Generative adversarial networks unlock new methods for cognitive science

Lore Goetschalckx[1,2] 🖂, Alex Andonian[3], & Johan Wagemans[1]

[1.] Department of Brain and Cognition
KU Leuven
3000 Leuven, Belgium

[2.] Carney Institute for Brain Science
Department of Cognitive Linguistic & Psychological Sciences
Brown University
Providence, RI 2912

[3.] Computer Science and Artificial Intelligence Laboratory (CSAIL)
MIT
Cambridge, MA 02139

🖂 Corresponding author: lore_goetschalckx@brown.edu (L. Goetschalckx), @L_Goetschalckx

Abstract

Generative adversarial networks (GANs) enable computers to learn complex data distributions and sample from these distributions. When applied to the visual domain, this allows artificial, yet photo-realistic images to be synthesized. Their success at this very challenging task triggered an explosion of research within the field of artificial intelligence (AI), yielding various new GAN findings and applications. After explaining the core principles behind GANs and reviewing recent GAN innovations, we illustrate how they can be applied to tackle thorny theoretical and methodological problems in cognitive science. We focus on how GANs can reveal hidden structure in internal representations and how they offer a valuable new compromise in the trade-off between experimental control and ecological validity.

Keywords: generative adversarial networks; image synthesis; visual stimuli; experimental control; natural images

The Advent of GANs

The number and nature of advances in artificial intelligence (AI) of the last decade are nothing short of revolutionary. Not only is AI increasingly changing our daily lives (e.g., through the phones in our pockets, the cars we drive, etc.), it also has a major impact on other scientific fields. The field of cognitive science is no exception, as illustrated by **deep learning** (see Glossary). Ever since a **convolutional neural network (CNN)** triumphed at the 2012 ImageNet Large Scale Visual Recognition Challenge [1], these discriminative networks have been studied extensively as a potential model for visual recognition in the brain [e.g., 2–6]. In cognitive science, more and more labs are incorporating deep learning into their core research interests and methodological toolkits. Here, we explore why generative models, specifically generative adversarial networks (GANs) [7], may be the next generation of deep-learning models to advance cognitive science.

Cognitive science deals with phenomena that are inherently difficult to study. The internal representations and mechanisms of the human mind are complex, determined by a multitude of factors that do not easily lend themselves to measurement or **experimental control**. Often, the need for experimental control leads to sacrifices in **ecological validity** or vice versa, a tension which is also evident in the choice of stimuli. Simple, artificial stimuli are easier to manipulate and offer more experimental control, but generally lack the complexity and richness that is representative of our real-world experiences. Rich stimuli face the opposite problem. GANs' ability to synthesize artificial, yet realistic content may relieve this tension. We focus on visual content, more specifically on images, which is where GANs have been most successful. In what follows, we first introduce what GANs are and discuss recent GAN findings, before demonstrating how they can benefit cognitive science. Finally, we reflect on remaining challenges and provide directions for future research.

How GANs Work

In general, the kind of probability distribution that is learned is what sets generative models apart from discriminative models performing classification (such as the CNNs mentioned earlier), which are already widely used in cognitive science. The latter typically learn conditional probabilities, mapping high-dimensional, complex sensory information to a class label. This allows them to *discriminate*, for example, an image of a cat from an image of a dog (P("cat"|X) vs. P("dog"|X), where X is the sensory input). Generative models, on the other hand, learn the probability distribution over the complex sensory data itself. This allows the model to *generate* new sensory samples of the target distribution (e.g., a new cat image) in addition to discriminating. Due to the mathematically challenging nature of this task, generative modelling had previously attracted little attention. Its popularity changed drastically, though, when Ian Goodfellow, “the GANfather”, proposed GANs as a solution [7].

The key innovation behind GANs is to pit two models against each other as adversaries. Figure 1 (Key Figure) provides a schematic of how they work. One model acts as a generator. Its task is to learn to generate samples of a certain image distribution. The other model, the discriminator, learns to tell apart real samples of the training set from fake ones generated by its adversary, the generator. The latter comes down to a simple binary classification task: P(“real”|X) vs. P(“fake”|X). Therefore, the discriminator can take the form of a classic CNN. The generator’s architecture looks similar but reversed (i.e., **upsampling** rather than **downsampling** the input, using transposed convolutions). This is accentuated by the mirrored trapezoids in Figure 1. The generator takes as input a random noise vector, and produces an output with the same shape as the training images (height x width x number of channels). The noise vector can be thought of, metaphorically, as the generator’s inspiration. It needs input for every output. While the discriminator updates its parameters in order to decrease its classification error, the generator will do exactly the opposite. It will optimize its parameters to obtain higher classification error from the discriminator. Put differently, the generator tries to produce output that will fool the discriminator into labelling it “real”. The two models are trained jointly, such that when the discriminator gets better at detecting fake samples, the generator is forced to generate more realistic output, and vice versa. The competition between the two types of models is what drives the success of GANs in generating realistic output. While applied in many domains (e.g., text generation, speech synthesis), they have been most popular and successful in image synthesis.

The original GAN paper [7] triggered an explosion of further research, some of which we will briefly review below. This new research has made GANs produce ever more realistic output (see Figure 2), to the point that it can fool most casual observers [8,9]. While new kinds of generative models [e.g., 10] are being proposed at a rapid pace, we focus on GANs because of the level of realism achieved across a wide range of applications and the public availability of many pretrained models. For the same reasons, we think the time is ripe for cognitive scientists to start taking full advantage of GANs in their research.

