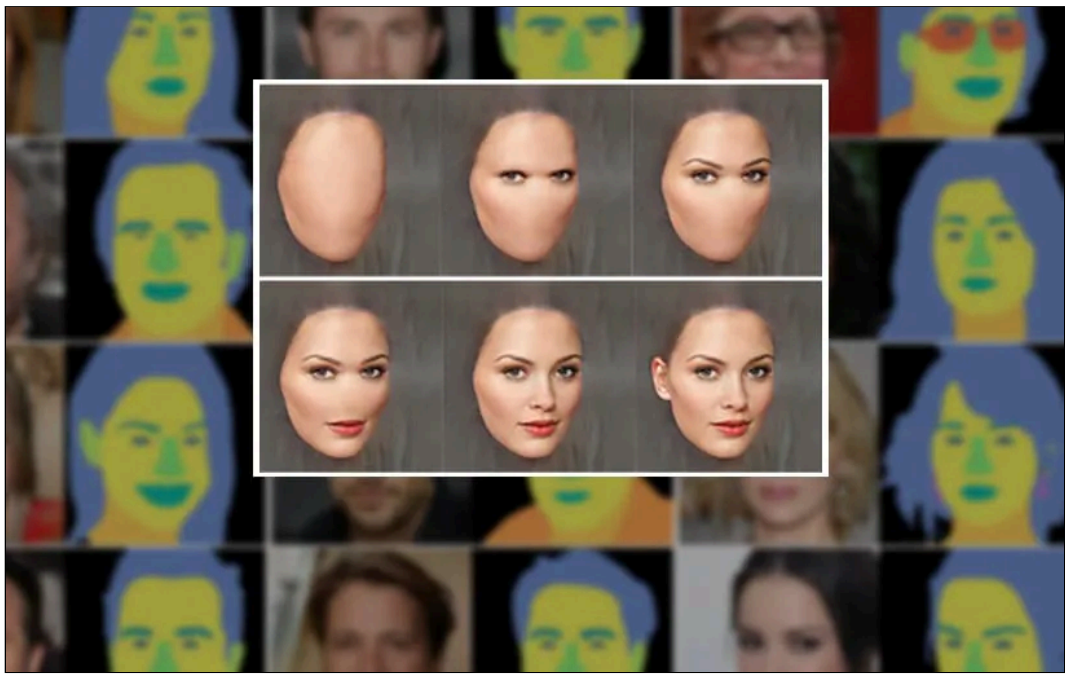
Orchestrating Facial Synthesis With Semantic Segmentation

Originally published at https://www.unite.ai/orchestrating-facial-synthesis-with-semantic-segmentation/

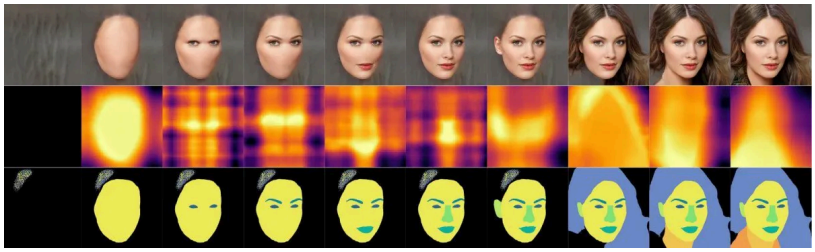


The problem with inventing human faces with a Generative Adversarial Network (GAN) is that the real-world data that fuels the fake images comes with unwelcome and inseparable accoutrements, such as hair on the head (and/or face), backgrounds, and various kinds of face furniture, such as glasses, hats, and ear-rings; and that these peripheral aspects of personality inevitably become bound up in a 'fused' identity.

Under the most common GAN architectures, these elements aren't addressable in their own dedicated space, but rather are quite tightly associated with the face in (or around) which they're embedded.

Neither is it usually possible to dictate or affect the appearance of *sub-sections* of a face created by a GAN, such as narrowing the eyes, lengthening the nose, or changing hair-color in the way that a police sketch artist could.

