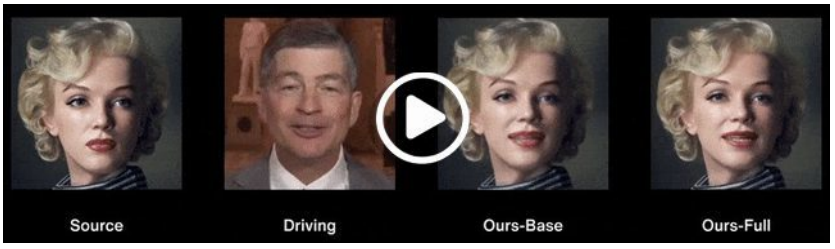
High-Quality Deepfake Puppetry in Thirty Seconds

Originally published December 16, 2022 at https://blog.metaphysic.ai/high-quality-deepfake-puppetry-in-thirty-seconds/

A new academic collaboration, including contributors from Microsoft, has developed a novel technique capable of fitting user-submitted images into a 'deepfake puppetry' workflow in only thirty seconds, with notably improved fidelity to the original identity.

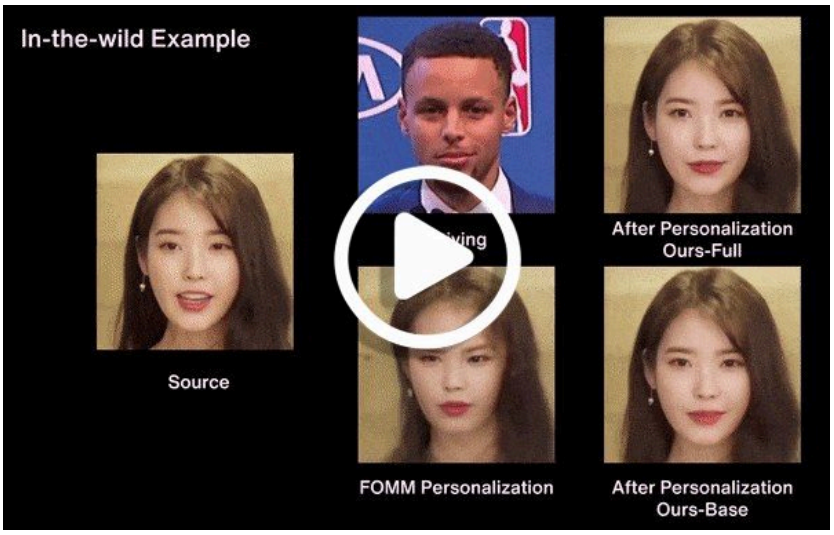


CLICK TO VIEW ANIMATED GIF

Marilyn, driven by only a single image (left) and a source video (second from left), with MetaPortrait. Please see source videos (embedded at the bottom of this article) for better resolution. Source: https://meta-portrait.github.io/

The new system, titled MetaPortrait, can be used both to recreate source video, or to take existing video footage and substitute the original identity with a new identity, based only on a single photo.

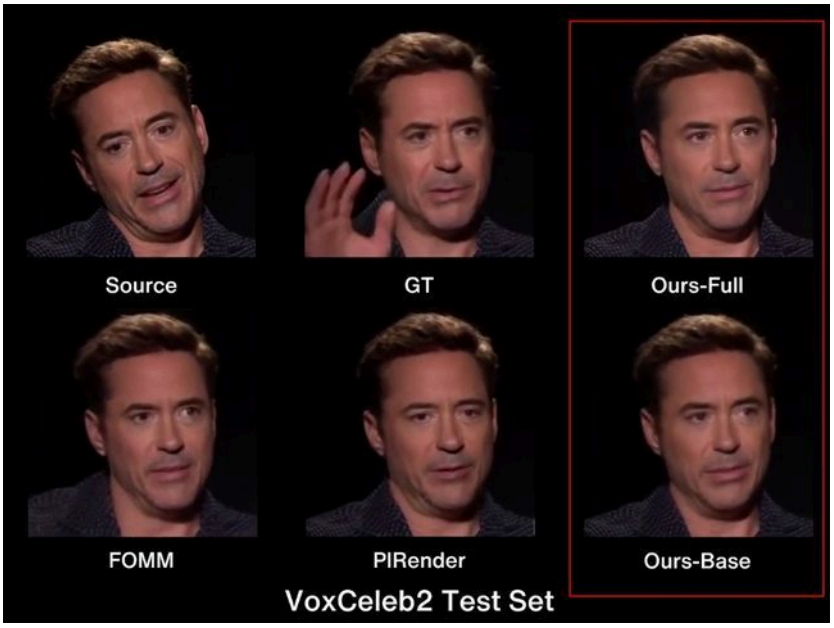
CONTINUED ON NEXT PAGE



CLICK TO VIEW ANIMATED GIF

Here we see MetaPortrait compared to the much-vaunted and oft-cited First Order Motion Model, obtaining clearly superior results whilst preserving fidelity to the input source photo.

