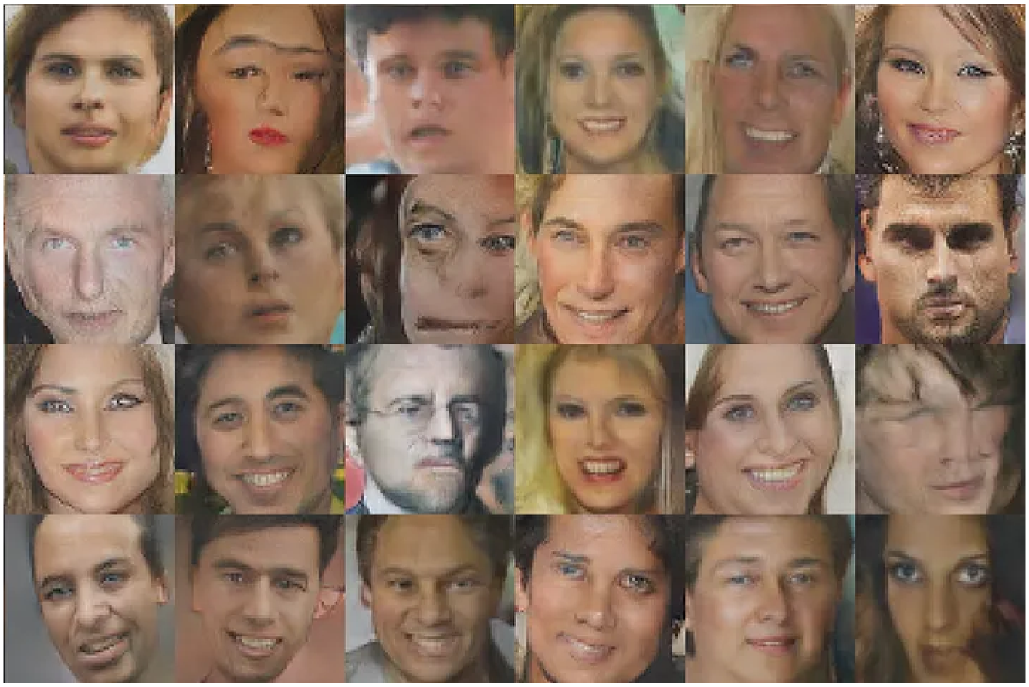
Kyle McDonald Follow
Dec 5, 2018 · 7 min read · Listen

How to recognize fake AI-generated images

In 2014 machine learning researcher Ian Goodfellow introduced the idea of generative adversarial networks or GANs. "Generative" because they output things like images rather than predictions about input (like "hotdog or not"); "adversarial networks" because they use two neural networks competing with each other in a "cat-and-mouse game", like a cashier and a counterfeiter: one trying to fool the other into thinking it can generate real examples, the other trying to distinguish real from fake.

The first GAN images were easy for humans to identify. Consider these faces from 2014.



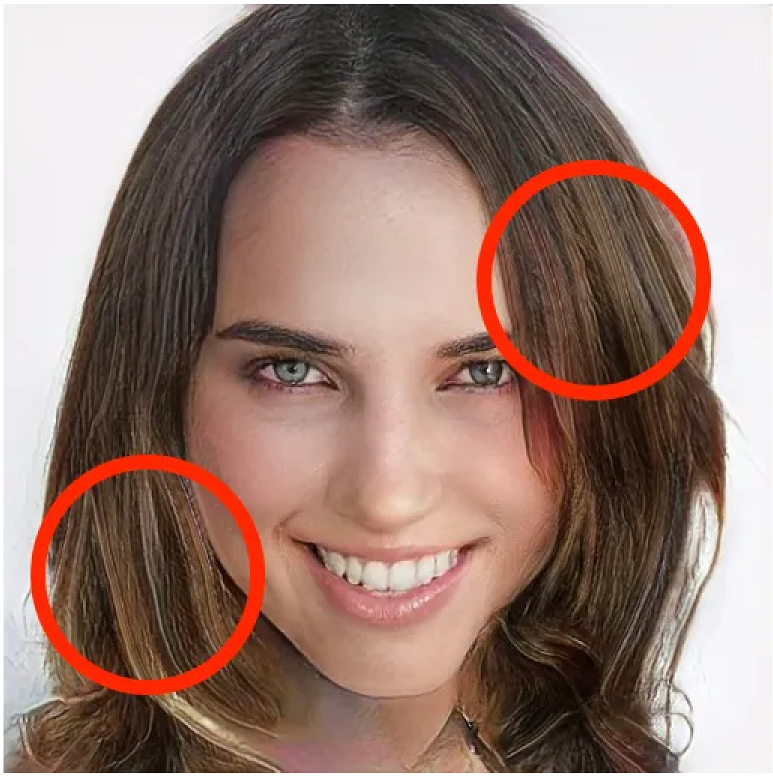
"Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks" (2014) by Radford et al, also known as DCGAN.