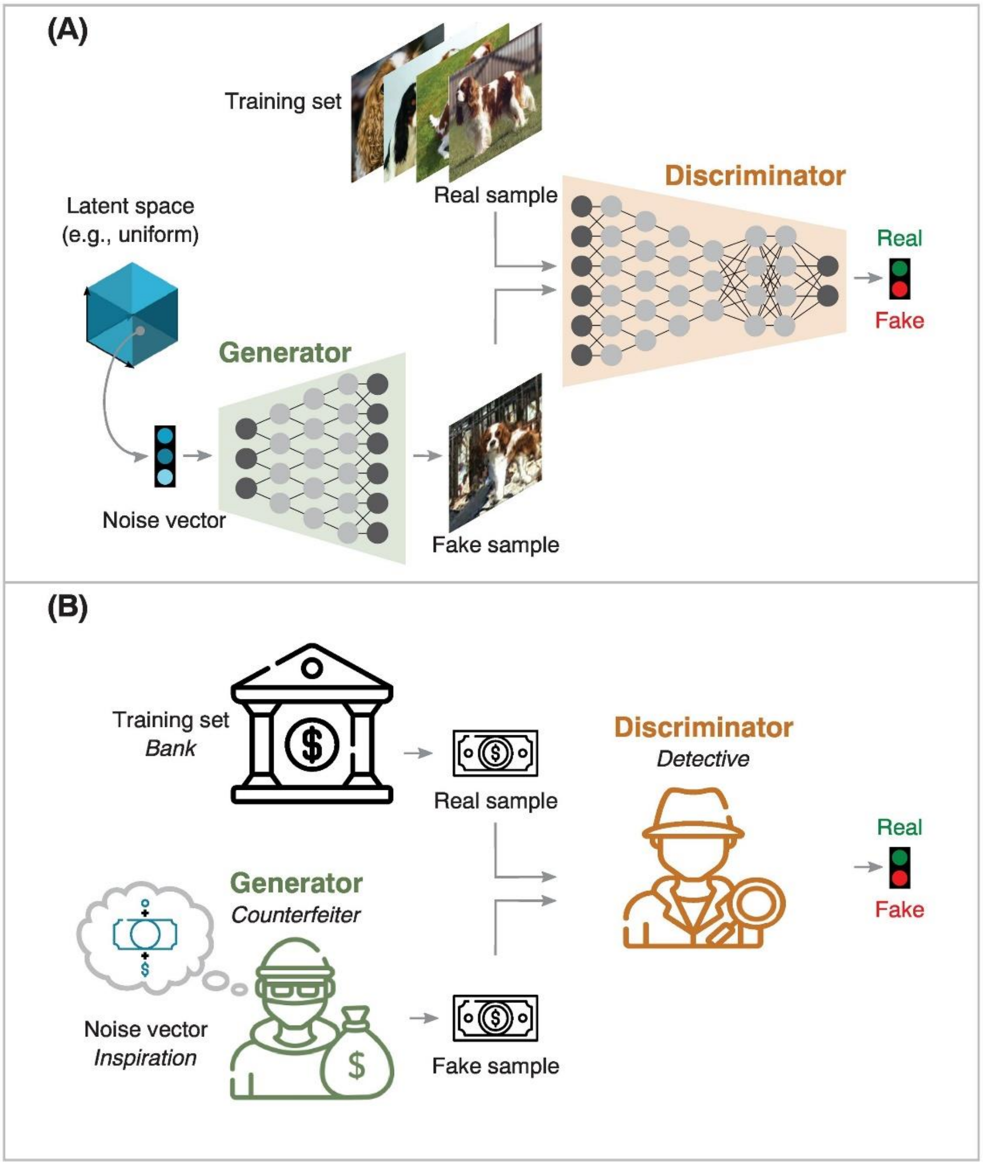


***Figure 1, Key Figure*. The GAN principle.** (A) A schematic of the generator and discriminator networks constituting a GAN. Note that this is a simplified representation. The generator upsamples noise vectors through multiple learned layers, producing image-like outputs. These are presented to the discriminator, intermixed with real training samples. The discriminator is trained to classify its input as coming from the training set (real) or the generator (fake). The generator is trained to have the discriminator incorrectly classify its output as real. The training samples shown are from [1]. The generated sample was synthesized with BigGAN [8]. (B) The counterfeiter-police/detective analogy illustrating the GAN logic [7]. The generator operates like a counterfeiter, producing fake money and wishing to improve their methods in order to fool the detective. The discriminator takes the role of the detective, who in turn also improves their methods to be more successful at catching the counterfeiter. This competition makes both of them increasingly better at their job. Icons made by Icongeek26 (dollar note) and Freepik (counterfeiter, detective, bank) from www.flaticon.com.

What GANs Can Do - Insights from AI

How does a GAN's latent space, where the noise vectors are sampled from (see Figure 1), map to the output image space? By itself, it is just random, uninterpretable noise, but a trained generator imposes structure and meaning on it, turning it into a lower-dimensional feature space. Moving through this space along certain (often just linear) directions has been found to result in smooth transitions of interpretable image attributes [e.g., 8,9,11,12] (see Figure 3). Moreover, performing simple arithmetic with noise vectors can have meaningful effects on the output image (e.g., smiling woman – neutral woman + neutral man = smiling man; [11]). It is also possible to interpolate between two semantic classes, creating peculiar new hybrid "breeds" [8] (see Figure 4.A). ArtBreeder[i], a collective artistic tool, is based on this principle.

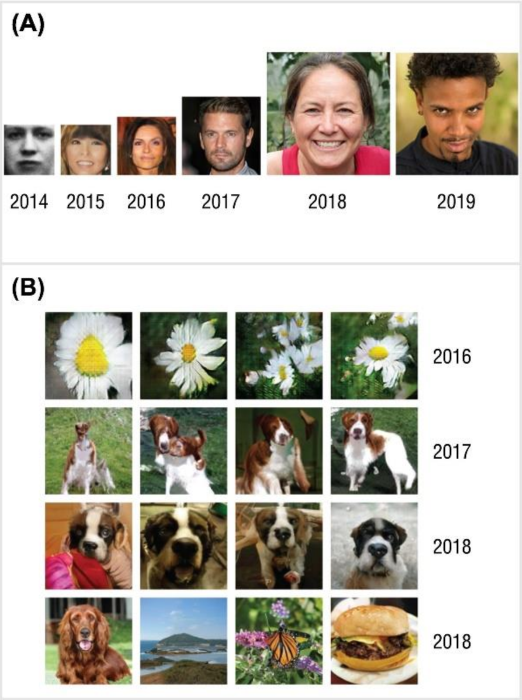


***Figure 2*. Recent progress by GANs trained on face images (A) and ImageNet [1] (B).** (A) Figure adapted from a tweet by Ian Goodfellow[viii]. From left to right, the images are from: GAN [7]; DCGAN [11]; CoGAN [99]; Progressive GAN [12], © 2018, NVIDIA Corporation; StyleGAN [9], © 2019, NVIDIA Corporation; StyleGAN2 [95], © 2019, NVIDIA Corporation. (B) Figure reprinted from [100]. From top to bottom, the images were produced by the GANs proposed in: [101], [102], [103], [8].

Inspired by such findings, different studies have developed ways to identify latent operations that can modify features of interest in generated images [13–17]. InterFaceGAN [15] lets one study the latent representations of facial attributes (e.g., age, gender, expression) in any pre-trained face GAN, as well as disentangle them such that they can be manipulated independently. Other work has identified linear directions matching simple image transformations (e.g., camera movements, color changes; [16]) in the latent space of BigGAN [8]. Finally, certain characteristics of generated scenes, like clutter and layout, can be modified through the method proposed in [17] (see Figure 4.B). Interestingly, the GAN they studied [9] inserts noise at every layer rather than only the input layer. Latent operations from early to late layers first control configuration and spatial layout, then object categories, and finally lower-level scene attributes like glossiness and color scheme [17].

