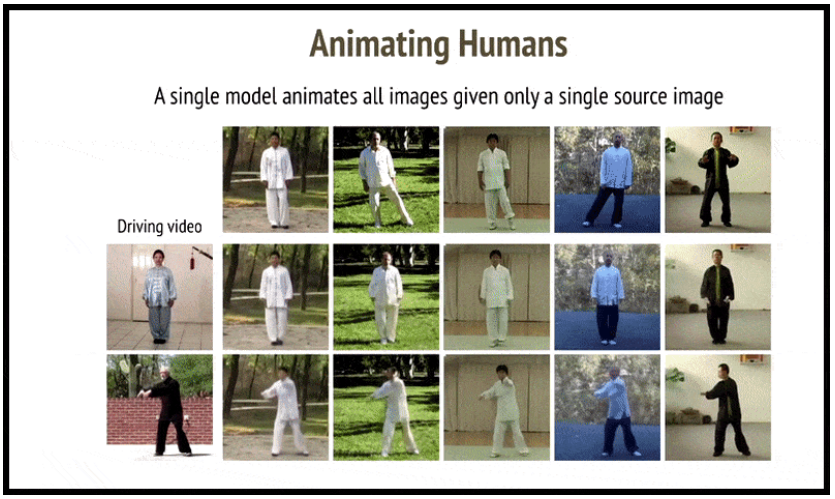
NVIDIA’s Implicit Warping Is a Potentially Powerful Deepfake Technique

Originally published at https://metaphysic.ai/nvidias-implicit-warping-is-a-potentially-powerful-deepfake-technique/

October 17, 2022

Over the past 10-20 years, and particularly in recent years, the computer vision research community has produced an abundance of frameworks capable of taking a single image and using it to perform ‘deepfake puppetry’ – the use of the facial and body movements of one person to simulate a secondary, fictitious identity.



A collaboration between researchers across Europe and Snap Inc., 2019's First Order Motion Model (FOMM) captured the imagination of some viewers at the time, even though it was clear that some of the movements it tried to simulate were literally a 'stretch' (see the middle row Mona Lisa above, which does not have enough base information to pull off the angle it is attempting) . Source: https://aliaksandrsiarohin.github.io/first-order-model-website/

This plethora of academic interest hails back to at least 2005, and includes projects such as Real-time Expression Transfer for Facial Reenactment, Face2Face, Synthesizing Obama, Recycle-GAN, ReenactGAN, Dynamic Neural Radiance Fields, and many others, diversely leveraging the limited available technologies, such as Generative Adversarial Networks (GANs), Neural Radiance Fields (NeRF) and autoencoders.

Not all of these initiatives attempt to derive video from a single frame; some of them perform computationally expensive calculations of each frame in a video, which is effectively exactly what deepfakes (in the sense of AI-powered viral celebrity impersonations) do. But since they are operating with *less information*, that kind of approach requires *per-clip training* – which is a step down from the open source approach of DeepFaceLab or FaceSwap, where one can train and use models capable of imposing an identity into *any* number of clips, not just one.

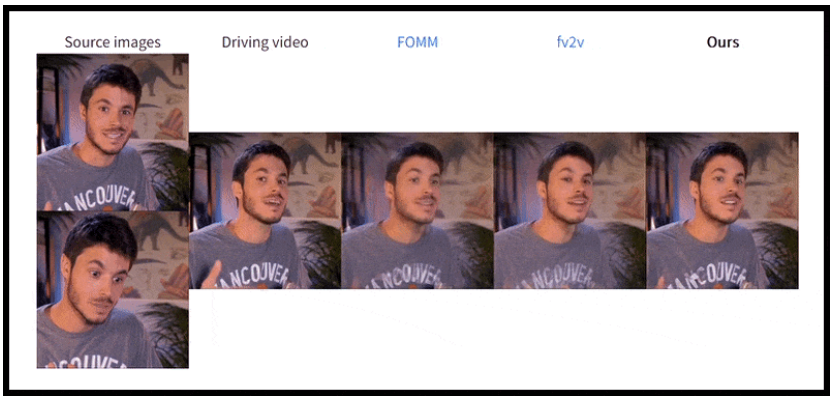
The others attempt to derive multiple poses and expressions from a *single face* or full-body representation; but this kind of approach usually only works with the most expressionless and immobile of subjects – and usually only in a relatively static ‘talking head’ situation, since there are no ‘sudden changes’ in facial expression or pose that the network will have to account for.