But the latest examples of GAN-generated faces, published in October 2017, are more difficult to identify.

“Progressive Growing of GANs for Improved Quality, Stability, and Variation” (2017) by Karras et al, also known as PGAN or ProGAN.

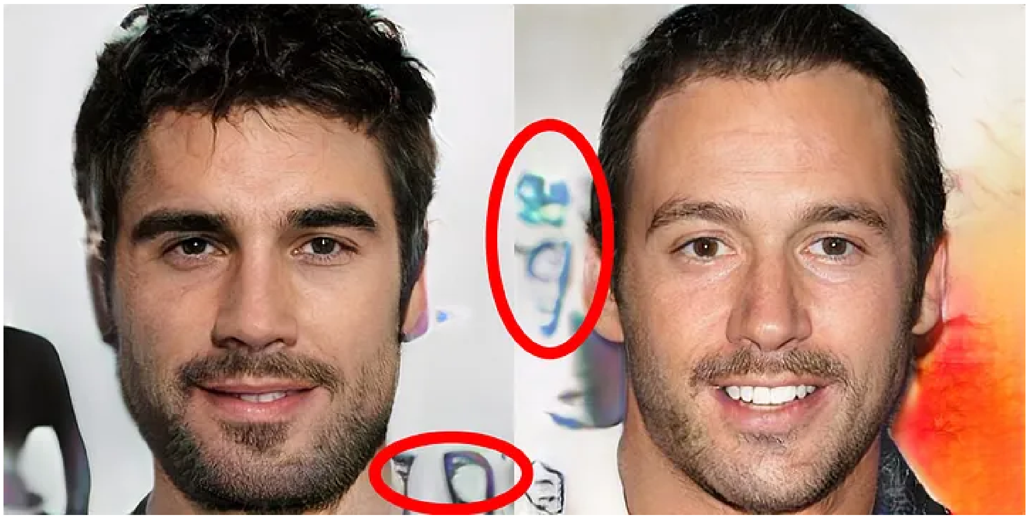
Here are some things you can look for when trying to recognize an image produced by a GAN. We’ll focus on faces because they are a common testing ground for researchers, and many of the artifacts most visible in faces also appear in other kinds of images.

Straight hair looks like paint



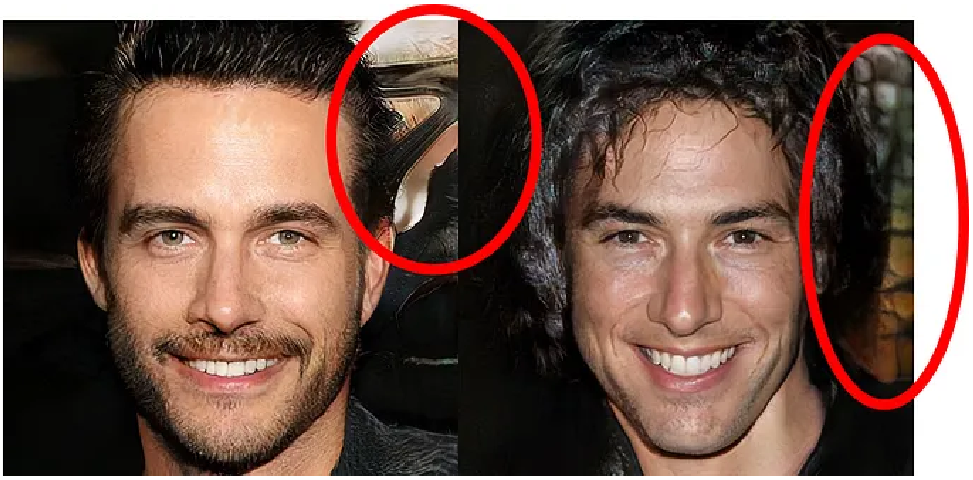
It’s common for long hair to take this hyper-straight look where a small patch seems good, but a long strand looks like someone smudged a bunch of acrylic with a palette knife or a huge brush.

Text is indecipherable



GANs trained on faces have a hard time capturing rare things in the background with lots of structure. Also, GANs are shown both original and mirrored versions of the training data, which means they have trouble modeling writing because it typically only appears in one orientation.