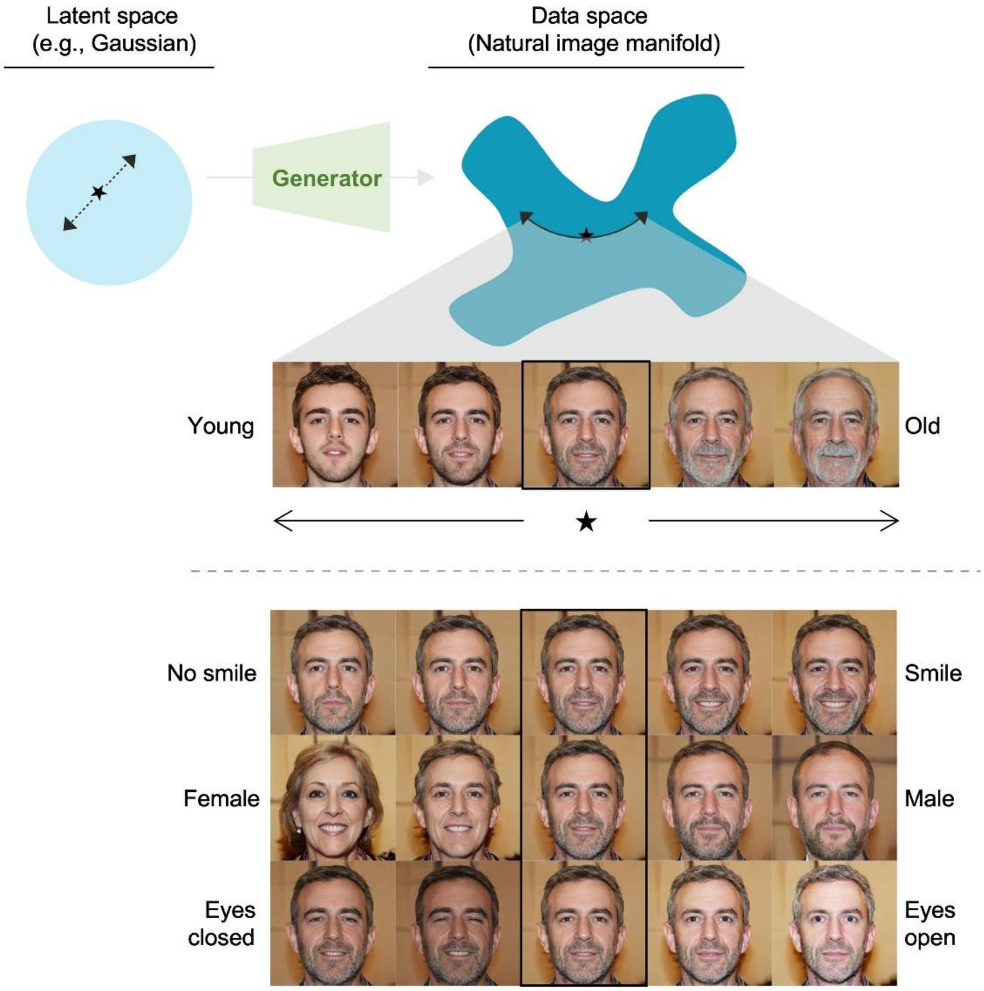
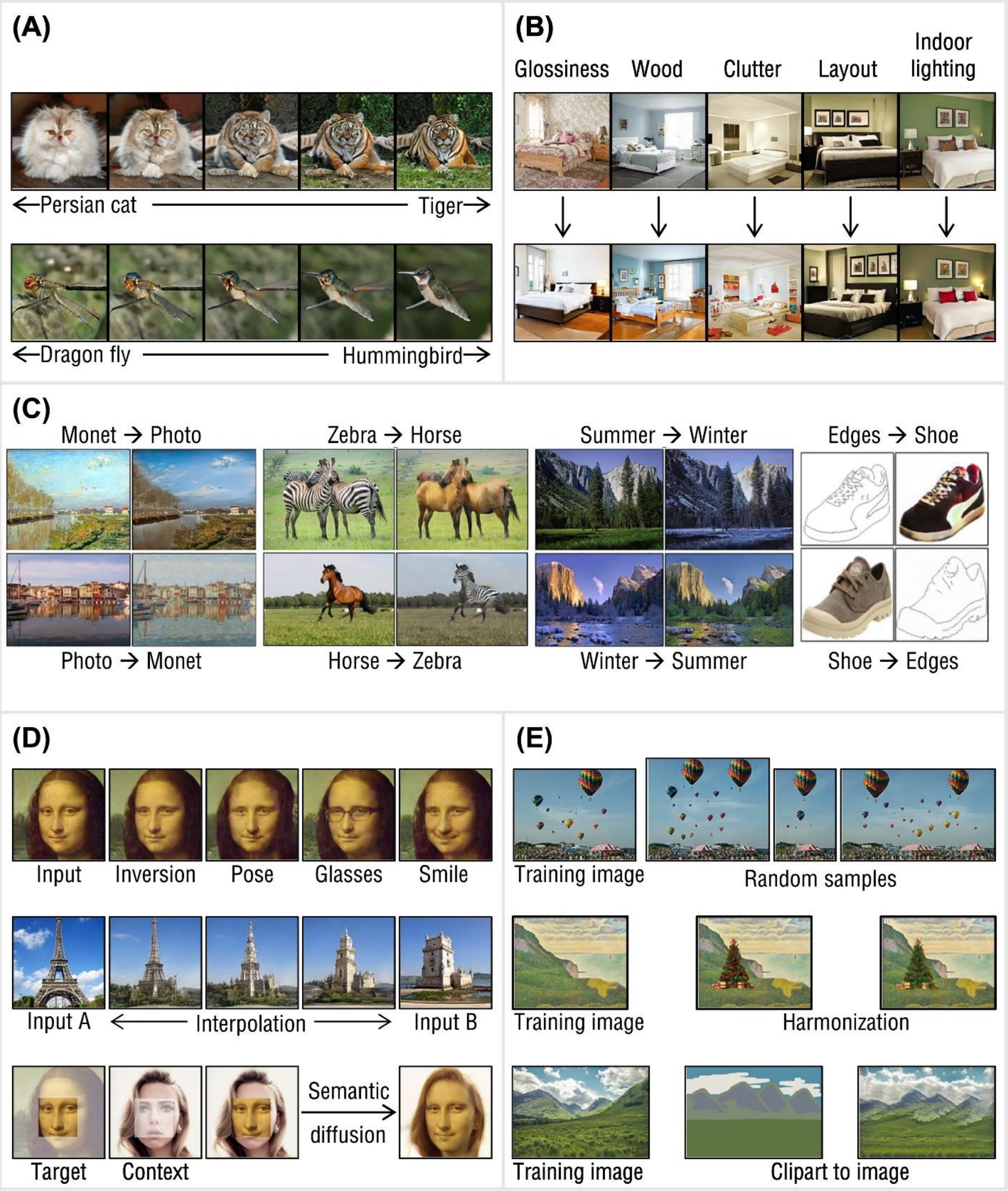


Figure 3. **Interpretable directions in the latent space of StyleGAN2**. By walking through the latent space of a GAN in a certain direction, one can vary for example the age of an output face image. Other latent directions affect other image attributes (e.g., smile, gender, and eye closure, as shown in the additional examples below the gray dotted line). The StyleGAN2 [95] directions as well as the example output images used for this figure were found by Robert Luxemburg[ix]. Note that in contrast to many more traditional GANs, StyleGAN2 inserts noise at every layer of the generator, not just the input layer. This makes for multiple latent spaces, but for simplicity, we only depict one here.

Note that this structure emerges even though GANs are not explicitly trained to represent information in this way, nor does structure emerge in the latent space alone. Recent work [18] studied how visual information is represented internally in the GAN's generator and discovered interpretable units in the network. These interpretable units are responsible for specific object concepts, such as "door units" and "tree units", and form the basis for the authors' scene editing tool: GANpaint [19, also see 20]. Extending this even further, the method proposed in [21] allows users to specify spatially localized modifications using free form descriptions in natural language.

Can we apply those same GAN manipulations to also edit real, user-supplied images? The missing ingredient has long been GAN inversion, the problem of reversely mapping a (real) image back to the GAN's latent space. If successful, feeding the resulting latent vector to the generator should faithfully reconstruct the original image. While this is far from a trivial task [22–24], substantial progress has been made [e.g., 19,22,25–27]. In one of the most recent contributions [22], the authors demonstrate how their inversion method enables modifying facial attributes, interpolation, and diffusing the inner parts of one face image into the outer parts of another (see Figure 4.D). This is achieved without altering the pretrained generator.

While so far we have discussed GANs trained to synthesize new images from scratch, GANs can also perform image-to-image translations [28]. In this scenario, the input is not a noise vector, but an image of a certain domain. The output is an image in another domain. Think of problems like turning grayscale to color, edges to image, photo to painting, etc. (see Figure 4.C). Before GANs, computer vision scientists had to come up with crafty **loss functions** that were tailored to a specific translation. However, GANs offer a more universal and often better solution [28]. CycleGAN [29], specifically, can actually be conceived as two GANs, one for each direction of the translation (e.g., horse → zebra, zebra → horse). Both have a discriminator that needs to be fooled into labelling the translated image as a real image from the respective target domain. An additional constraint is that when an input (from either domain) is cycled through both GANs (i.e., translated and back-translated), the output should be similar to the original input. Such image-to-image GANs can even turn anyone into a graceful ballerina [30] (by transferring movements from a professional dance video onto an amateur in another video), as well as give rise to new art forms[ii,iii]. Furthermore, the related problem of text to image translation has also been tackled with GANs [e.g., 31–33], although the most astonishing results on that task are from the Dall-E model [10]. Moreover, GANs have been employed to translate fMRI data back to the presented visual stimulus that evoked it [34–40], which is useful for uncovering internal neural representations.



(See figure legend at the top of the next page.)

Figure 4. **Examples of recent GAN innovations.** (A) Interpolation between two BigGAN [8] classes, creating new "breeds". Figure created using @genekogan's fork of the BigGAN colab notebook[x]. (B) Modifying visual attributes of generated scenes at different layers of a semantic hierarchy. Figure adapted from [17]. (C) CycleGAN performing image-to-image translations. Figure adapted from [29]. (D) In-domain GAN inversion and example applications [22]. The images labeled "Input", "Input A", "Input B", "Target" and "Context" are real, user-supplied images. Once the images are inverted into a GAN's latent space, one can apply GAN modifications. Each row represents an example application. Figure adapted from [22]. (E) Example applications of SinGAN [42]. SinGAN is trained on a single image (shown on the left). Each row represents an example application. Note that the random samples from the single training image in the top row have arbitrary sizes. Figure adapted from [42].

Finally, there is the idea of training a GAN on a single natural image, as was first adopted in InGAN [41]. Typical GANs learn the distribution of different images in a dataset. However, a single image in itself also constitutes a distribution, one of image patches. By learning this distribution at various patch scales (coarse and fine), GANs can capture a single image's internal statistics or "DNA" [41]. This technique can be used to retarget natural images to different output sizes, aspect ratios, and shapes without distorting its DNA. The more recent SinGAN [42] can also transform a clipart into a photo-realistic image or blend a new object with the training image, for example (see Figure 4.E). Note however that these GANs remain oblivious to most semantic aspects, which might result in semantically implausible output [41].

