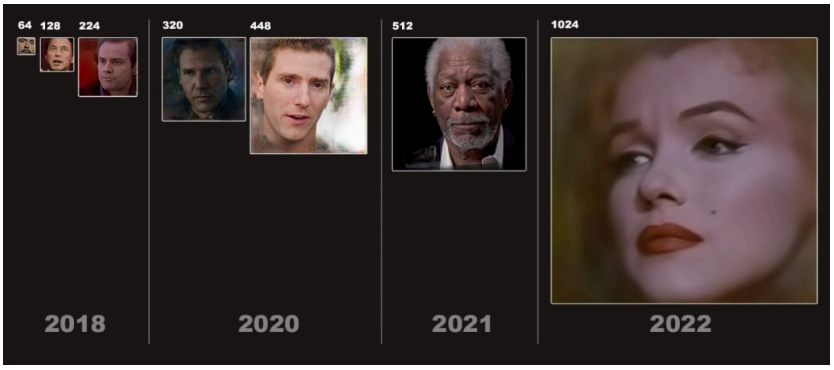
Deepfakes Go High-Res – But Can Deepfakers Handle It?

Originally published at https://metaphysic.ai/deepfakes-go-high-res-deepfakers-handle/

July 11, 2022

Towards the end of our new feature on the future of autoencoder-based deepfakes, we preview a new capability in FaceSwap, set for release this Saturday, that's potentially capable of creating the highest-resolution deepfakes ever, with the model processing and producing faces at native 1024px resolution, and able to train on a card as humble as a GTX 1080 with 8GB of VRAM (if you have a *lot* of patience).
To put this into perspective, some of the highest-resolution deepfakes currently being undertaken are in the 384/448px range, requiring very well-specced GPUs and formidable training times, undertaken on the sparsest settings.
Training deepfake models at input/output 1024px effectively doubles even the most 'experimental' native resolutions being tested in the deepfake community at this time.

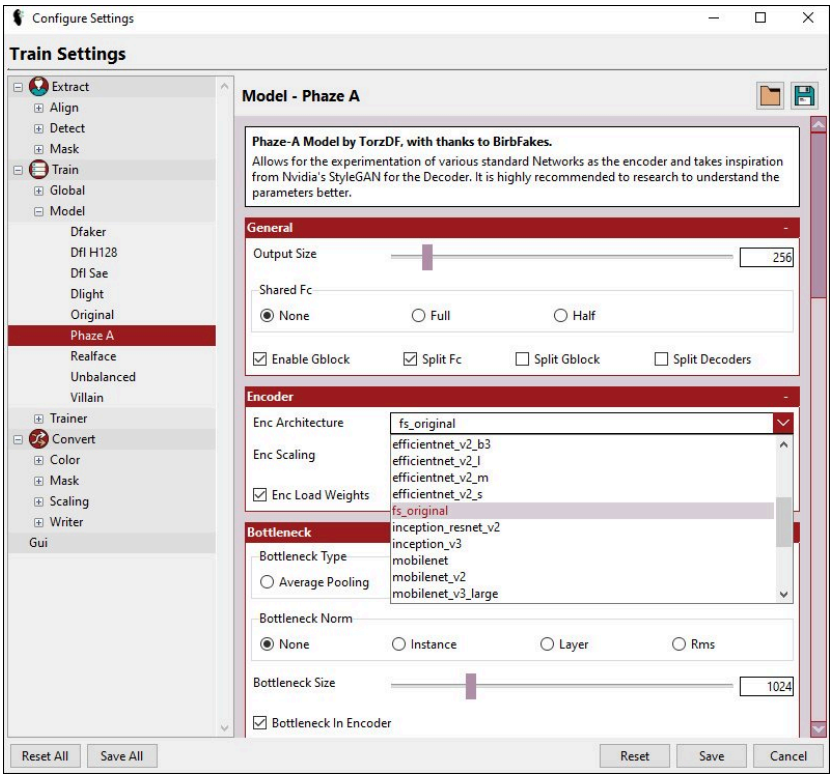


DeepFaceLab's table of increase in training image size, compared to 1024px native training resolution. These are broadly representative input/output sizes in the five years since deepfakes emerged, though better-specced deepfakers also use 384, 768, and various other resolutions.

Currently the input/output resolution range for DeepFaceLab maxes out at 640px. As one prominent DeepFaceLive proponent points out, that setting is likely to require 24GB of VRAM on the best NVIDIA consumer card available.