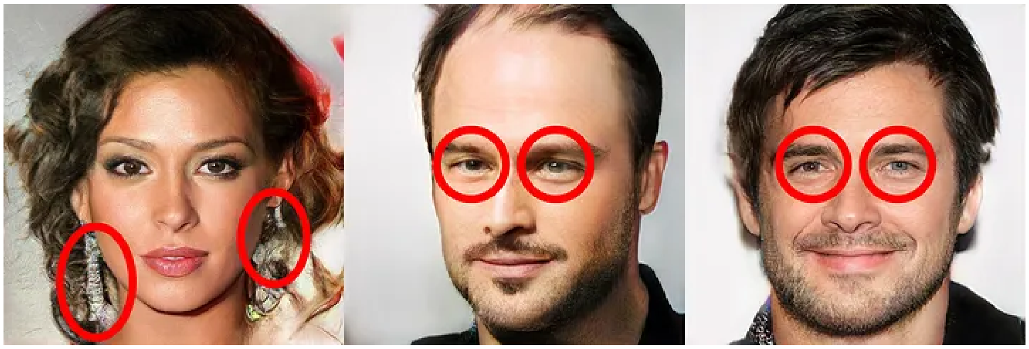
Background is surreal



One reason the faces from a GAN look believable is because all the training data has been centered. This means that there is less variability for the GAN to model when it comes to, for example, the placement and rendering of eyes and ears. The background, on the other hand, can contain anything.

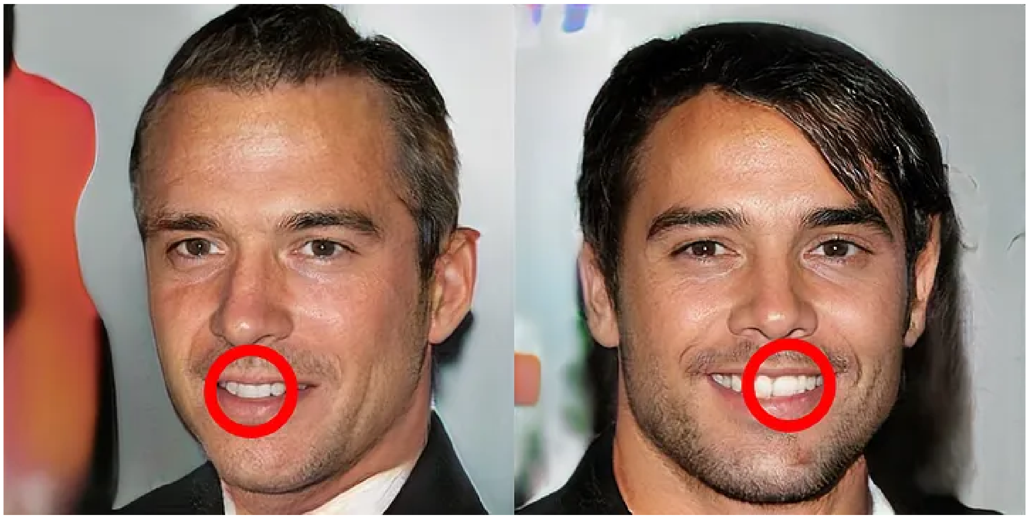
This is too much for the GAN to model and it ends up replicating general background-like-textures rather than "real" background scenes.

Asymmetry



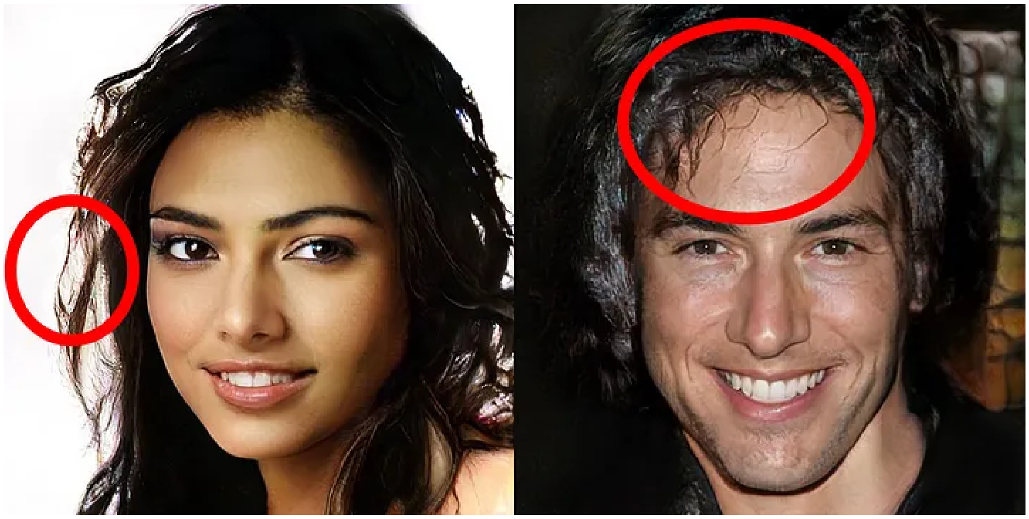
It can be difficult for a GAN to manage long-distance dependencies in images. While paired accessories like earrings usually match in the dataset, they don't in the generated images. Or: eyes tend to point in the same direction and they are usually the same color, but the generated images are very frequently crosseyed and heterochromatic. Asymmetry is also commonly visible in ears being at very mismatched heights or sizes.