The approach uses a variety of novel and prior techniques to both speed up and improve recreation and identity transfer, including an elaborate identity-preserving framework that offers a significant diminution in 'identity bleed', in cases where one person's motion is powering the movement of a 'target' identity – or even when the system is tasked with recreating a source video neurally:

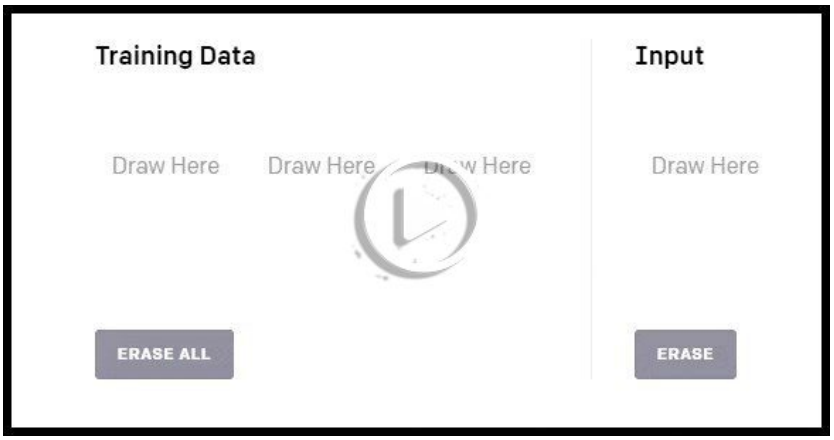


Prior approaches have suffered from loss of authentic identity during neural recreation, whereas MetaPortrait's dedicated identity-preserving module succeeds in more accurately reproducing the source video. Note that the original frame ('GT' is 'ground truth) is the middle image in the upper row. Source: https://arxiv.org/pdf/2212.08062.pdf

The ability to fit an arbitrary photo to the network is enabled by the use of Model-Agnostic Meta-Learning (MAML), a technique pioneered by the University of California at Berkeley in concert with OpenAI, who went on to develop the technique into the scalable meta-learning algorithm REPTILE (also leveraged in MetaPortrait).

REPTILE is a first-order MAML-based approach that constitutes a one-shot classification system, and is able to make accurate predictions based on live and unseen (i.e., experimental) data, without time-consuming pre-training or dedicated training.

CONTINUED ON NEXT PAGE



CLICK TO VIEW ANIMATED GIF

MAML, here demonstrated in a simple REPTILE demo from OpenAI, learns dynamically, enabling faster fitting in MetaPortrait. Source: https://openai.com/blog/reptile/

This approach radically cuts down on fitting requirements, enabling MetaPortrait to achieve the 30-second adaptation time, and allowing for a more deployable and low-latency workflow, compared to prior works such as the influential 2019 work, First-Order Motion Model (FOMM).

From the new paper, various historical personages (and even fictitious people) animated through MetaPortrait.

Testing the new system against comparable prior works, MetaPortrait outperforms all the baselines for image fidelity (how much the 'deepfake' or recreation resembles the target), and performs comparably in terms of motion transfer.

Methods	Self Reconstruction (256 × 256)					Cross Reenactment (256 × 256)			
	FID↓	LPIPS↓	ID Loss↓	AED↓	APD↓	FID↓	ID Loss↓	AED↓	APD↓
X2Face [44]	45.2908	0.6806	0.9632	0.2147	0.1007	91.1485	0.6496	0.3112	0.1210
Bi-layer [52]	100.9196	0.5881	0.5280	0.1258	0.0139	127.7823	0.6336	0.2330	**0.0208**
FOMM [34]	12.1979	0.2338	0.2096	0.0964	**0.0100**	80.1637	0.5760	0.2340	0.0239
PIRender [31]	14.4065	0.2639	0.3024	0.1080	0.0162	78.8430	0.5440	**0.2113**	0.0214
Ours	**11.9528**	**0.2262**	**0.1296**	**0.0942**	0.0124	**77.5048**	**0.2944**	0.2524	0.0258
Methods	Self Reconstruction (512 × 512)					Cross Reenactment (512 × 512)			
StyleHEAT [50]	44.5207	0.2840	0.4112	0.1155	0.0131	111.3450	0.4720	**0.2505**	**0.0218**
Ours	**21.4974**	**0.2079**	**0.0832**	**0.0904**	**0.0121**	**49.6020**	**0.1952**	0.2737	0.0242

Results of comparisons of MetaPortrait against prior works. Further details of frameworks and metrics below.