https://www.youtube.com/watch?v=3Mi1Ofdc5t4

Though some of these techniques and approaches gained public traction in the less desensitized time before the wider proliferation of deepfake technologies and – lately – latent diffusion image synthesis methods, they all seemed to arrive at the same end via slightly differing means, with their applicability limited and their versatility in question.
To be honest, we're a little immured to this kind of thing now, and more dazzling innovations have diverted our attention.

Reframing the Challenge

NVIDIA's computer vision research division has been developing a similar kind of system over the past few years, and lately the company has presented it in such a dull context (i.e., the by-now formulaic exact recreation of source videos via machine learning, which typifies this strand of research), that many may not have noticed how significant it could be.

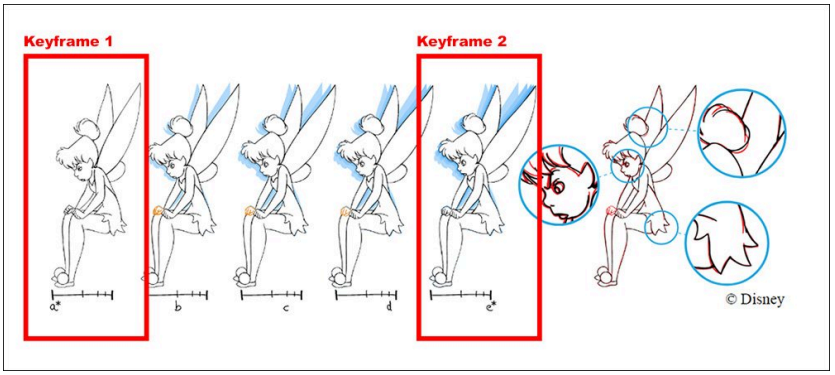


Rather than trying to get all the necessary pose information from a single frame, Implicit Warping can 'tween' between multiple frames, and even as few as two, in accordance with some of the oldest principles of traditional animation – a facility that is either absent or works very poorly in all rival or prior frameworks. Source: https://deepimagination.cc/implicit_warping/

NVIDIA's recently-published paper *Implicit Warping for Animation with Image Sets*, has done little to attract further attention to the project; likewise the extensive accompanying videos at the main project page and the additional results page – because, ironically, the more you succeed at recreating a source video by methods of this nature, the less evident the significance of the achievement is, with the results appearing redundant and repetitive of previous efforts.
In fact, Implicit Warping has extraordinary potential to create hyper-realistic deepfake motion, to an extent that none of its predecessors have been equipped to do.

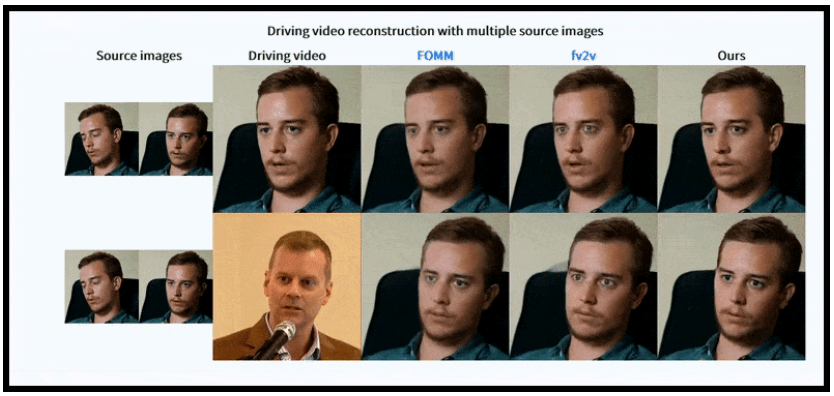
EbSynth on Steroids

The difference with the new method, titled *Implicit Warping*, is that, harking back to the earliest days of animation, it can 'tween' two (or any arbitrary number of) keyframes, instead of attempting to torment a single image into a range of dynamic poses and expressions that no single image can possible yield.