Weird teeth



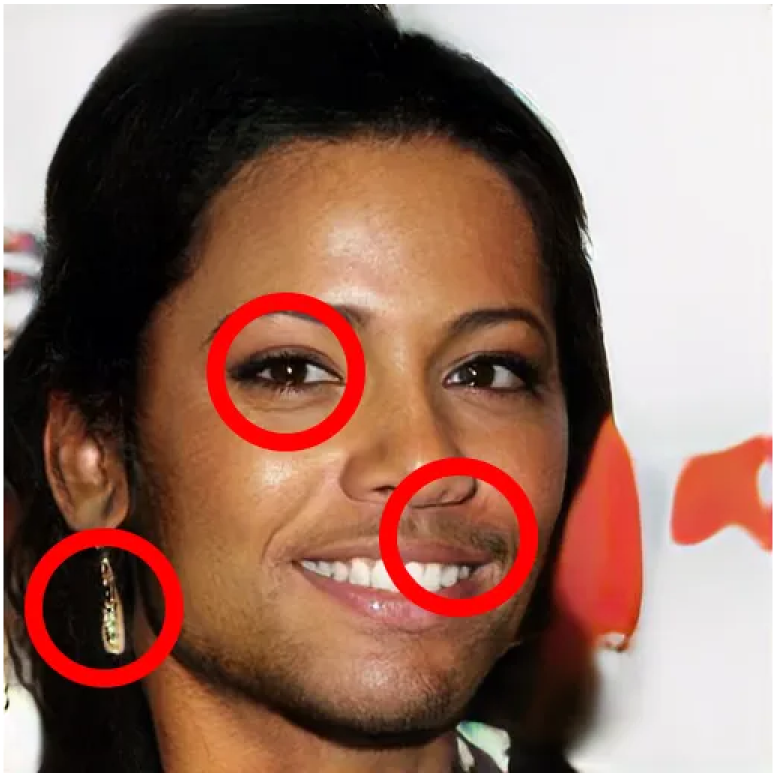
GANs can assemble a general scene, but currently have difficulty with semi-regular repeating details like teeth. Sometimes a GAN will generate misaligned teeth, or it will stretch or shrink each tooth in unusual ways. Historically this problem has shown up in other domains like texture synthesis with images like bricks.

Messy hair



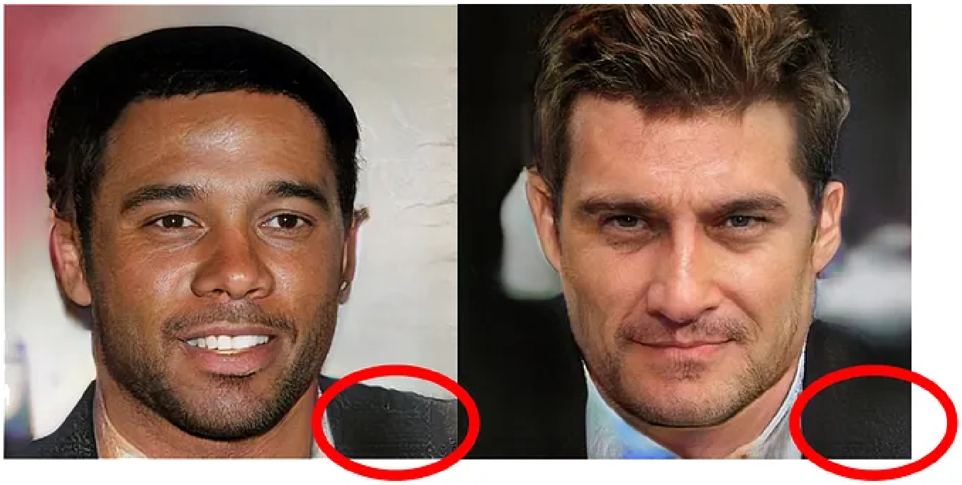
This is one of the quickest ways to identify a GAN-generated image. Typically a GAN will bunch hair in clumps, create random wisps around the shoulders, and throw thick stray hairs on foreheads. Hair styles have a lot of variability, but also a lot of detail, making it one of the most difficult things for a GAN to capture. Things that aren't hair can sometimes turn into hair-like textures, too.