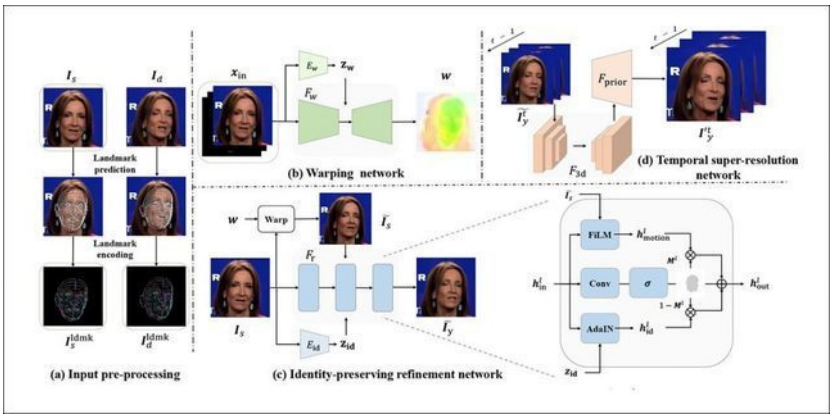
Where prior approaches are able to take the lead in this table, it must be considered that none of them can approach the speed of the new technique, and all of them have inferior identity representation.

Output from the system runs at 512x512px, with a dedicated upsampling super-resolution network that, innovatively, considers multiple adjacent frames in order to maintain temporal consistency, and also employs StyleGAN and 3D convolution for quality enhancement.

The new paper is titled *MetaPortrait: Identity-Preserving Talking Head Generation with Fast Personalized Adaptation*, and comes from nine researchers at the University of Science and Technology of China (USTC), the Hong Kong University of Science and Technology, and Microsoft.

Approach

The central premise behind the work is that dense facial landmarks are needed in order to preserve identity. MetaPortrait uses a landmark estimator that generates 669 landmarks defining the entire head, including problematic areas such as teeth, eyeballs and ears. More common approaches use less than 70 landmarks, leaving these 'details' to be handled at the pixel evaluation level when the information is fed to the warping network that will attempt to 'pair up' the source landmarks (i.e., the single photo of Marilyn Monroe in the earlier image) with those extracted from the driving source video.



An overview of MetaPortrait's one-shot framework.

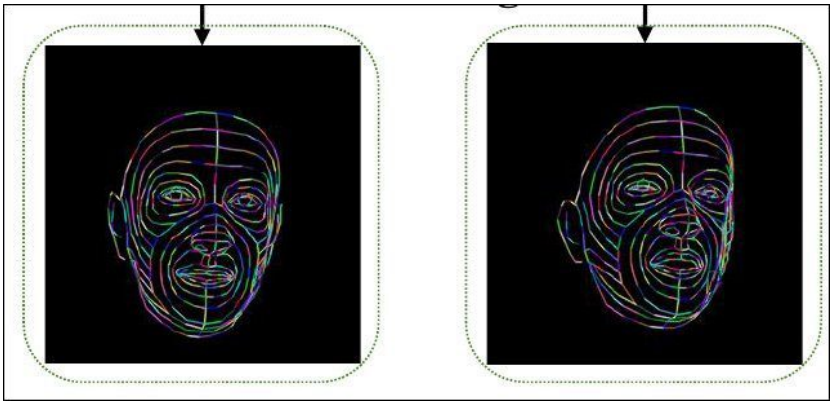
A pretrained landmark detector derives the point data, which is then fed into the warping flow between the source image and each individual driving frame in the video source. The result from this is then shunted into the ID-preserving network, before being upscaled from 256x256px to 512x512px.

The landmark estimator is trained on synthetic data from Microsoft's *Fake It Til You Make It*, Blender-centric system (also used in another new HKUST project of the last week).

The estimator itself is a deployment of Microsoft's own *3D Face Reconstruction with Dense Landmarks* (see video below), which scales up facial landmark capture into a far more complex and granular system that's better-suited to the current challenges of facial neural reconstruction than many of the popular older methods, such as the Facial Alignment Network (FAN) system used in popular open source deepfakes repositories such as DeepFaceLab and FaceSwap.

CLICK TO PLAY VIDEO ON SITE

The authors note that actually making use of such a high volume of landmark data is 'non-trivial', and corral the information by drawing neighboring landmark points in which each connection is color-coded:



Despite these challenges, the paper asserts that it is this improved landmark density (which would ordinarily slow down such a system) that's the key to the improved identity-fidelity in MetaPortrait.

