

GANALYZE: TOWARD VISUAL DEFINITIONS OF COGNITIVE IMAGE PROPERTIES

A PREPRINT

Lore Goetschalckx*
MIT, KU Leuven
lgoetsch@mit.edu

Alex Andonian*
MIT
aandonia@mit.edu

Aude Oliva
MIT
oliva@mit.edu

Phillip Isola
MIT
phillipi@mit.edu

August 10, 2019

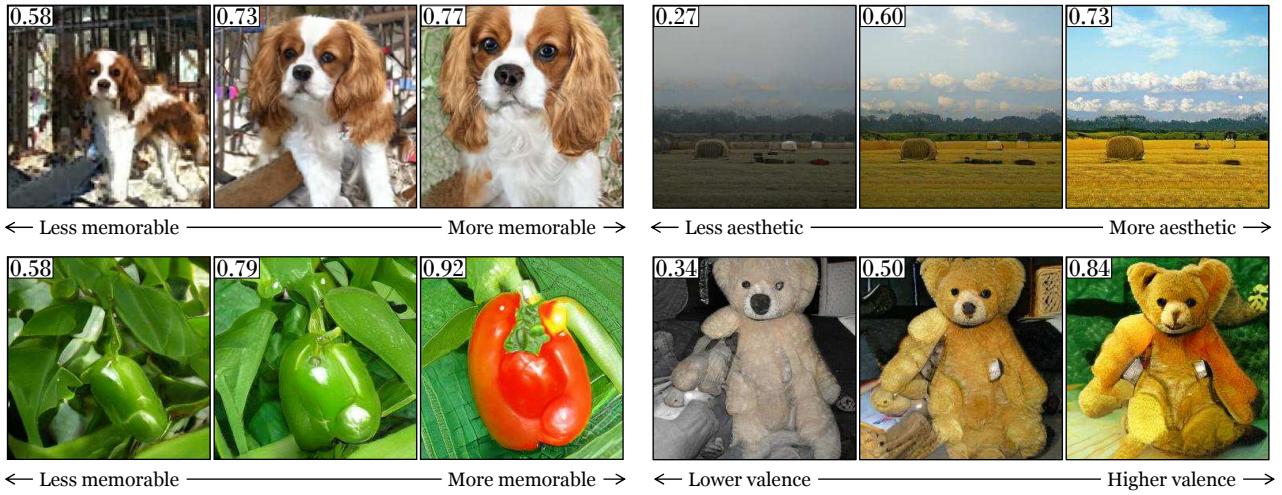


Figure 1: **Visualizations produced by the proposed GANalyze framework.** The middle columns represent generated images serving as the original seed. The originals are then modified to be characterized more (right) or less (left) by a given property of interest (memorability, aesthetics, or emotional valence). The images’ respective property scores are presented in their top left corner.

ABSTRACT

We introduce a framework that uses Generative Adversarial Networks (GANs) to study cognitive properties like memorability, aesthetics, and emotional valence. These attributes are of interest because we do not have a concrete visual definition of what they entail. What does it look like for a dog to be more or less memorable? GANs allow us to generate a manifold of natural-looking images with fine-grained differences in their visual attributes. By navigating this manifold in directions that increase memorability, we can visualize what it looks like for a particular generated image to become more or less memorable. The resulting “visual definitions” surface image properties (like “object size”) that may underlie memorability. Through behavioral experiments, we verify that our method indeed discovers image manipulations that causally affect human memory performance. We further demonstrate that the same framework can be used to analyze image aesthetics and emotional valence. Visit the GANalyze website at <http://ganalyze.csail.mit.edu/>.

Keywords Generative Adversarial Networks · visualizations · memorability · aesthetics · emotional valence

1 Introduction

Why do we remember the things we do? Decades of work have provided numerous explanations: we remember things that are out of context [1, 2], that are emotionally salient [3], that involve people [4], etc. But a picture is, as they say, worth a thousand words. What does it *look like* to make an image more or less memorable? The same questions can be asked for many cognitive visual properties: what visual changes can take a bland foggy seascape and add just the right colors and tones to make it serenely beautiful.

Attributes like memorability, aesthetics, and emotional valence are of special interest because we do not have concrete definitions of what they entail. This contrasts with attributes like "object size" and "smile". We know exactly what it means to zoom in on a photo, and it's easy to imagine what a face looks like as it forms a smile. It's an open question, on the other hand, what exactly do changes in "memorability" look like? Previous work has built powerful predictive models of image memorability [4, 5] but these have fallen short of providing a fine-grained visual explanation of what underlies the predictions.

In this paper, we propose a new framework, GANalyze, based on Generative Adversarial Networks (GAN) [6], to study the **visual features** and properties that underlie high-level cognitive attributes. We focus on image memorability as a case study, but also show that the same methods can be applied to study image aesthetics and emotional valence.

Our approach leverages the ability of GANs to generate a continuum of images with fine-grained differences in their visual attributes. We can learn how to navigate the GAN's latent space to produce images that have increasing or decreasing memorability, according to an off-the-shelf memorability predictor [5]. Starting with a seed image, this produces a sequence of images of increasing and decreasing predicted memorability (see Figure 1). By showing this visualization for a diverse range of seed images, we come up with a catalog of different image sequences showcasing a variety of visual effects related to memorability. We call this catalog a *visual definition* of image memorability. GANalyze thereby offers an alternative to the non-parametric approach in which real images are simply sorted on their memorability score to visualize what makes them memorable (example shown in Figure S1). The parametric, fine-grained visualizations generated by GANalyze provide much clearer visual definitions.