Non-stereotypical gender presentation



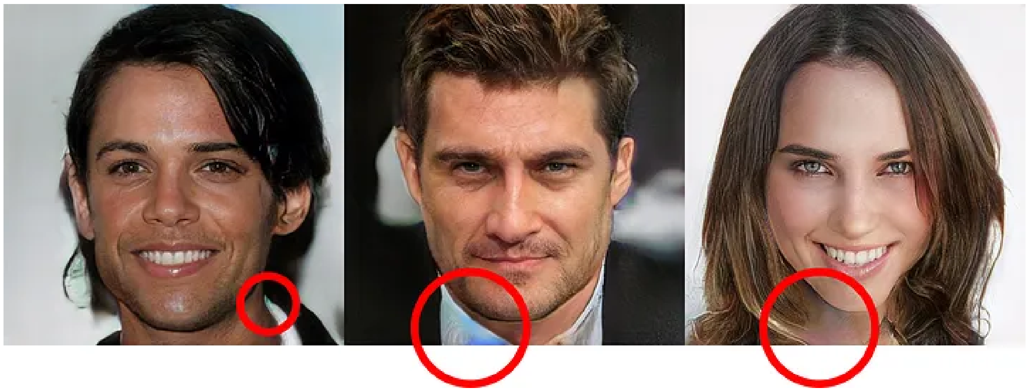
This GAN was trained on a subset of CelebA, which contains 200k images of 10k celebrity faces. In this dataset, I haven't seen an example of someone with facial hair, earrings, and makeup; but the GAN regularly mixes different attributes from stereotypical gender presentations. More generally, I think this is because GANs don't always learn the same categories or binaries that humans socially reinforce (in this case "male vs female"). It's important to be clear here: like asymmetry, non-stereotypical gender presentation is not inherently an indicator that an image isn't "real". Unlike messy hair, it is less of a visual artifact present in individual images, and more of a disparity in matching statistics across a large collection of images.

Semi-regular noise



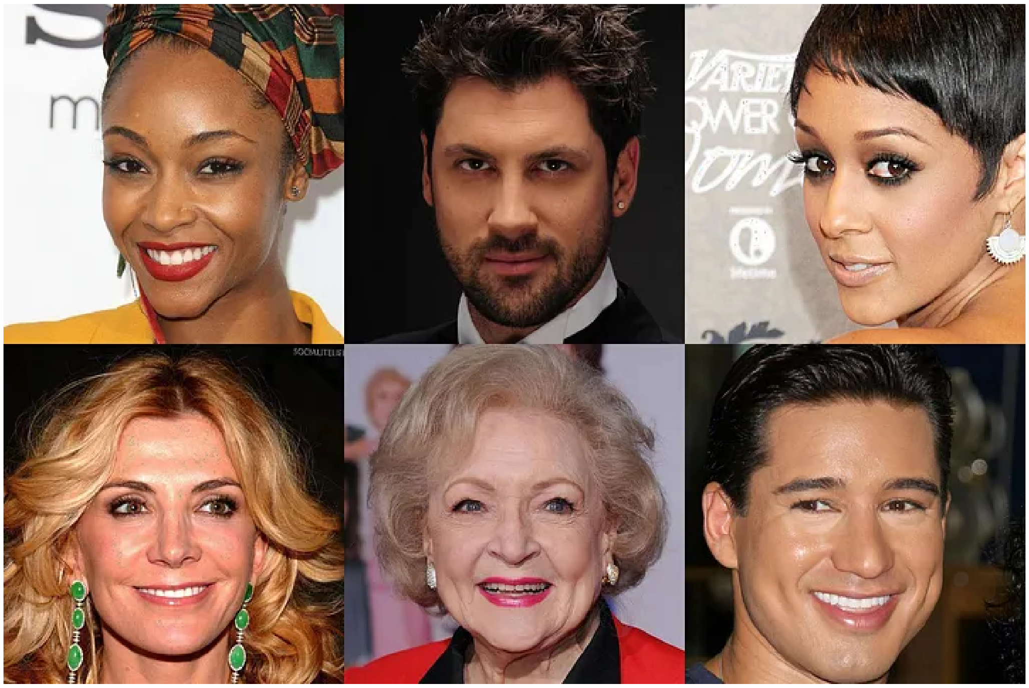
Some areas that are otherwise monochrome may exhibit semi-regular noise with horizontal or vertical banding. In the cases above, this is probably the network trying to imitate the texture of cloth. Older GANs have a much more prominent noise pattern that is usually described as checkerboard artifacts.

Iridescent color bleed



Some areas with lighter solid colors have a multi-hued cast, including collars, necks, and eye whites (not shown).

Examples of real images



Check out that clear background text, those matching earrings, those equally sized teeth, detailed hairstyles. With all these tricks in mind, try playing this game that tests your ability to distinguish real from fake and see how many you get right. *Note: some people have had problems clicking "start".*