What GANs Can Do for Cognitive Science

The previous section highlighted recent GAN work that showcases their potential to generate content that is rich, complex and realistic, yet still allows for considerable control. By reconciling these two desirable but often conflicting experimental properties, GANs open up new opportunities for cognitive science. Next, we will support this by discussing concrete roles that GANs have fulfilled in our field or could fulfill in future work.

Exploiting and Understanding Representations

Many key questions in cognitive science are about mental representations, from the representations supporting perception to how experiences are encoded into memory. Common paradigms to probe representations in visual working memory (VWM) rely on the availability of a continuous stimulus space, such as a color space (e.g., to manipulate target-foil similarity, to create a color wheel for continuous reports) [43,44]. While these paradigms have brought important insights into the capacity, fidelity and feature encoding in VWM, they have mostly been limited to simple features. Recent work [45] proposed that GANs offer a way to extend this research toward more complex and ecologically valid scene images. Creating a scene wheel from a GAN's latent space, the study showed that participants' perceived similarity was indeed a function of distance on the wheel. Moreover, participants' error patterns in a memory experiment using the scene wheel for continuous reports mimicked those acquired with simpler stimuli. Other work [46] likewise proposed to exploit a GAN's latent space to obtain a window into the representations of mental categories.

Future work will have to further elucidate the psychological validity of the latent space as a representational space, especially considering that GANs rely on CNNs which are themselves still imperfect models of the visual system [5,47–49]. Nevertheless, the work discussed above constitutes a valuable and promising first step. In particular, exploiting GANs to study mental representations has two main advantages: (i) it is less constrained than only using hand-picked features while still offering some interpretability, (ii) the generator allows points in the latent space to be converted to stimuli(as also noted in [50]).

We also see potential for it to advance theory on topics beyond VWM and categorization. Increasing our understanding of how we recognize faces and are able to tell that two different images or views of a face are still the same person, is one example. Using a well-trained face GAN [e.g., 9] (an option also explored in [50]), future work could complement important studies like [51], which relied on hand-picked features and manual photo editing. Another example is a study [52] investigating how we represent shapes, which is crucial for object recognition and many other tasks. The authors of this study present ShapeComp [52], a space with 100 hand-picked features, and even trained a GAN to generate new shapes for validation purposes. Future work, however, could study the qualities of their GAN's latent space as a shape space in its own right.

Data-driven Hypothesis Generation

GANs' synthesizing abilities make them a suitable tool for data-driven hypothesis generation. GANalyze [53], for example, is a framework to help discover underlying features driving abstract cognitive image properties, like memorability. Indeed, images can be ranked reliably on their likelihood of being remembered [54]. While CNNs can accurately predict this ranking (e.g., MemNet [55]), they do not readily tell us what gives memorable images their high score. Guided by MemNet, GANalyze identified a memorability direction in a pretrained GAN's latent space. By walking along this direction, it gradually visualized what it means for a (generated) image to become more memorable (see Figure 5) according to MemNet. The resulting visualizations put forward candidate features that were class-orthogonal and indeed correlated with memorability (e.g., object size, colorfulness), thereby complementing previous work that had mostly stressed class-related effects (e.g., people are more memorable than landscapes). Related GAN work furthermore suggested that turning up a generated image's memorability score comes with more interpretable semantics [56]. Future work could apply the same framework to study other image properties (e.g., aesthetics, emotional valence; see Figure 5), provided that they can be quantified in a way that offers optimization gradients (e.g., through a CNN) and that the GAN has been trained on a relevant image set.

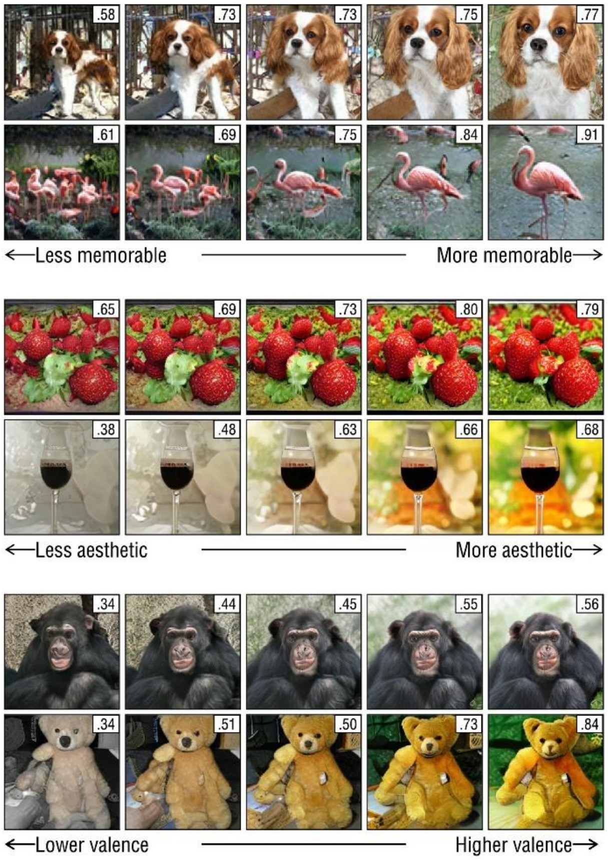


Figure 5. **Visual definitions of cognitive image properties produced by GANalyze [53].** By gradually moving along a direction in a GAN's latent space correlated to the property of interest, GANalyze provides "visual definitions" of what it means for an image to be characterized more (or less) by the given property. This process is steered by a CNN that predicts the property of interest. The predicted scores are indicated in an image's top right corner on a 0–1 scale. Aside from visualizing, the framework can also be used to generate a custom stimulus set to study the effect of a variable of interest on a behavioral or neural outcome.

Somewhat similarly, GANs can help formulate hypotheses on the features driving activations in CNNs and other artificial neural networks through "activation maximization" [e.g., 57]. This is of interest because such networks are popular (although somewhat controversial) models of the ventral visual stream and knowing their preferred stimulus tells us something about their internal representations. Recent work [58] was able to amplify the spiking activity of monkey V4 neurons by presenting images synthesized to maximize the activation in matched CNN units, which implies model-to-brain similarity. However, rather than synthesizing by directly optimizing pixel values to obtain activation maximization, one can optimize the latent code of a GAN, as proposed in [59]. The GAN acts as a naturalistic image prior making it easier to interpret the output and pinpoint candidate features. On the downside, the results will not only depend on the targeted unit, but also on the choice of the GAN and its training set.

Finally, XDREAM [60,61] targets the activity in monkey IT-neurons directly and evolves preferred stimuli through a combination of a **genetic algorithm** and a GAN in an online feedback loop. Using a GAN avoids to the need to hand-pick stimuli and lowers the risk of missing critical features when sampling from fully defined but more constrained parametric spaces. While mostly laying the groundwork for future studies, the initial experiments already suggested that a neuron's preferred stimulus need not be something that the monkey has ever encountered, that invariance to transformations might not be a fixed neuron feature, and that there is a major role for experience (e.g., some neurons evolved images resembling the monkey's caregivers). Finally, ongoing work [62] is making GANs synthesize preferred stimuli for higher order visual brain areas in humans (e.g., FFA, PPA, LOC) using a CNN trained on fMRI data for the optimization (as a proxy for neural activations). This could help elucidate to what extent these areas are truly category-specific and if so, for which category [e.g., 63,64].

Exploiting Natural Variation in Image Attributes

Another application is to use GANs to vary and experimentally test candidate features (whether surfaced by a GAN, derived from theory or any other source). For instance, one can train GANalyze to find an "object size" direction in the latent space. Generated images modified through this method to have the main object take up more space were indeed more likely to be remembered by participants in a memory task [53]. While this manipulation might still be automatable without GANs (e.g., through smart cropping), others might be less straightforward. Consider varying an image's aesthetic appeal. Sticking with the memorability question, one hypothesis could be that aesthetic appeal will make an image memorable. Previous studies had to rely on naturally varying images and found no or negative correlations [54,55]. GANalyze, however, efficiently varied generated images' predicted aesthetic score while at the same time keeping the class label constant. This had a small but positive effect on memorability, which suggests that the beauty of what is depicted (as reflected by the class) and how it is depicted might both affect memorability but probably in different ways.