From a 2010 Disney paper, the earliest principles of inbetweening – where more junior animators would receive 'master frames' or keyframes from senior animators, and would be tasked with producing the interstitial frames. Source: https://media.disneyanimation.com/uploads/production/publication_asset/120/asset/Whi10.pdf

Tested against prior versions, the researchers of the new paper found that the quality of results from the older approaches actually *deteriorates* with extra 'keyframes', whereas the new method, in line with the logic of animation itself, improves in a quite linear manner as the number of keyframes rises.
But, impressively, Implicit Warping can recreate video with as little as two frames, depending on the motion in question.



The video recreation in the far right column uses only the two frames depicted in the first column, while Implicit Warping derives the entire motion from a combination of the source clip and the face information in the frames.

If something abrupt should occur in the middle of the clip, such as an event or expression that is not represented either in the starting or end frame, Implicit Warping can *add a frame at that point*, and the added information will feed into the clip-wide attention mechanisms for the entire clip.

This kind of keyframed approach is currently being pursued both by amateurs and professional developers interested in expanding the video potential of the Stable Diffusion text-to-image synthesis system, and many (including myself – scroll down at this link) have experimented with using the non-AI tweening software EbSynth to create deepfake puppetry for complex and changing motion, by adding multiple Stable Diffusion renders to a video sequence powered by a real person.

Examples of Stable Diffusion output, animated via EbSynth, from users at Reddit.*

The power and potential of Implicit Warping notably outstrips not only prior works, but also EbSynth itself, which was not designed for this task, and, arguably, is difficult to adapt to it.

It seems likely that the researchers have chosen not to demonstrate actual 'transformations' of this kind due to a growing timidity in the image synthesis research sector regarding techniques that could as easily be used for deepfaking as for their chosen purpose. Acknowledging this, the by-now standard deepfake disclaimer in the new paper steers the customary path between enthusiasm and caution:

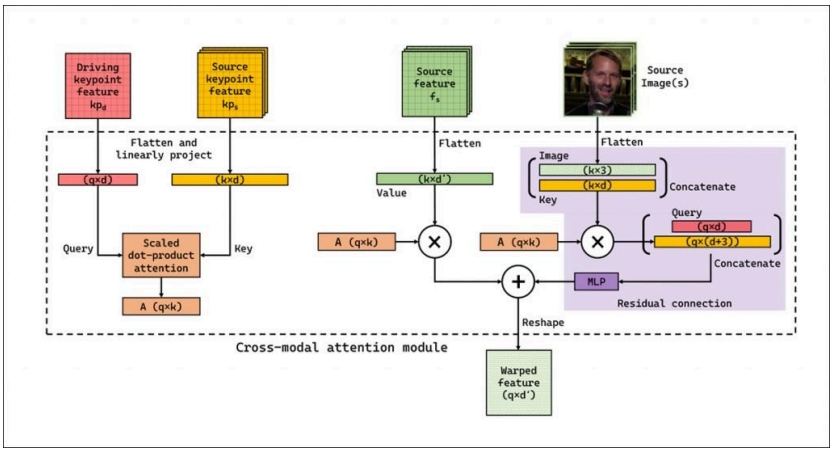
'Our method has the potential for negative impact if used to create deepfakes. Via the use of cross-identity transfer and speech synthesis, a malicious actor can create faked videos of a person, resulting in identity theft or dissemination of fake news. However, in controlled settings, the same technology can also be used for entertainment purposes.'

The paper also notes the potential of the system to power neural video reconstruction, in frameworks such as Google's Project Starline, where the work of recreating the correspondent appears to occur primarily on the client-side, using sparse motion information from the person at the other end. This schema is of growing interest to the research community, and is intended also to enable low bandwidth teleconferencing, by sending either pure motion data, or sparsely-intervalled keyframes that will be interpreted and interpolated into full, HD video on arrival.