Phaze-A 1024px

The new 1024px FaceSwap capability is not a dedicated model in itself, but a setting for FaceSwap's innovative Phaze-A framework, which allows the user to develop highly customized deepfake neural networks, and is capable of utilizing any of nine popular and powerful open source encoder architectures.



Only part of the extensive settings for Phaze-A (see the extra space in the scrollbar!), which allows the user to create custom neural architectures. Phaze-A offers so many possibilities that it now comes with a selection of loadable presets for users to employ as a starting point for new models.

Phaze-A's complexity and configurability has led to the need to create loadable presets, the most demanding of which, to date, has been *StoJo* (named after the two personalities used as default testing identities), which is difficult to run on less than 11GB of VRAM.
Despite its ability to pass unprecedented image-sizes through a consumer-level GPU for deepfake training, the new Phaze-A 1024setting is far more tolerant. FaceSwap developer Matt Tora tells us that he has gotten the Phaze-A 1024 preset (not its official name – it currently has none) training on the 8GB of VRAM available in the venerable GTX 1080, under Linux (which does not steal VRAM for system usage, unlike Windows).

Training Marilyn

Little is yet known of the capabilities of Phaze-A 1024. The most extensive training anyone has done with the model, which was made available early to Patreon consumers, has been by notable deepfaker Deep Homage.
Deep Homage trained a 1024-native model of actresses Theresa Russell (playing a tacit incarnation of Marilyn Monroe in the 1985 Nick Roeg outing *Insignificance*) and Marilyn Monroe to just 21,000 iterations.

<YOUTUBE LINK OMITTED>

It's worth considering that 21,000 iterations is considered to be a *very* early phase in training session. A typical model is not likely to reach convergence (i.e. to become effective and convincing) until it has reached 500k-1 million iterations, while many of the best models require two million iterations to resolve the most acute detail.
Deep Homage comments 'I hesitate to send these 1024 training previews, because the model has only been trained for 21 thousand iterations at batch size 8. The model will need at least a million iterations or more to really show its full potential.'
The 1024 Russell/Monroe model was trained on a NVIDIA RTX A6000 with 48GB VRAM, with an average selling price of around $5000 USD.
The full-size, non-masked PNG preview panel provided by Deep Homage weighs in at 40mb, with even a JPEG version reaching an ungainly 12MB. Since that's too large to host, below is a full-size version of the previews with the real source training images removed (5.62MB, click to enlarge).

Training previews for the Russell/Monroe session. Images courtesy of Deep Homage.

There are no 'real' images here – all the images of Russell and Monroe are the latest attempted recreations from the emerging model. Odd-numbered columns left-to-right are direct recreations of the real training photos, while even-numbered columns left-to-right are interpretations of that same pose with the target identity.
The most common poses (such as front views) typically form better, faster, and earlier, due to being well-represented in the datasets. Acute angles and profile views will be among the last to resolve well, far later in training, while details such as eye-glint and teeth are not likely to become clear until at least mid-way through training.
The content depicted is dumped from the previews panel in FaceSwap, which refreshes to show the model's progress after each save iteration is completed.

Getting to 1024

Phaze-A 1024 is, to the best of our knowledge, the first native 1024 input/output model to be made available in an open source deepfakes package. Though the rumor mill suggests that a number of VFX studios are developing off-shoots of 2017-era autoencoder architectures that are capable of 1024, the only relatively hard evidence to date is from PAGI studios in Sacramento, CA, where former security developer Dogan Kurt has developed a proprietary framework that builds on existing public code to enable impressive native 1024 resolution deepfakes, exhibiting remarkable consistency and authenticity:

https://www.youtube.com/watch?v=UuK9tjEQ2EU

(see our full-length feature for our chat with Dogan about this method, and for further details).
However, as best Dogan's pipeline can be understood, the resulting model is very lighting-specific and locked to the source material, with limited capacity to adapt to a wide range of footage.
By contrast, a fully-trained Phaze-A 1024 model could, like any other typical FaceSwap or DeepFaceLab model, adapt well to depicting the target personality across a wide range of video clips.
Matt Tora is not disclosing all the details of the new architecture, but has commented to us:
'The 1024 preset is a symmetrical encoder/decoder network that trades off filter count for dimensional space. It also replaces the fully-connected layers with convolutions to reduce the memory overhead at the center of the model.
'This helps to create a model that can train to high resolutions with a relatively small memory footprint. Whilst it is possible, within the correct circumstances, to train this model on an 8GB card at very low batch-sizes, it is recommended to use a GPU with more than 8GB.'