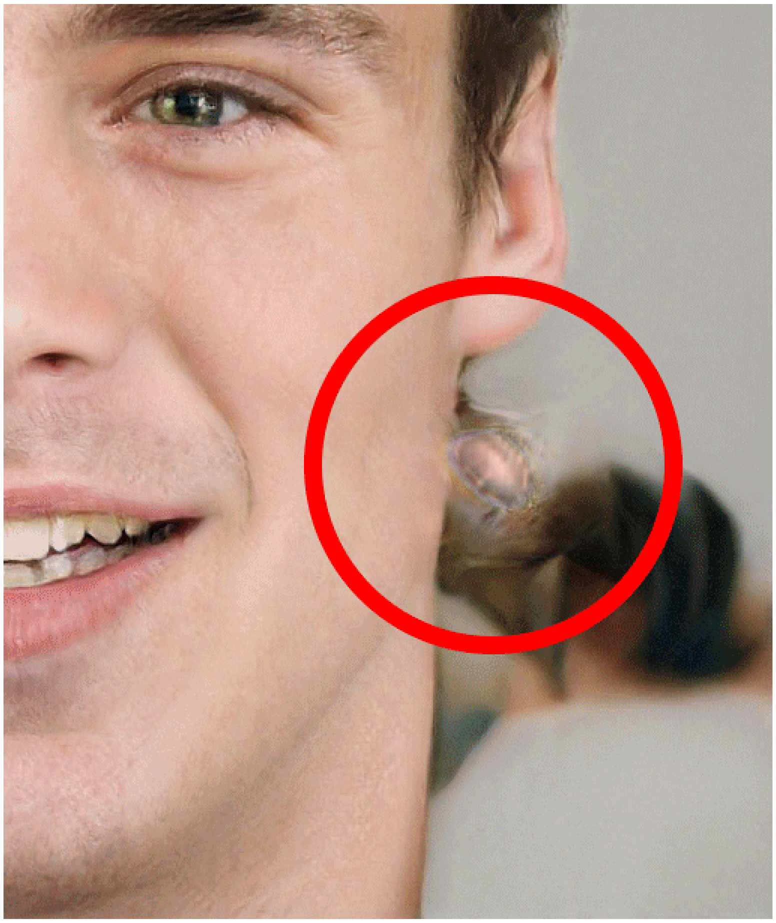
Update (December 13, 2018)

One year after "Progressive Growing of GANs" which produce the above images, the same researchers have published "A Style-Based Generator Architecture for GANs". Check out the video. This new work improves on many of the issues above.

Faces generated by “A Style-Based Generator Architecture for GANs”

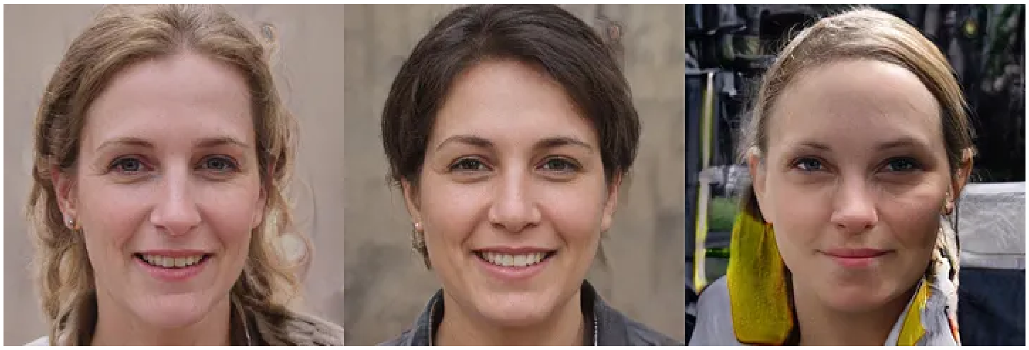
At low resolutions, almost all the images in the paper are indistinguishable from photographs. There are only a few artifacts that stand out to me that I will try to address.