Development and Method

Implicit Warping departs from prior approaches such as FOMM, Monkey-Net, and NVIDIA's own face-vid2vid, which use explicit warping to map out a temporal sequence into which information extracted from the source face and the controlling motion must be adapted, and to which it must conform. The final mapping of keypoints is fairly rigid under this regime.

By contrast, Implicit Warping uses a cross-modal attention layer that produces a workflow with less pre-defined bootstrapping, and which can adapt to input from multiple frames. Neither does the workflow require warping on a per-keypoint basis, which allows the system to select the most apposite features from a range of images.



The workflow for Implicit Warping.

Nonetheless, the new system repurposes the keypoint prediction components in the prior FOMM framework, ultimately encoding the derived spatial driving keypoint representations with a simple U-net. A separate U-net is used to encode the source image in tandem with the derived spatial representation, and both networks can operate at a range of resolutions at 64px (for 256px square output), to 384x384px.

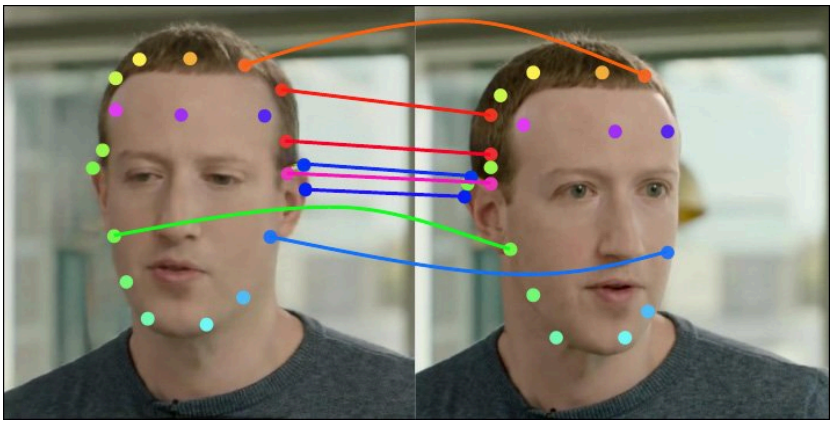
Because this 'mechanization' can't automatically account for all the possible variations of pose and movement in any given video, additional necessary, keyframes can be added *ad hoc*. Without this ability to intervene, keys with inadequate point-similarity to the target motion would automatically be uprated, lowering the quality of output.

The researchers explain:

'While it is the key most similar to the query in the given set of keys, it may not be similar enough to produce a good output. For example, suppose the source image has a face with lips closed, while the driving image has one with lips open and teeth exposed.

'In this case, there will be no key (and value) in the source image appropriate for the mouth region of the driving image. We overcome this issue by allowing our method to learn additional image-independent key-value pairs, which can be used in the case of missing information in the source image. These additional keys and values are concatenated to the keys and values obtained from the source image.'

Though the current implementation is quite fast, at around 10FPS on 512x512px images, the researchers believe that the pipeline can be optimized in future versions by a factored I-D attention layer, or a spatial-reduction attention (SRA) layer (i.e., a pyramid vision transformer).



Here Implicit Warping has derived a frontal image (left) from the genuine source image, with the mapped points indicated in various colors. Because Implicit Warping uses global rather than local attention, it can anticipate factors that previous efforts cannot, such as objects that are about to become dis-occluded.

Data and Tests

The researchers tested the system on the VoxCeleb2 dataset, the more challenging TED Talk dataset, and, for ablation studies, the TalkingHead-1KH set, comparing baselines between 256x256px and the full 512x512px resolution. Metrics used were Frechet Inception Distance (FID), LPIPS over AlexNet, and peak signal-to-noise ratio (PSNR).

Competing frameworks used for the tests were FOMM and face-vid2vid, in addition to AA-PCA. Since prior methods had little or no capacity to use multiple keyframes, which is the primary innovation of Implicit Warping, the researchers devised like-for-like testing methodologies.

First, Implicit Warping was tested on the 'home ground' of the former methods – as a way to derive motion from a single keyframe.