Standing on the Shoulders of Giants

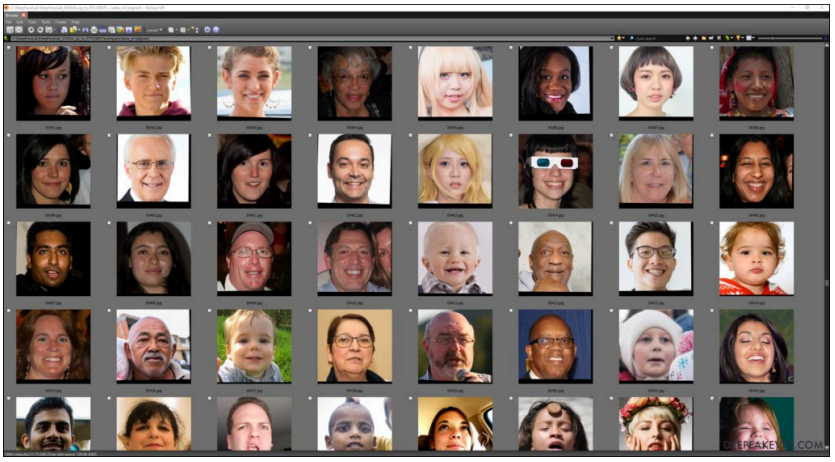
Computing autoencoder loss values for such a high volume of large source images is a staggering task. Though the forthcoming 40XX series of NVIDIA graphics cards are set to improve training performance to an extent, they're not likely to make much of a dent in the training times that will be needed for 1024px native workflows.
It seems at the moment that the only realistic route to rational 1024px training, even for well-specced VFX studios, would be the re-use of semi-trained models, or else the loading of weights from fully-trained models, either of which can 'kick start' a model and notably accelerate training through Transfer Learning.
Here, the DeepFaceLab and FaceSwap projects diverge a little, in terms of common practices. The DFL community is broadly committed to saving on training times by re-using semi-trained models that have already calculated most of the overarching identity features.
Pretrained models that have assimilated thousands of identities are used to pre-inject into a model a high level of coverage for the highest possible number of feature relationships between two identities.

https://www.youtube.com/watch?v=Uicc8ioJiog

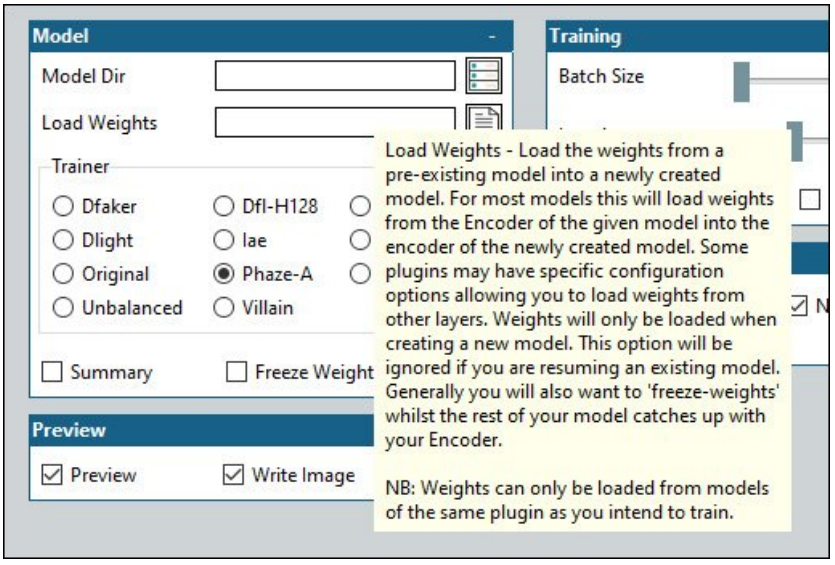
Users then share and download these pre-trained models, replacing the source and target identity with the two identities that they want to swap. So long as the pre-trained model is still flexible enough to adapt to the new data, it's possible to bypass a massive stretch of the training process.

https://www.youtube.com/watch?v=N6XN0fjHZSA

DeepFaceLab now incorporates NVIDIA Labs' Flicker Faces HQ (FFHQ) open-source dataset as a native pretrained resource, while previous versions used the popular CelebA dataset.