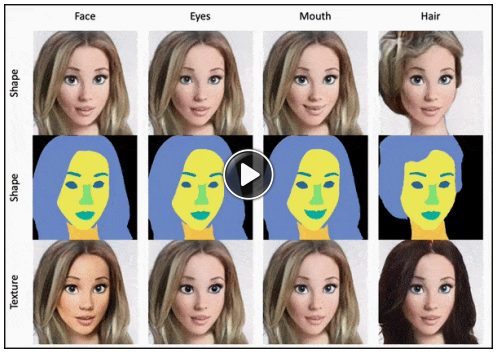
However, the image synthesis research sector is working on it:



New research into GAN-based face generation has separated the various sections of a face into distinct areas, each with their own 'generator', working in concert image. In the middle row, we see the orchestrating 'feature map' building up additional areas of the face. Source: https://arxiv.org/pdf/2112.02236.pdf

In a new paper, researchers from the US arm of Chinese multinational tech giant ByteDance have used semantic segmentation to break up the constituent parts of the face into discrete sections, each of which is allocated its own generator, so that it's possible to achieve a greater degree of disentanglement. Or, at least, *perceptual* disentanglement.

The paper is titled *SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing*, and is accompanied by a media-rich project page featuring multiple examples of the various fine-grained transformations that can be achieved when facial and head elements are isolated in this way.



CLICK TO VIEW ANIMATED GIF

Facial texture, hair style and color, eye shape and color, and many other aspects of once-indissoluble GAN-generated features can now be de facto *disentangled, though the quality of separation and level of instrumentality is likely to vary across cases.* Source: https://semanticstylegan.github.io/

The Ungovernable Latent Space

A Generative Adversarial Network trained to generate faces – such as the StyleGan2 generator that powers the popular website thispersondoesnotexist.com – forms complex interrelationships between the 'features' (not in the facial sense) that it derives from analyzing thousands of real-world faces, in order to learn how to make realistic human faces itself.

These clandestine processes are 'latent codes', collectively the *latent space*. They are difficult to analyze, and consequently difficult to instrumentalize.

Last week a different new image synthesis project emerged that attempts to 'map' this near-occult space during the training process itself, and then to use those maps to interactively navigate it, and various other solutions have been proposed to gain deeper control of GAN-synthesized content.

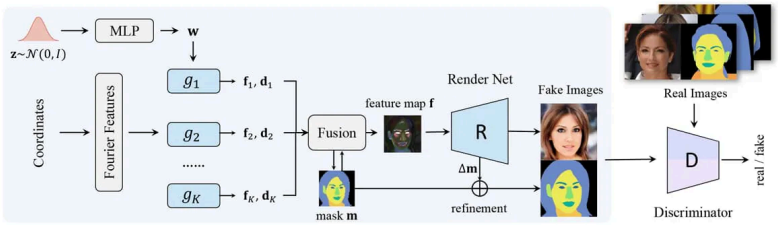
Some progress has been made, with a diverse offering of GAN architectures that attempt to 'reach into' the latent space in some way and control the facial generations from there. Such efforts include InterFaceGAN, StyleFlow, GANSpace, and StyleRig, among other offerings in a constantly-productive stream of new papers.

What they all have in common is limited degrees of disentanglement; the ingenious GUI sliders for various facets (such as 'hair' or 'expression') tend to drag the background and/or other elements into the transformation process, and none of them (including the paper discussed here) have solved the problem of temporal neural hair.

Dividing and Conquering the Latent Space

In any case, the ByteDance research takes a different approach: instead of trying to discern the mysteries of a single GAN operating over an entire generated face image, SemanticStyleGAN formulates a layout-based approach, where faces are 'composed' by separate generator processes.

In order to achieve this distinction of (facial) features, SemanticStyleGAN uses Fourier Features to generate a semantic segmentation map (crudely colored distinctions of facial topography, shown towards the lower-right of the image below) to isolate the facial areas which will receive individual, dedicated attention.



Architecture of the new approach, which imposes an interstitial layer of semantic segmentation onto the face, effectively turning the framework into an orchestrato different facets of an image.