These past works highlight GANs' possible advantage over two alternative strategies to test candidate features in the context of rich, complex images. A first strategy is to select real images that naturally vary along the critical dimension but are matched on others. Of course, this is labor-intensive and it is hard to satisfy all criteria. Arguably, GANs could handle this more efficiently. A second strategy is to manipulate images using existing photo editing techniques, which in many cases requires graphical skills and manual effort that GANs could avoid. We illustrate this further below by discussing example studies that could benefit from a GAN approach in future work. A caveat with this approach, though, is the fact that different features could be “entangled” in the latent space, which can hamper the possibility to manipulate candidate features in a fully independent fashion (see further).

How do children and adults recognize the material properties of objects based on visual input? To address this question, a recent study [65] designed a two alternative forced choice task asking which of two objects was real food. The stimuli were pairs of real food objects and a toy version of a different material (e.g., plastic), matched for shape and other features. As the authors note, the matching was only done “to the extent possible” [65, p. 3]. Here and in similar studies, CycleGAN could be used to translate images from one material domain to another (e.g., food → plastic), for easier matching and larger stimulus sets.

Various studies on face perception currently manipulate face images using morphing software, which can be effortful. In [66], the authors had to select over 100 anchor points to vary facial expression and study how perceived emotion might interact with a face’s perceived gender. These studies could benefit from GANs’ ability to efficiently modify an almost unlimited range of starting images [15,67,68], as noted before in [50]. Indeed, there is ongoing work from cognitive neuroscientists who are building a virtually infinite, parameterized face dataset that is based on GANs[iv,v]. They are looking to identify relevant feature dimensions and their corresponding GAN manipulation (e.g., identity, orientation, expression, age).

Beyond Natural Variation

Finally, we envision GANs unlocking possibilities for more dedicated types of stimuli, outside the range of natural variance, for use in specific paradigms.

First, studies on art perception attempt to find out what makes an artwork engaging. It is believed that ambiguity and indeterminacy play an important role, but as noted in [69], finding the right stimuli to test this can be challenging. Real art works suffer from possible confounds introduced by historical, stylistic or contextual factors. Simplified alternatives, such as Mooney images, do not offer the same richness. A recent contribution [69] proposed to use GAN art from the ArtBreeder[i] project instead. The work showed that GAN art spans a wide range of ambiguity, with ambiguity quantified as a function of the diversity in crowd-sourced image descriptions. While this work was mostly methodological and did not test the ambiguity hypothesis, it did provide a helpful starting point.

Second, there is work from AI (see above) demonstrating GANs' ability to generate a multitude of morph sequences, for instance, between two image classes (see Figure 4.A and [8]), two generated face identities [25], or even two user-supplied images (see Figure 4.D and [22,25]). As argued in [70], morph sequences have been employed to address a variety of questions in cognitive science, like how autistic-like traits correlate with the ability to flexibly switch from one concept to another [71], how serial dependence in face perception supports face recognition [72,73], and more [70,74–76]. However, compared to GAN-based morphs, current methods have the disadvantage of only allowing for a limited number of stimulus sets, often with stimuli that are not very rich or complex. At the same time, a possible limitation with GANs is that its morphing is less parameterized, and the result would also depend on the type of GAN and its training set (e.g., using BigGAN, see Figure 4.A, will also morph the background, which may or may not be desirable).

Third, we see potential for GANs to generate stimulus sets of "composite" faces, such as for behavioral and neural studies on holistic face processing [77–81]. Relevant GAN techniques are presented in [22], where inner face parts of one image are semantically diffused into the context of another by inversion into a GAN's latent space (see Figure 4.D).

A fourth example concerns studies asking whether scene-incongruent objects are attended to faster and recognized more easily [e.g., 82–84], remembered better [e.g., 85,86], etc. than scene-congruent objects (or vice versa). Perhaps GANs could arrive at realistically looking incongruent stimuli more efficiently compared to editing the stimuli manually in Adobe Photoshop CS 9.0 (Adobe, San Jose, CA) [87] or staging actual incongruent scenes and photographing them [88]. SinGAN [42] can blend (foreign) objects and backgrounds (see also [89,90]).

Remaining Challenges

Despite GANs' advantages, there remain some challenges or caveats to keep in mind. First, it is not yet clear how homologous GANs' latent representations are to perceptual or brain representations. The results in [45] (see above) show that distance in the latent space relates to perceived similarity and memory performance. Furthermore, the authors of a GAN-based method to recover face stimuli from brain activity [34] speculate that the method's success might be due to a topological similarity between the latent space and face representations in the brain. Nevertheless, future work should address this more systematically, especially because a variety of GANs exist. Indeed, a second challenge is understanding how the results in cognitive science studies might depend on the GANs' training set, and perhaps also its architecture. Some research questions might therefore require training a custom GAN, which brings a third challenge: "they are a devil to train" (quoting Alexei Efros[vi]). Fortunately, there is ongoing work to facilitate training (e.g., allowing for smaller training sets [91]). Fourth, how would we evaluate a GAN's performance (e.g., in terms of achieved realness)? This is still an open question. Different automated metrics have been put forward [92,93], but their psychological validity is unclear. An EEG-based alternative metric was proposed in [94]. Moreover, given a good metric, what level of realness would be required to safely compare results obtained with GAN stimuli to more traditional ones? Even the most qualitatively impressive GANs [e.g., 8,95] still contain some artefacts in their output. Finally, one should be wary of possible entanglements in the latent space when trying to manipulate features of interest (e.g., a latent space operation meant to remove a beard from a face, may inadvertently render other face attributes more feminine too). Such entanglements might represent real world correlations, but also dataset bias. Thankfully, this is actively being addressed in AI and different disentanglement strategies have already been proposed [e.g., 13,15]. Most of these challenges are currently hot topics of research, and it is possible that significant progress will be made soon.

Concluding Remarks

GANs achieve both realism and control in their visual output, thereby providing a middle ground in a trade-off faced by many cognitive scientists. In the short time they have been around, GANs have shed light on important research questions (e.g., in visual memory, neurophysiology). However, there still is much more potential to tap. In this review, we offered a primer of relevant concepts and findings from AI, and discussed directions on how those could advance cognitive science. Since GAN-based methods still have their own challenges, we see them as complementing, not replacing traditional methods. Future work could further improve GANs' deployability by addressing some of these challenges (see Outstanding Questions). Of course, as artificial neural networks continue to develop, it will be important to explore what other types of generative models beyond GANs [e.g., 10] can offer cognitive scientists. Finally, one can also speculate about the existence and relevance of generative models of/in the brain[vii] [e.g., 96]. Recent work [97] proposed a generative adversarial framework for probabilistic computation in the brain. Perhaps generative models also represent visual information in a more brain-like way than discriminative models do [e.g., 98].

Glossary

- **Convolutional neural network (CNN):** a kind of machine learning algorithm called artificial neural networks, which are loosely inspired by biology. They have been particularly successful in learning visual tasks, such as image classification. Typically, a CNN consists of a succession of layers of units (reminiscent of neurons), where the first layers apply learned local filters to the image through convolutions. The first layers produce feature maps that later layers learn to recombine globally to arrive at an accurate image prediction (e.g., a class label).

- **Deep learning:** a kind of machine learning in which complex data, such as image data, is modelled by different, successive levels of representation. Each level adds more abstraction from the original, raw input creating a hierarchy that bears similarity with the flow of information in the brain. A typical CNN is an example of a deep learning model. Each convolutional layer extracts learned features that are fed to the next, more high-level layer.

- **Downsampling:** producing output of reduced width and height compared to the original, image-like input. Going through the different layers of a typical CNN, the resulting feature maps get smaller and smaller.

- **Ecological validity:** the extent to which the results of a study can be generalized beyond the test settings, especially to more real-life settings. Simple, artificial visual stimuli are less representative of what we encounter in daily life than naturally occurring stimuli such as images. Therefore, they have lower ecological validity.

- **Experimental control:** the extent to which a study can prevent factors other than the ones being studied from affecting and thus confounding the results. The variation in simple, artificial visual stimuli can more easily be constrained to factors of interest than in naturally varying stimuli such as images.