By contrast, FaceSwap has developed a *Load Weights* feature, which imports into a new model the learned relationships from a previous, fully-trained model.



In FaceSwap, it's possible to import the weights of a previously-trained model to a new model. The 'new' identity's weights will be frozen (see the 'Freeze Weights' option in image above) until it catches up with the previously trained data.

Matt Tora explains:
'Basically, you're importing well-refined encodings for faces. Whilst specific to your first face-pairings, most of that encoding information will be useful for any *face pairing.*
'The main difference is that DeepFaceLab is training on many identities, while FaceSwap is re-using weights that have been trained on just two identities, which may help to obtain a more accurate resemblance.
'Due to the large training times involved, it's difficult to conduct the lengthy testing that could really establish that one method is superior to the other, so that remains an open question.
FaceSwap co-developer Bryan Lyon adds 'Basically, the Load Weights feature loads the encoder only. The Encoder in an ideal world will not contain *any* identity data, and should work for any given faces'
Thus, the future paradigm for 1024 training seems likely to be that a 'generic' high-resolution model will be trained (probably for months) on a highly capable GPU, which will then contribute its weights and learned knowledge to users with lesser hardware, either through being used as a pre-trained model, or else 'donating' those weights that were so incredibly expensive and time-consuming to learn.
In this sense, deepfaking may eventually become more like hyperscale Natural Language Processing (NLP) projects such as GPT-3, and massively expensive image synthesis frameworks like DALL-E 2, and Google's Imagen – a scenario where the software itself is trivial, and where expensive, industrially-trained weights represent the major value proposition.

Other Barriers to Entry

So there are several primary reasons why Phaze-A 1024 (or any similarly functional deepfake model) is not a 'plug-and-play' solution to higher-res deepfake output for the casual hobbyist.
Firstly, though training the 1024 setting on a low-end card is technically possible, it can only be done on Linux (due to the Windows 10 VRAM-appropriation), and only at a batch size of 2.

A batch size that low can be very useful for obtaining better detail in the later stages of training (combined with a lower learning rate), but can make the model rather 'myopic' in the earliest stages, and impede generalization. Anyone using the model on a higher-end card will be able to start at a more sensible setting (such as batch size 8), and ramp down in the usual way, likely obtaining a superior overall resemblance, and better detail.
Secondly, the training times involved for budget GPUs are likely to be an insuperable barrier to the use of this new preset. Deep Homage himself had to stop his Phaze-A 1024 experiment at 21,000 iterations, after the first three days of training, due to pressure of work – and that was on the mighty A6000. For a typical 8-11GB card, it isn't unreasonable to expect many months of training time in order to arrive at a usable 800k-1.5m iterations, in the same scenario.
However, once a culture of weight-loading or sharing of pre-trained models is established (see 'Standing on the shoulders of giants', above), it will no longer be necessary to train a 1024 model from zero, making high-resolution deepfaking a little more attainable – though still effectively impractical for low-end setups.
Finally, the customary difficulty in obtaining adequately-sized face images for deepfake training sets becomes critical when considering a 1024 pipeline.
In order to get images that do not need to be upscaled inside the model architecture (which would affect the quality of the output), it's necessary to find source face images that are not only high quality, but uncommonly high-resolution.
Matt Tora says:
'Basically, to get the best out of a 1024px model, you are going to want faces extracted that are, at a minimum, 1024px across each size. On 1080p footage, that is nigh-on impossible except in extreme close-ups.'
Consequently, viral deepfakers looking to recast Hollywood movies by inserting alternate actors into movie clips would need to extract source material from 4K sources. Many of those sources are likely to have been encoded with High Dynamic Range (HDR), which makes extracting faces from the source a notable challenge.
Tora says 'There are techniques that attempt to map HDR footage to LDR, but these could best be described as hit-and-miss'.
These barriers to entry could limit the use of Phaze-A 1024 (and probably any other similar framework) to actual VFX professionals, who will have the necessary resources to curate high-resolution source material, and to train models in a reasonable time-frame; or at the very least, to more casual deepfakers who are generating their own high-resolution source material, rather than ripping it from copyrighted sources.
Tora comments:
'The main problem with 1024px is that your average user just does not have the time to either find useful data for it, nor the time to train it.'
The new Phaze-A 1024 preset is set to enter the main branch of FaceSwap on Saturday 16th July.