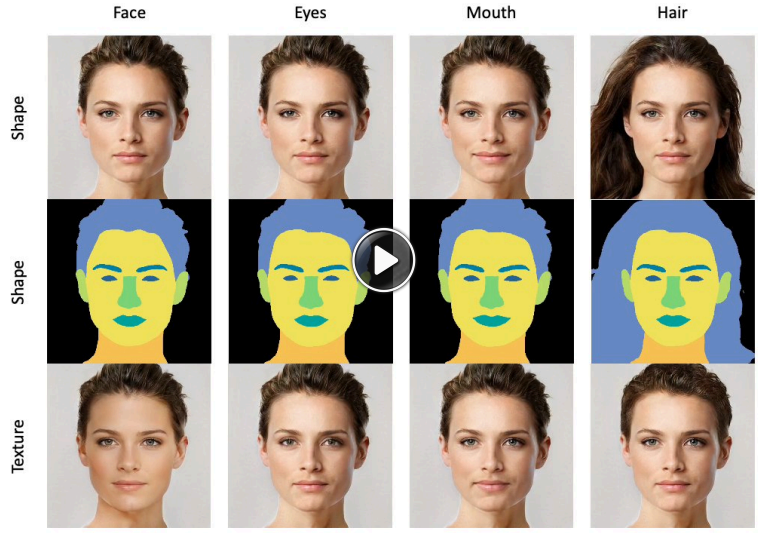
The segmentation maps are generated for the fake images that are systematically presented to the GAN's discriminator for evaluation as the model improves, and to the (non-fake) source images used for training.

At the start of the process, a Multi-Layer Perceptron (MLP) initially maps randomly-chosen latent codes, which will then be used to control the weights of the several generators that will each take control of a section of the face image to be produced.

Each generator creates a feature map and a simulated depth-map from the Fourier features that are fed to it upstream. This output is the basis for the segmentation masks.

The downstream render network is only conditioned by the earlier feature maps, and now knows how to generate a higher-resolution segmentation mask, facilitating the final production of the image.

Finally, a bifurcated discriminator oversees the concatenated distribution of both the RGB images (which are, for us, the final result) and the segmentation masks that have allowed them to be separated.

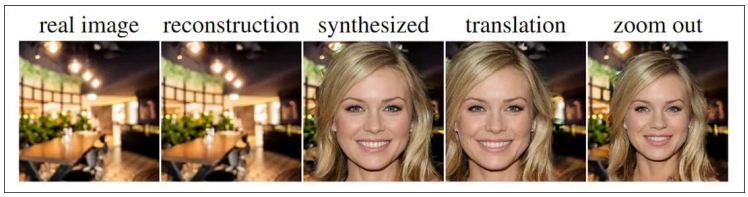


CLICK TO PLAY VIDEO ON SITE

With SemanticStyleGAN, there are no unwelcome visual perturbations when 'dialing in' facial feature changes, because each facial feature has been separately trained within the orchestration framework.

Substituting Backgrounds

Because the intention of the project is to gain greater control of the generated environment, the rendering/composition process includes a background generator trained on real images.



One compelling reason why the backgrounds do not get dragged into facial manipulations in SemanticStyleGAN is that they are sitting on a more distant layer, and partially hidden by the superimposed faces.

Since the segmentation maps will result in faces without backgrounds, these 'drop-in' backgrounds not only provide context, but are also configured to be apposite, in terms of lighting, for the superimposed faces.

Training and Data

The 'realistic' models were trained on the initial 28,000 images in CelebAMask-HQ, resized to 256×256 pixels to accommodate the training space (i.e. the available VRAM, which dictates a maximum batch size per iteration).

A number of models were trained, and diverse tools, datasets and architectures experimented with during the development process and various ablation tests. The project's largest productive model featured 512×512 resolution, trained over 2.5 days on eight NVIDIA Tesla V100 GPUs. After training, generation of a single image takes 0.137s on a lobe GPU without parallelization.

The more cartoon/anime-style experiments demonstrated in the many videos on the project's page (see link above) are derived from various popular face-based datasets, including Toonify, MetFaces, and Bitmoji.

A Stopgap Solution?

The authors contend that there is no reason why SemanticStyleGAN could not be applied to other domains, such as landscapes, cars, churches, and all the other 'default' test domains to which new architectures are routinely subjected early in their careers.

However, the paper concedes that as the number of classes rises for a domain (such as *'car'*, '*street-lamp'*, *'pedestrian'*, *'building'*, *'car'* etc.), this piecemeal approach might become unworkable in a number of ways, without further work on optimization. The CityScapes urban dataset, for instance, has 30 classes across 8 categories.

It's difficult to say if current interest in conquering the latent space more directly is as doomed as alchemy; or whether latent codes will eventually be decipherable and controllable – a development that could render this more 'externally complex' type of approach redundant.

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