- **Genetic algorithm:** search or optimization algorithm inspired by natural selection in biological evolution. Individuals of a population (e.g., latent GAN vectors) are evaluated on some fitness criterion (e.g., neural activation) and only the fittest individuals are selected to be recombined, with a possibility for mutations, into a next, fitter generation.

- **Latent space:** Many machine learning algorithms will learn to ENcode their highly complex input data into a different format that helps solve the task at hand. In doing so, they learn a mapping between the observable input space (e.g., image data) and a so-called latent (i.e., hidden) space (e.g., the feature space in a CNN). This space is often compressed. You can imagine a GAN's generator doing the opposite: its input comes from a compressed latent space and it DEcodes this into an observable output space (e.g., image data).

- **Loss function:** the function a machine learning algorithm needs to minimize, also known as cost function. It is a quantification of the dissimilarity between the algorithm's output and the desired output. GANs use adversarial loss, meaning that another model, an adversary that differentiates generated samples from real samples of the target distribution, defines the loss. A GAN's generator minimizes the (log) probability of the adversary being correct.

- **Upsampling:** the opposite of downsampling (see above). In contrast to a typical CNN, the layers of a GAN's generator upsample their input, meaning the height and width increase.

Outstanding Questions

- To what extent are visual representations in a GAN's latent space psychologically relevant and how similar are they to brain representations?

- How can a GAN's achieved realness be quantified, especially at the item level, and to what extent do automated measures reflect human's perception of realness?

- Which level of realness would be required to safely compare results obtained with GAN stimuli to more traditional ones? How good does the GAN have to be, especially in those cases when retraining is needed to suit very specific cognitive science designs?

- Can we extend the current methods used to disentangle image attributes in the latent space? Entangled attributes can result from real world correlations, but can also introduce unwanted bias or confounds. Different disentanglement strategies have already been proposed, but these are yet to cover the wider range of attributes and image types that cognitive science is interested in.

- How can we overcome some of the practical hurdles involved in training GANs when the available pretrained GANs are not a good fit for a particular research question? Fortunately, many training scripts are publicly available, but choosing the right training settings, using smaller datasets, avoiding typical pitfalls (e.g., the discriminator taking a large lead on the generator, halting further learning), etc. can still be challenging.

- To what extent can generative neural networks (GANs or others) model visual object recognition and other kinds of visual processing in the brain and how would this complement what is already known about discriminative models versus the brain? In addition, would they let us move towards modelling more open-ended generative processes (e.g., involving imagination, creativity)?

Highlights

- The internal mechanisms of the human mind are often complex, determined by a multitude of factors that do not lend themselves well to be measured or controlled experimentally. In many cases, the need for experimental control leads to sacrifices in ecological validity or vice versa.
- GANs offer a valuable compromise in the well-known trade-off between experimental control and ecological validity, thereby unlocking new methods for cognitive science that may lead to answers to questions that could not be addressed sufficiently before.
- GANs have recently achieved a huge leap forward in realistic image synthesis. They generate a continuous space of realistic looking images that offer surprising levels of control.
- GANs could become the next AI breakthrough to have a major impact on our field, after discriminative deep neural networks.

Acknowledgements

JW is supported by long-term structural funding from the Flemish Government (METH/14/02).

Resources

i. Simon, J. (n.d.) *Artbreeder* [Creative and collaborative online tool]. https://www.artbreeder.com

ii. Akten, M. (2017) *Learning to see: Gloomy sunday* [HD Video]. http://www.memo.tv/works/gloomy-sunday/

iii. Sarin, H. (2018, September 13) Playing a game of GANstruction. *The Gradient*. https://thegradient.pub/playing-a-game-of-ganstruction/

iv. Kietzmann, T. [@TimKietzmann] (2020, June 19) Face perception researchers, if you could design a new large-scale image dataset... [Tweet]. *Twitter*. https://twitter.com/TimKietzmann/status/1274026658148696065

v. Kietzmann, T. [@TimKietzmann] and Doerig, A. [@AdrienDoerig] (2020, June 5) Working on a new GAN-based face dataset with @AdrienDoerig. Emotions intensify. [Tweet]. *Twitter*. https://twitter.com/TimKietzmann/status/1268893738753024001

vi. Efros, A. (2019, May 31) *Reasons to love GANs* [Recorded presentation]. GANocracy: Workshop on Theory, Practice and Artistry of Deep Generative Modeling, Cambridge, MA. http://web.mit.edu/webcast/quest/sp19/

vii. Konkle, T. *et al.* (2021, April 13) *Testing generative models in the brain* [Recorded panel discussion]. Center for Brains, Minds, and Machines (CBMM) Panel Discussions. https://www.youtube.com/watch?v=g_KBUXU_UPM

viii. Goodfellow, I. [@goodfellow_ian] (2019, January 15) 4.5 years of GAN progress on face generation [Tweet]. *Twitter*. https://twitter.com/goodfellow_ian/status/1084973596236144640?lang=en

ix. Luxemburg, R. [@robertluxemburg] (2019, December 17) StyleGAN2 latent directions. [Tweet]. *Twitter*. https://twitter.com/robertluxemburg/status/1207087801344372736

x. https://t.co/8r7o5VVBMB?amp=1

References

1. Russakovsky, O. *et al.* (2015) ImageNet large scale visual recognition challenge. *International Journal of Computer Vision 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

2. Cichy, R. M. and Kaiser, D. (2019) Deep neural networks as scientific models. *Trends in Cognitive Sciences 23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

3. Kriegeskorte, N. and Golan, T. (2019) Neural network models and deep learning. *Current Biology 29*(7), R231–R236. https://doi.org/10.1016/j.cub.2019.02.034

4. Saxe, A. *et al.* (2021) If deep learning is the answer, what is the question? *Nature Reviews Neuroscience 22*(1), 55–67. https://doi.org/10.1038/s41583-020-00395-8

5. Serre, T. (2019) Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science 5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

6. Yamins, D. L. K. and DiCarlo, J. J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience 19*(3), 356–365. https://doi.org/10.1038/nn.4244

7. Goodfellow, I. *et al.* (2014) Generative adversarial nets. In *Advances in Neural Information Processing Systems 27* (Z. Ghahramani et al. eds.), pp. 2672–2680. Curran Associates, Inc. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

8. Brock, A. *et al.* (2019, May) *Large scale GAN training for high fidelity natural image synthesis* [Conference paper]. 7th International Conference on Learning Representations (ICLR), New Orleans, LA. https://openreview.net/forum?id=B1xsqj09Fm

9. Karras, T. *et al.* (2019) A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

10. Ramesh, A. *et al.* (2021) Zero-shot text-to-image generation. *ArXiv:2102.12092 [Cs]*,

Published online on February 26, 2021. https://arxiv.org/abs/2102.12092

11. Radford, A. *et al.* (2016, May) *Unsupervised representation learning with deep convolutional generative adversarial networks* [Conference paper]. 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico. http://arxiv.org/abs/1511.06434
12. Karras, T. *et al.* (2018, April) *Progressive growing of GANs for improved quality, stability, and variation* [Conference paper]. 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada. https://openreview.net/forum?id=Hk99zCeAb
13. Härkönen, E. *et al.* (2020) GANSpace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems 33* (H. Larochelle et al. eds.), pp. 9841–9850. Curran Associates, Inc. https://papers.nips.cc/paper/2020/file/6fe43269967adbb64ec6149852b5cc3e-Paper.pdf
14. Voynov, A. and Babenko, A. (2020) Unsupervised discovery of interpretable directions in the GAN latent space. In *Proceedings of the 37th International Conference on Machine Learning, in PLMR 119* (H. D. III & A. Singh eds.), pp. 9786–9796. MLResearchPress. http://proceedings.mlr.press/v119/voynov20a.html
15. Shen, Y. *et al.* (2020) Interpreting the latent space of GANs for semantic face editing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9240–9249. https://doi.org/10.1109/CVPR42600.2020.00926
16. Jahanian, A. *et al.* (2020, April) *On the “steerability” of generative adversarial networks* [Conference paper]. 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia. https://openreview.net/forum?id=HylsTT4FvB
17. Yang, C. *et al.* (2021) Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-020-01429-5

18. Bau, D. *et al.* (2019, May) *GAN dissection: Visualizing and understanding generative adversarial networks* [Conference paper]. 7th International Conference on Learning Representations (ICLR), New Orleans, LA. https://openreview.net/forum?id=Hyg_X2C5FX

19. Bau, D. *et al.* (2019) Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH) 38*(4), 1–11. https://doi.org/10.1145/3306346.3323023

20. Ashual, O. and Wolf, L. (2019) Specifying object attributes and relations in interactive scene generation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4560–4568. https://doi.org/10.1109/ICCV.2019.00466

21. Bau, D. *et al.* (2021) Paint by word. *ArXiv:2103.10951 [Cs]*, Published online on March 19, 2021. http://arxiv.org/abs/2103.10951

22. Zhu, J. *et al.* (2020) In-domain GAN inversion for real image editing. In *Computer Vision – ECCV 2020* (A. Vedaldi et al. eds.), pp. 592–608. Springer International Publishing.