The Missing Earring

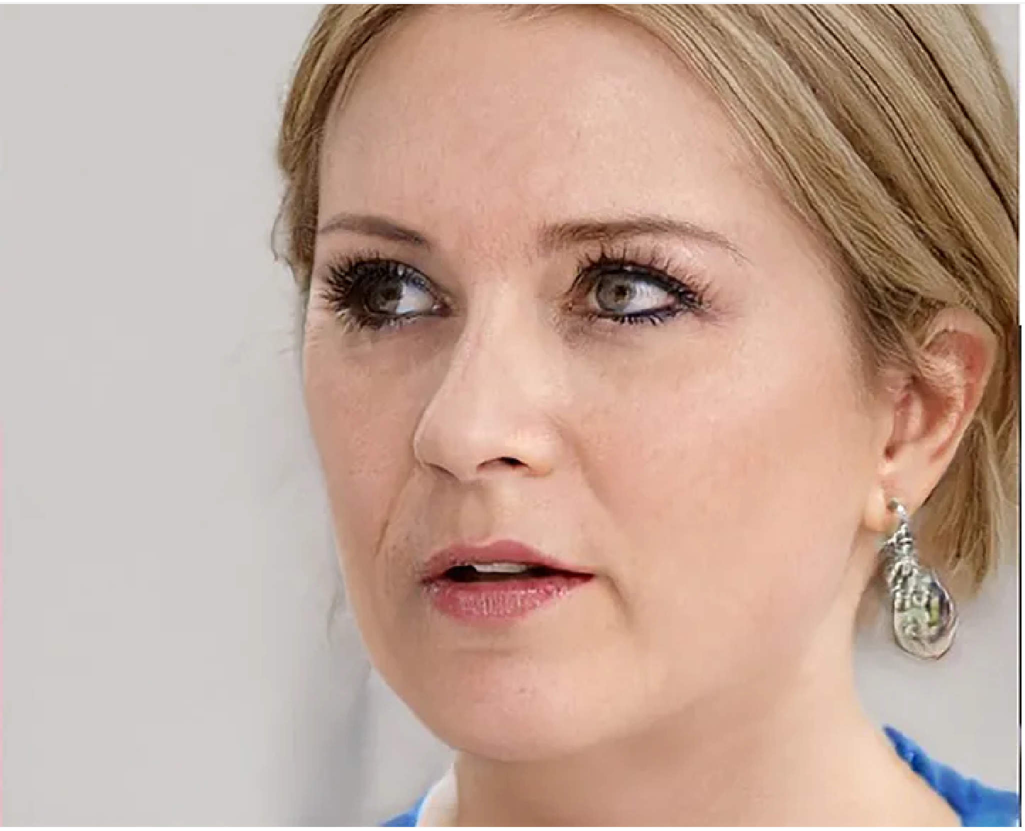


This glitch shows up in a few images in the exact same spot. This might have to do with the neural net trying to generate earrings and failing, because they all come from the same "source" image and in one case when mixed with a "middle style" showing a feminine face an earring appears in this spot. It could also be unrelated, because another example shows a similar glitch across multiple images in a totally different location.

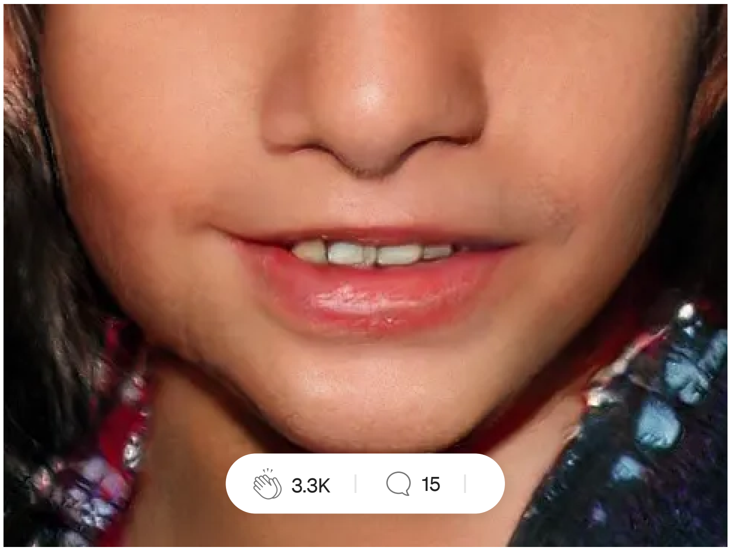
Asymmetry



In the center is the "average face" from the dataset, based on 70k photos from Flickr users all over the world. There appears to be an earring in the right ear (left side of image), but not in the left ear. This is not a judgement about whether having an earring in one ear is “right” or “wrong”, but about whether this kind of asymmetry is equally common in the dataset. The mismatched ear sizes in the right image is another example of asymmetry. Another example of overly frequent asymmetry might be this face that appears to have some strabismus: one eye seems to point in a different direction than the other.