	TalkingHead-1KH [41] 512 × 512 faces			VoxCeleb2 [26] 256 × 256 faces			TED Talk [28] 384 × 384 upper bodies		
	FOMM [26]	fv2v [41]	Ours	FOMM [26]	fv2v [41]	Ours	FOMM [26]	AA-PCA [28]	Ours
PSNR ↑	23.23	23.27	**23.32**	**24.15**	23.66	23.54	24.17	25.14	**25.24**
$\mathcal{L}_1$ ↓	12.86	**12.30**	12.86	10.99	**11.10**	11.40	8.22	**6.87**	7.69
LPIPS ↓	0.16	0.16	**0.15**	0.12	0.12	**0.11**	0.16	0.13	**0.12**
AKD (MTCNN) ↓	4.25	3.83	**3.48**	**1.69**	1.70	1.75	–	–	–
(AKD, MKR)@0.1 ↓	–	–	–	–	–	–	(7.44, 0.01)	(5.10, 0.005)	**(4.39, 0.005)**
(AKD, MKR)@0.5 ↓	–	–	–	–	–	–	(5.93, 0.06)	(4.13, 0.030)	**(3.31, 0.023)**
FID ↓	18.15	18.09	**16.44**	9.53	5.76	**4.68**	24.69	19.17	**18.27**

Results for motion transfer from a single frame, against networks specifically designed for this task. For up arrows, larger is better; down arrows, smaller is better.

Here Implicit Warping outperforms most of the competing methods across most of the metrics, but is not allowed to get to fifth gear, and loses some position to architectures optimized to the task.

Next, the researchers tested for multiple-keyframe reconstruction, using sequences of at most 180 frames, and selected interstitial frames. Here Implicit Warping achieves a convincing overall victory:

TalkingHead1KH (512 × 512 faces)

#Source frames	1			2			3		
	FOMM [26]	fv2v [41]	Ours	FOMM [26]	fv2v [41]	Ours	FOMM [26]	fv2v [41]	Ours
PSNR ↑	24.260	24.273	**24.276**	25.433	25.557	**26.094**	26.044	25.843	**26.768**
$\mathcal{L}_1$ ↓	11.075	**10.338**	11.152	9.023	8.292	**7.990**	8.238	8.330	**7.347**
LPIPS ↓	0.139	0.135	**0.130**	0.116	0.110	**0.089**	0.109	0.118	**0.081**
AKD (MTCNN) ↓	5.039	4.792	**4.370**	4.662	4.384	**4.029**	4.483	6.201	**3.957**
FID ↓	17.454	17.174	**15.658**	19.647	16.846	**9.850**	19.495	19.216	**9.097**

TED Talk (384 × 384 upper bodies)

#Source frames	1		2		3	
	FOMM [26]	Ours	FOMM [26]	Ours	FOMM [26]	Ours
PSNR ↑	24.096	**25.188**	24.546	**26.528**	24.835	**26.884**
$\mathcal{L}_1$ ↓	8.290	**7.737**	7.726	**6.597**	7.398	**6.319**
LPIPS ↓	0.156	**0.119**	0.143	**0.088**	0.137	**0.081**
(AKD, MKR)@0.1 ↓	(7.618, 0.010)	(**4.438, 0.005**)	(6.898, 0.010)	(**4.182, 0.004**)	(6.518 0.010)	(**3.989, 0.004**)
(AKD, MKR)@0.5 ↓	(6.041, 0.060)	(**3.349, 0.025**)	(5.325, 0.067)	(**3.264, 0.021**)	(5.064, 0.067)	(**3.173, 0.021**)
FID ↓	24.324	**17.746**	29.053	**13.822**	30.838	**13.128**

The researchers note:
'As the number of source images increases, our method obtains better reconstructions as indicated by the improving scores on all metrics. However, reconstructions by prior work get worse as the number of source images increases, contrary to expectation.'