23. Bau, D. *et al.* (2019) Seeing what a GAN cannot generate. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4501–4510. https://doi.org/10.1109/ICCV.2019.00460

24. Li, K. and Malik, J. (2018) On the implicit assumptions of GANs. *ArXiv:1811.12402 [Cs, Stat]*, Published online on November 29, 2018. http://arxiv.org/abs/1811.12402

25. Abdal, R. *et al.* (2019) Image2StyleGAN: How to embed images into the StyleGAN latent space? *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4431–4440. https://doi.org/10.1109/ICCV.2019.00453

26. Anirudh, R. *et al.* (2020) MimicGAN: Robust projection onto image manifolds with corruption mimicking. *International Journal of Computer Vision 128*(10), 2459–2477. https://doi.org/10.1007/s11263-020-01310-5

27. Creswell, A. and Bharath, A. A. (2019) Inverting the generator of a generative adversarial

network. *IEEE Transactions on Neural Networks and Learning Systems 30*(7), 1967–1974. https://doi.org/10.1109/TNNLS.2018.2875194

28. Isola, P. *et al.* (2017) Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632
29. Zhu, J.-Y. *et al.* (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244
30. Chan, C. *et al.* (2019) Everybody dance now. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5932–5941. https://doi.org/10.1109/ICCV.2019.00603
31. Reed, S. *et al.* (2016) Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning, in PMLR 48* (M. F. Balcan & K. Q. Weinberger eds.), pp. 1060–1069. MLResearchPress. http://proceedings.mlr.press/v48/reed16.html
32. Zhang, H. *et al.* (2019) StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*(8), 1947–1962. https://doi.org/10.1109/TPAMI.2018.2856256
33. Li, W. *et al.* (2019) Object-driven text-to-image synthesis via adversarial training. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12166–12174. https://doi.org/10.1109/CVPR.2019.01245
34. VanRullen, R. and Reddy, L. (2019) Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology 2*(1), 1–10. https://doi.org/10.1038/s42003-019-0438-y
35. Lin, Y. *et al.* (2019, May 27) *DCNN-GAN: Reconstructing realistic image from fMRI* [Conference paper]. 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan. https://doi.org/10.23919/MVA.2019.8757985

36. Le, L. *et al.* (2021) Brain2Pix: Fully convolutional naturalistic video reconstruction from brain activity. *BioRxiv*, Published online on February 3, 2021. https://doi.org/10.1101/2021.02.02.429430

37. Qiao, K. *et al.* (2020) BigGAN-based Bayesian reconstruction of natural images from human brain activity. *Neuroscience 444*, 92–105. https://doi.org/10.1016/j.neuroscience.2020.07.040

38. Seeliger, K. *et al.* (2018) Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage 181*, 775–785. https://doi.org/10.1016/j.neuroimage.2018.07.043

39. Shen, G. *et al.* (2019) Deep image reconstruction from human brain activity. *PLoS Computational Biology 15*(1). https://doi.org/10.1371/journal.pcbi.1006633

40. St-Yves, G. and Naselaris, T. (2018) Generative adversarial networks conditioned on brain activity reconstruct seen images. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1054–1061. https://doi.org/10.1109/SMC.2018.00187

41. Shocher, A. *et al.* (2019) InGAN: Capturing and retargeting the “DNA” of a natural image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4491–4500. https://doi.org/10.1109/ICCV.2019.00459

42. Shaham, T. R. *et al.* (2019) SinGAN: Learning a generative model from a single natural image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4569–4579. https://doi.org/10.1109/ICCV.2019.00467

43. Luck, S. J. and Vogel, E. K. (2013) Visual Working Memory Capacity: From Psychophysics and Neurobiology to Individual Differences. *Trends in Cognitive Sciences 17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

44. Schurgin, M. W. *et al.* (2020) Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour 4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

45. Son, G.et al.(2021) Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *Behav. Res. Methods.* https://doi.org/10.3758/s13428-021-01630-5

46. Peterson, J. C. *et al.* (2018, July 25) *Capturing human category representations by sampling in deep feature spaces* [Conference paper]. 40th Annual Meeting of the Cognitive Science Society (CogSci), Madison, WI. http://cocosci.princeton.edu/jpeterson/pdf/peterson-cogsci2018-deep-mcmcp.pdf

47. Xu, Y. and Vaziri-Pashkam, M. (2021) Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications 12*(1), 2065. https://doi.org/10.1038/s41467-021-22244-7

48. Kar, K. *et al.* (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience 22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

49. Spoerer, C. J. *et al.* (2017) Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology 8*. https://doi.org/10.3389/fpsyg.2017.01551

50. Suchow, J. W. *et al.* (2018) Learning a face space for experiments on human identity. *ArXiv:1805.07653 [Cs]*, Published online on May 19, 2018. http://arxiv.org/abs/1805.07653

51. Abudarham, N. and Yovel, G. (2016) Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision 16*(3), 40. https://doi.org/10.1167/16.3.40

52. Morgenstern, Y. *et al.* (2020) An image-computable model of human visual shape similarity. *BioRxiv*, Published online on January 11, 2020. https://doi.org/10.1101/2020.01.10.901876

53. Goetschalckx, L. *et al.* (2019) GANalyze: Toward visual definitions of cognitive image

properties. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5743–5752. https://doi.org/10.1109/ICCV.2019.00584

54. Isola, P. *et al.* (2014) What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*(7), 1469–1482. https://doi.org/10.1109/tpami.2013.200

55. Khosla, A. *et al.* (2015) Understanding and predicting image memorability at a large scale. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2390-2398. https://doi.org/10.1109%2Ficcv.2015.275

56. Kyle-Davidson, C. *et al.* (2020) Generating memorable images based on human visual memory schemas. *ArXiv:2005.02969 [Cs]*, Published online on May 6, 2020. http://arxiv.org/abs/2005.02969

57. Yosinski, J. *et al.* (2015, July 10) *Understanding neural networks through deep visualization* [Conference paper]. Deep Learning Workshop at the International Conference on Machine Learning (ICML), Lille, France. https://8109f4a4-a-62cb3a1a-s-sites.googlegroups.com/site/deeplearning2015/46.pdf?attachauth=ANoY7crvt-JqBaqUUjczmLGesJe2mpIolmS83fs3FxSylShO6fRVeifzzd7OwG3moG5Gg-sRzuG5q4-uuKAH2las-j2reGpqFRXgoj7BpVj_ZLeqGkPcwGM3Nekgmyk2lKmCMK9qo0_y5KB6JRh-FWeBhggIrw-9SGDZAKGQ62fHfA-5tbQznM4BPD-WtSUQcGG_8PYqirgBt4LqOpTZTztGINRlVRrTXA%3D%3D&attredirects=0

58. Bashivan, P. *et al.* (2019) Neural population control via deep image synthesis. *Science 364*(6439). https://doi.org/10.1126/science.aav9436

59. Nguyen, A. *et al.* (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems 29* (D. Lee et al. eds.). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/5d79099fcdf499f12b79770834c0164a-

Paper.pdf

60. Ponce, C. R. *et al.* (2019) Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell 177*(4), 999-1009.e10. https://doi.org/10.1016/j.cell.2019.04.005

61. Xiao, W. and Kreiman, G. (2020) XDream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Computational Biology 16*(6), e1007973. https://doi.org/10.1371/journal.pcbi.1007973

62. Roth, J. *et al.* (2021, February 24) *Synthesizing preferred stimuli for individual voxels in the human visual system* [Conference abstract]. Cosyne, Online.