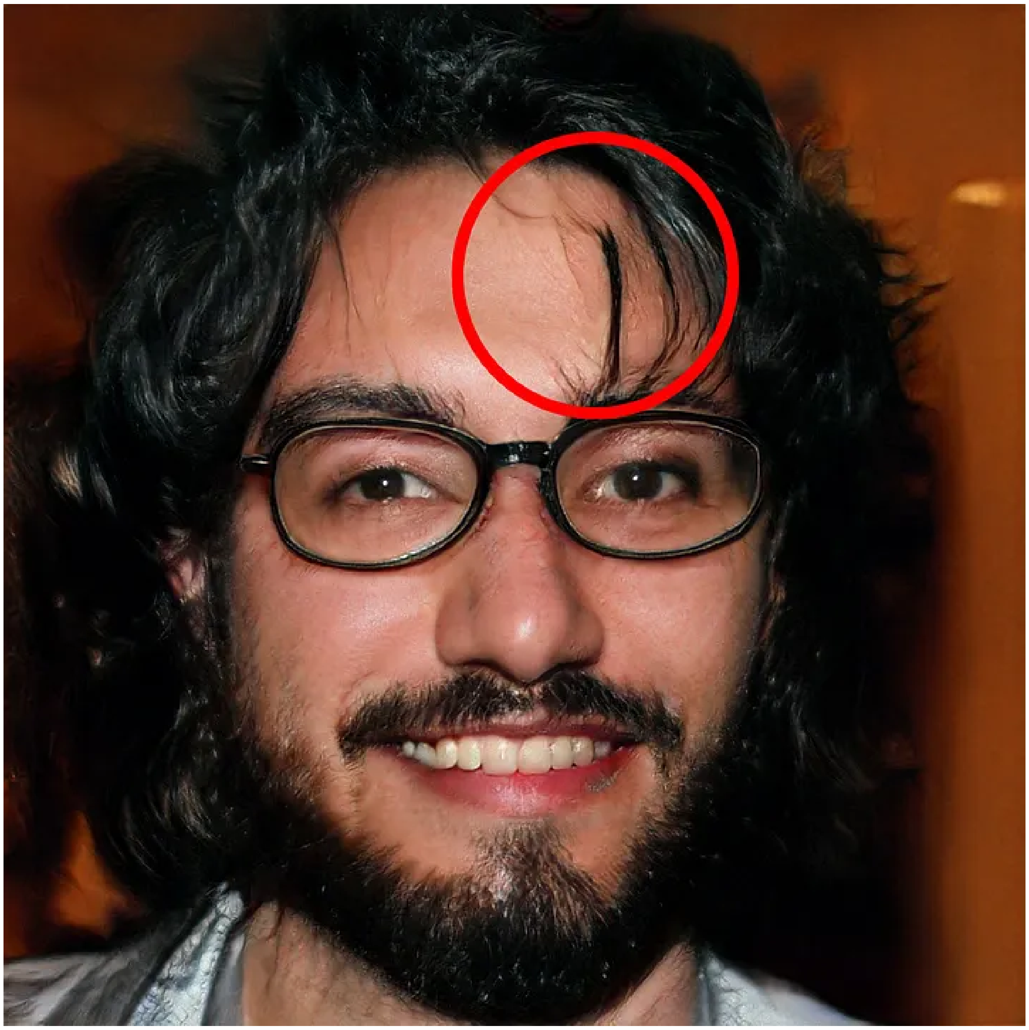


Weird teeth



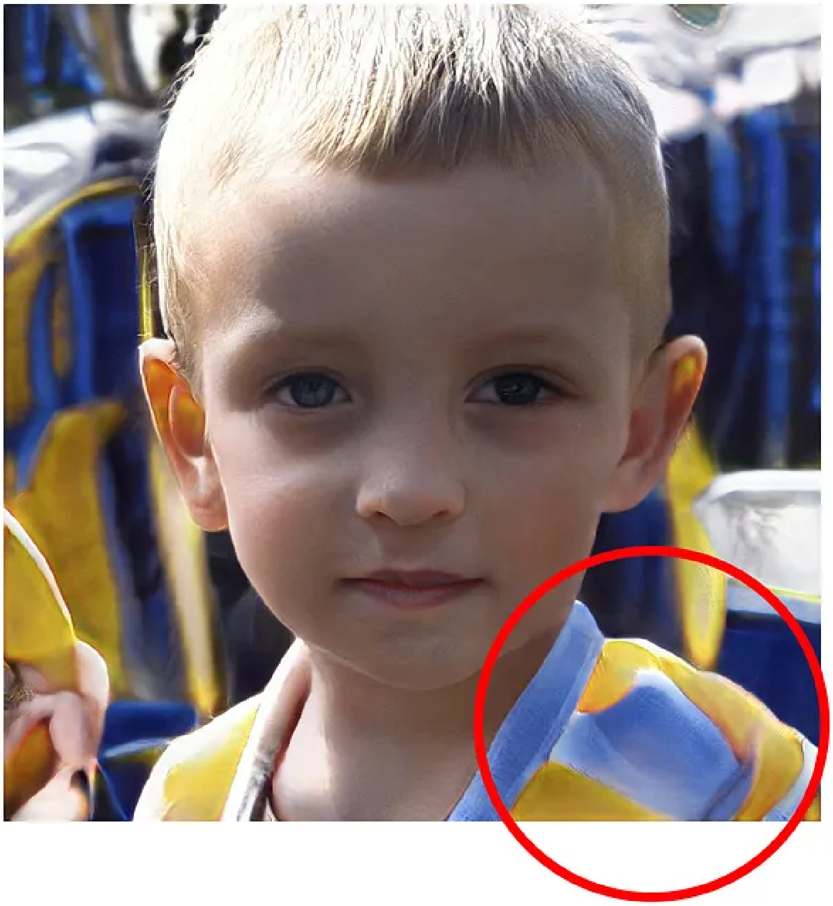
They’re still there, but you might have to look a little closer. In this example one tooth has a space in the middle. In other images they show all the teeth sliding to one side.

Messy hair



Also still there, but usually blending in a little better.

Painterly rendering



This one image has an unusual watercolor aesthetic. It’s not clear why this might appear. In their previous work, they used a super-resolution network to preprocess the training images. If they used the same system here. In the other “coarse styles copied” image, this region appears as some variant of a brightly colored shirt.

Artificial Intelligence Machine Learning Fake News Generative Adversarial