MetaPortrait's architecture also incorporates an implementation of Feature-wise Linear Modulate (FiLM), a conditioning method that aids visual reasoning in computer vision workflows (not to be confused with Google Research's apparently apposite but unrelated *FILM: Frame Interpolation for Large Motion*).

Further, a multi-scale patch discriminator, NVIDIA's popular SPADE network, is used to improve the rendered quality of challenging areas such as the mouth and hard eye regions, aided by additional reconstruction losses.

Improved Personalization

In prior works, personalized fine-tuning has been a key method of improving fidelity and temporal coherency. Fine-tuning involves the partial retraining of a trained network, so that it conforms to the exact data at hand. An example of this might be that a generative model trained on many subjects would be later trained additionally on very specific, user-contributed data that was not present in the original training data.

Though many of the pitfalls of fine-training, such as the possibility of loss of general detail refinement, can be mitigated (or may not be a factor in a 'disposable' fine-tuned model that will only ever be called on to perform one specific task), the technique takes time; and the longer the source video is, the more time it will take.

For this reason the researchers have adopted MAML (see above), producing a model with 'ductile' weights that are more easily adaptable to new data, and permitting shorter fitting times.

Data, Experiments and Metrics

Following the methodology of the summer 2022 MegaPortraits paper, a Samsung-led academic collaboration, the authors trained MetaPortrait's warping and refinement networks on cropped material from the University of Oxford's VoxCeleb2 dataset, resulting in 256x256px input data.

500 videos were selected from VoxCeleb2 for evaluation, with the base MetaPortrait model fine-tuned on the HDTF dataset (and here fine-tuning is appropriate, since the operation leads to an implementable algorithm, and is not aimed at end-users).

HDTF provided 410 videos covering 300 different identities, and for the project's purposes these were split into 400 training videos and ten test videos (a typical configuration, where a certain portion of the data is 'held back' from training in order to be evaluated against the trained model – since the 'held back' data is definitely not 'out-of-distribution', or OOD, if the algorithm fails against it, then it has truly failed).

After meta-training (see above), the temporal resolution model (which upscales the output to a more usable 512px) was also trained on HDTF, which contains 300-frame videos at 512x512px resolution.

Downsampled 256px frames were then passed to the aforementioned warping and refinement networks and used as inputs for the subsequent super-resolution module.

Metrics used for evaluation were Frechet Inception Distance (FID) and LPIPS, with motion transfer quality evaluated by average expression distance (AED) and average pose distance (APD).

MetaPortrait was compared against four state-of-the-art networks: Oxford's X2Face, the above-mentioned FOMM, Bi-Layer (a collaboration between the University of Toronto and Google Research), and PIRender (from AIIT and Tencent).

Qualitative results of self-reconstruction for MetaPortrait

Of the qualitative results, the authors comment:

'Since our landmarks have a better decomposition of identity and motion and our refinement network is identity-aware, our method is the only one that well preserves the identity of source image.
'In contrast, previous methods suffer from appearance data leakage directly from the driver and generate faces with a similar identity to the driving image.'

There is one other similar project capable of outputting 512x512px resolution – StyleHEAT, a 2022 collaboration between the Tsinghua Shenzhen International Graduate School and Tencent. Therefore the MetaPortrait authors also tested their system against this.

CONTINUED ON NEXT PAGE

CLICK TO VIEW ANIMATED GIF

Comparison of MetaPortrait with StyleHEAT.

Regarding the results of the comparison, the authors observe:

'[StyleHEAT] fails to synthesize sharp and accurate teeth in our experiment, and the identity of the output image is quite different from the source portrait due to the limitations of GAN inversion. In contrast, our refinement network is identity-aware and we leverage pretrained StyleGAN in our temporal super-resolution module to fully exploit its face prior knowledge.'

As for the other technical details of the project, the 256px base model was trained on the VoxCeleb2 dataset at a batch size of 48 under the Adam optimizer on eight Tesla V100 GPUs. The warping network was trained for 200,000 iterations before being trained jointly with the refinement network for a further 200,000 iterations.

The authors conclude:

'Our model is able to provide state-of-the-art generation quality with high temporal fidelity on both self-reconstruction and cross-reenactment tasks. Moreover, the videos of fast personalization illustrate the strong adaptation capability of our meta-learned model. The in-the-wild examples also demonstrate the generalized ability of the proposed model.'

At the time of writing, no source code or additional supplementary material was available for the project.

EDIT – 5th January, 2023 – The three YouTube embeds were set to private by the researchers. You can now view the project's associated videos at the official page.