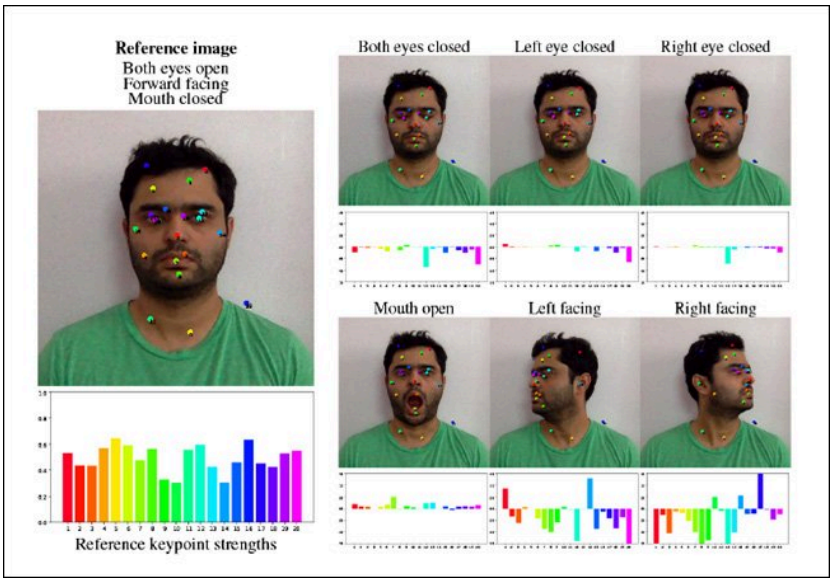
The challenge here is to recreate the 'driving' image (second from left) using only the information from the source image (far left). The third and fourth pictures show how previous frameworks compromised either on detail or positioning (or both), while Implicit Warping, far right, has successfully recreated the frame.

The system is not infallible: in the case of very extreme angles of a head, and where no keyframe offers a more confrontational pose, Implicit Warping has difficulty interpreting a view; however, as we have noted elsewhere, without the relevant data, this is practically an impossible task for any framework.
The results were presented also in a qualitative survey to Amazon Mechanical Turk (AMT) workers, who rated Implicit Warping's results above the rival methods.

TalkingHead-1KH	vs FOMM	vs fv2v
1 source	64.1%	53.5%
2 sources	65.8%	59.9%
TED Talk	**vs FOMM**	**vs AA-PCA**
1 source	81.5%	63.3%
2 sources	91.4%	–

Results from the qualitative user study.

Each worker was shown a pair of videos from a total of 360 and 128 clips, from TalkingHead-1KH and Ted Talk, respectively.



The user interface presented to the AMT workers for the qualitative study.

What Could We Do With Implicit Warping?

Given access to this kind of framework, users would be able to produce far more coherent and longer video simulations and full-body deepfake footage, all of which could feature a far greater range of motion than any of the frameworks that the system has been trialed against.

The value of this kind of approach is in the extent to which a keyframe may be *difficult to produce* – not something which the paper addresses, since the authors choose only to recreate existing footage.

Reproducing some kind of extraordinary figure in an adequate number of poses to populate the keyframe necessary for complex motion may involve, for instance, the elaborate construction of CGI models; else the training (brief though it is) of temporally consistent DreamBooth models for Stable Diffusion, capable of depicting a character in different poses without any other physical changes in their appearance (which is otherwise a challenge in latent diffusion models, which may give you something 'a little different' every time).

Thus, a system such as Implicit Warping could enable simulated or deepfaked clips of a length and variability that no other interpretive framework has yet offered. As extraordinary actions occur in the driving source clip, additional keyframes could cover that extra data as necessary, without either needing continuous and contiguous rendering, or to hope that a single frame of data might be enough to populate the clip.

It's a potentially powerful animation tool, apparently masquerading as yet another constrained 'talking head' generator, and is perhaps being downplayed by NVIDIA for 'optical' rather than practical reasons.

* *Sources: https://old.reddit.com/r/StableDiffusion/comments/x8gdtu/stable_difussion_img2img_ebsynth_is_a_very/*

https://old.reddit.com/r/StableDiffusion/comments/xn2p5s/playing_with_ebsynth_stable_diffusion/

https://old.reddit.com/r/StableDiffusion/comments/xka1do/i_used_sd_and_ebsynth_for_some_horror_makeup_vfx/