63. Burns, E. J. *et al.* (2019) P-curving the fusiform face area: Meta-analyses support the expertise hypothesis. *Neuroscience and Biobehavioral Reviews 104*, 209–221. https://doi.org/10.1016/j.neubiorev.2019.07.003

64. Rajimehr, R. *et al.* (2011) The “Parahippocampal Place Area” Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. *PLOS Biology 9*(4), e1000608. https://doi.org/10.1371/journal.pbio.1000608

65. Balas, B. *et al.* (2020) Children’s use of local and global visual features for material perception. *Journal of Vision 20*(2), 10. https://doi.org/10.1167/jov.20.2.10

66. Harris, D. A. *et al.* (2016) What’s in a Face? How face gender and current affect influence perceived emotion. *Frontiers in Psychology 7*. https://doi.org/10.3389/fpsyg.2016.01468

67. Geng, Z. *et al.* (2020) Towards photo-realistic facial expression manipulation. *International Journal of Computer Vision 128*(10), 2744–2761. https://doi.org/10.1007/s11263-020-01361-8

68. Ververas, E. and Zafeiriou, S. (2020) SliderGAN: Synthesizing expressive face images by sliding 3D blendshape parameters. *International Journal of Computer Vision 128*(10), 2629–2650. https://doi.org/10.1007/s11263-020-01338-7

69. Wang, X. *et al.* (2020) Toward quantifying ambiguities in artistic images. *ACM Transactions*

on Applied Perception 17(4). https://doi.org/10.1145/3418054

70. Stöttinger, E. *et al.* (2016) Assessing perceptual change with an ambiguous figures task: Normative data for 40 standard picture sets. *Behavior Research Methods 48*(1), 201–222. https://doi.org/10.3758/s13428-015-0564-5

71. Burnett, H. G. and Jellema, T. (2013) (Re-)conceptualisation in Asperger's Syndrome and Typical Individuals with Varying Degrees of Autistic-like Traits. *Journal of Autism and Developmental Disorders 43*(1), 211–223. https://doi.org/10.1007/s10803-012-1567-z

72. Liberman, A. *et al.* (2014) Serial dependence in the perception of faces. *Current Biology 24*(21), 2569–2574. https://doi.org/10.1016/j.cub.2014.09.025

73. Turbett, K. *et al.* (2019) Individual differences in serial dependence of facial identity are associated with face recognition abilities. *Scientific Reports 9*(1), 18020. https://doi.org/10.1038/s41598-019-53282-3

74. Hartendorp, M. O. *et al.* (2010) Categorical perception of morphed objects using a free-naming experiment. *Visual Cognition 18*(9), 1320–1347. https://doi.org/10.1080/13506285.2010.482774

75. Verstijnen, I. M. and Wagemans, J. (2004) Ambiguous figures: Living versus nonliving objects. *Perception 33*(5), 531–546. https://doi.org/10.1068/p5213

76. Newell, F. N. and Bülthoff, H. H. (2002) Categorical perception of familiar objects. *Cognition 85*(2), 113–143. https://doi.org/10.1016/S0010-0277(02)00104-X

77. Knowles, M. M. and Hay, D. C. (2014) The role of inner and outer face parts in holistic processing: A developmental study. *Acta Psychologica 149*, 106–116. https://doi.org/10.1016/j.actpsy.2014.03.012

78. Andrews, T. J. *et al.* (2010) Internal and external features of the face are represented holistically in face-selective regions of visual cortex. *Journal of Neuroscience 30*(9), 3544–3552. https://doi.org/10.1523/JNEUROSCI.4863-09.2010

79. Hills, C. *et al.* (2014) An adaptation study of internal and external features in facial

representations. *Vision Research 100*, 18–28. https://doi.org/10.1016/j.visres.2014.04.002

80. Logan, A. J. *et al.* (2019) From individual features to full faces: Combining aspects of face information. *Journal of Vision 19*(4), 23. https://doi.org/10.1167/19.4.23

81. Peters, J. C. *et al.* (2018) From coarse to fine: Interactive feature processing precedes local feature analysis in human face perception. *Biological Psychology 138*, 1–10. https://doi.org/10.1016/j.biopsycho.2018.07.009

82. Greene, M. R. *et al.* (2015) What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics 77*(4), 1239–1251. https://doi.org/10.3758/s13414-015-0859-8

83. Truman, A. and Mudrik, L. (2018) Are incongruent objects harder to identify? The functional significance of the N300 component. *Neuropsychologia 117*, 222–232. https://doi.org/10.1016/j.neuropsychologia.2018.06.004

84. De Graef, P. *et al.* (1990) Perceptual effects of scene context on object identification. *Psychological Research 52*(4), 317–329. https://doi.org/10.1007/BF00868064

85. Friedman, A. (1979) Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General 108*(3), 316–355. http://dx.doi.org.kuleuven.ezproxy.kuleuven.be/10.1037/0096-3445.108.3.316

86. Bainbridge, W. A. *et al.* (2020) Disrupted object-scene semantics boost scene recall but diminish object recall in drawings from memory. *BioRxiv*, Published online on May 14, 2020. https://doi.org/10.1101/2020.05.12.090910

87. Leroy, A. *et al.* (2020) Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. *Scientific Reports 10*(1), 8447. https://doi.org/10.1038/s41598-020-65158-y

88. Coco, M. I. *et al.* (2019) Fixation-related brain potentials during semantic integration of object–scene information. *Journal of Cognitive Neuroscience 32*(4), 571–589.

https://doi.org/10.1162/jocn_a_01504

89. Chai, L. *et al.* (2021) Using latent space regression to analyze and leverage compositionality in GANs. *ArXiv:2103.10426 [Cs]*, Published online on March 18, 2021. http://arxiv.org/abs/2103.10426

90. Azadi, S. *et al.* (2020) Compositional GAN: Learning image-conditional binary composition. *International Journal of Computer Vision 128*(10), 2570–2585. https://doi.org/10.1007/s11263-020-01336-9

91. Karras, T. *et al.* (2020) Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems 33* (H. Larochelle et al. eds.), pp. 12104–12114. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf

92. Salimans, T. *et al.* (2016) Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29* (D. Lee et al. eds.). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf

93. Heusel, M. *et al.* (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30* (I. Guyon et al. eds.). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

94. Wang, Z. *et al.* (2020) Use of neural signals to evaluate the quality of generative adversarial network performance in facial image generation. *Cognitive Computation 12*(1), 13–24. https://doi.org/10.1007/s12559-019-09670-y

95. Karras, T. *et al.* (2020, June) Analyzing and improving the image quality of StyleGAN. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR).

96. Al-Tahan, H. and Mohsenzadeh, Y. (2021) Reconstructing feedback representations in the ventral visual pathway with a generative adversarial autoencoder. *PLOS Computational Biology 17*(3), e1008775. https://doi.org/10.1371/journal.pcbi.1008775

97. Gershman, S. J. (2019) The generative adversarial brain. *Frontiers in Artificial Intelligence 2*. https://doi.org/10.3389/frai.2019.00018

98. Golan, T. *et al.* (2020) Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences 117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

99. Liu, M.-Y. and Tuzel, O. (2016) Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems 29* (D. D. Lee et al. eds.), pp. 469–477. Curran Associates, Inc. http://papers.nips.cc/paper/6544-coupled-generative-adversarial-networks.pdf

100. Odena, A. (2019) Open questions about generative adversarial networks. *Distill*. https://doi.org/10.23915/distill.00018

101. Odena, A. *et al.* (2017) Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning, in PMLR 70* (D. Precup & Y. W. Teh eds.), pp. 2642–2651. MLResearchPress. http://proceedings.mlr.press/v70/odena17a/odena17a.pdf

102. Miyato, T. and Koyama, M. (2018, April 30) *CGANs with projection discriminator* [Conference paper]. 6th International Conference on Learning Representations, Vancouver, BC, Canada. https://openreview.net/pdf?id=ByS1VpgRZ

103. Zhang, H. *et al.* (2019) Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning, in PMLR 97*, pp. 7354–7363. MLResearchPress. http://proceedings.mlr.press/v97/zhang19d/zhang19d.pdf