These visualizations surface several correlates of memorability that have been overlooked by prior work, including "object size", "circularity", and "colorfulness". Most past work on modeling image memorability focused on semantic attributes, such as object category (e.g., "people" are more memorable than "trees") [4]. By applying our approach to a class-conditional GAN, BigGAN [7], we can restrict it to only make changes that are orthogonal to object class. This reveals more fine-grained changes that nonetheless have large effects on predicted memorability. For example, consider the cheeseburgers in Figure 4. Our model visualizes more memorable cheeseburgers as we move to the right. The apparent changes go well beyond semantic category – the right-most burger is brighter, rounder, more canonical, and, we think, looks tastier.

Since our visualizations are learned based on a *model* of memorability, a critical step is to verify that what we are seeing really has a causal effect on human behavior. We test this by running a behavioral experiment that measures the memorability of images generated by our GAN, and indeed we find that our manipulations have a causal effect: navigating the GAN manifold toward images that are predicted to be more memorable actually results in generating images that are measurably more memorable in the behavioral experiment.

Our contributions include the following:

- Introducing GANalyze, a framework that uses GANs to provide a *visual definition* of image properties, like memorability and aesthetics, that we can measure but are not easy, in words, to define.
- Showing that this framework surfaces previously overlooked attributes that correlate with memorability.
- Demonstrating that the discovered transformations have a causal effect on memorability.
- Showing that GANalyze can be applied to provide visual definitions for aesthetics and emotional valence.

1.1 Related work

Generative Adversarial Networks or GANs. GANs [6] introduced a revolutionary framework to synthesize natural-looking images [8, 7, 9, 10, 7]. Among the many applications for GANs are style transfer [11], visual prediction [12], and "sim2real" domain adaptation [13]. Here, we show how they can also be applied to the problem of understanding high-level, cognitive image properties, such as memorability.

Understanding CNN representations The internal representations of a CNN can be unveiled using methods like network dissection [14, 15, 16] including for a CNN trained on memorability [5]. For instance, Khosla et al. [5] showed that units with strong positive correlations with memorable images specialized for people, faces, body parts, etc., while

those with strong negative correlations where more sensitive to large regions in landscapes scenes. Here, our framework introduces a new way of defining what memorability, and aesthetic, variability look like.

Modifying Memorability. The memorability of an image, like faces, can be manipulated using warping techniques [17]. Concurrent work has also explored using a GAN for this purpose [18]. Another approach is a deep style transfer [19] which taps into more artistic qualities. Now that GANs have reached a quality that is often almost indistinguishable from real images, they offer a powerful tool to synthesize images with different cognitive qualities. As shown here, our GANalyze framework successfully modified GAN-generated images across a wide range of image categories to produce a second generation of GAN realistic photos with different mnemonic qualities.

2 Model

2.1 Formulation

We start with a pretrained Generator G , who takes a noise vector $\mathbf{z} \in \mathbb{R}^{1 \times M}$ and a one-hot class vector $\mathbf{y} \in \{0; 1\}^{1 \times C}$ as input and generates a photo-realistic image $G(\mathbf{z}, \mathbf{y})$. Assumed is also an Assessor function A that assesses an image property of interest, in this case memorability. Our goal was to learn to transform any given noise vector \mathbf{z} of any class \mathbf{y} such that the memorability of its resulting, generated image increases (or decreases) with a certain amount α . The transformation is achieved by a Transformer function, who moves the input \mathbf{z} along a certain direction $\theta \in \mathbb{R}^{1 \times M}$ in the latent space. We express the objective as:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}, \mathbf{y}, \alpha} [(A(G(T_\theta(\mathbf{z}, \alpha), \mathbf{y})) - (A(G(\mathbf{z}, \mathbf{y})) + \alpha))^2] \quad (1)$$

Note that this is simply the MSE loss between the target memorability score, i.e. the seed image's score $A(G(\mathbf{z}, \mathbf{y}))$ increased by α , and the memorability score of the transformed clone image $A(G(T_\theta(\mathbf{z}, \alpha), \mathbf{y}))$. The scalar α acts as a metaphorical knob with which one can use to turn up or turn down memorability. The optimizing problem is $\theta^* = \operatorname{argmin}_\theta \mathcal{L}(\theta)$. The Transformer T is defined as:

$$T_\theta(\mathbf{z}, \alpha) = \mathbf{z} + \alpha\theta \quad (2)$$

Figure 2 presents a schematic of the model. Finally, note that when $\alpha = 0$, T becomes a null operation and $G(T_\theta(\mathbf{z}, \alpha), \mathbf{y})$ then equals $G(\mathbf{z}, \mathbf{y})$.

2.2 Implementation

For the results presented here, we used the Generator of BigGAN [7], which generates state-of-the art GAN images and is pretrained on ImageNet [20]. The Assessor was implemented as MemNet [5], a CNN predicting image memorability. Note, however, that training our model with different Generators or different Assessors can easily be achieved by substituting the respective modules. We discuss an Assessor for image aesthetics in Section 4. Furthermore, we present additional results for implementations with a StyleGAN [8] Generator in the supplementary materials.

To train our model and find θ^* , we built a training set by randomly sampling 400K \mathbf{z} vectors from a standard normal distribution truncated to the range $[-2, 2]$. Each \mathbf{z} was accompanied by an α value, randomly drawn from a uniform distribution between -0.5 and 0.5, and a randomly chosen \mathbf{y} . We used a batch size of 4 and an Adam optimization procedure.

In view of the behavioral experiments (see Section 3), we restricted the test set to 750 randomly chosen ImageNet classes and two \mathbf{z} vectors per class. Each \mathbf{z} vector was then paired with five different α values: $[-0.2, -0.1, 0, 0.1, 0.2]$. Note that this includes an α of 0, representing the original image $G(\mathbf{z}, \mathbf{y})$. Finally, the test set consisted of 1.5K sets of five images, or 7.5K test images in total.

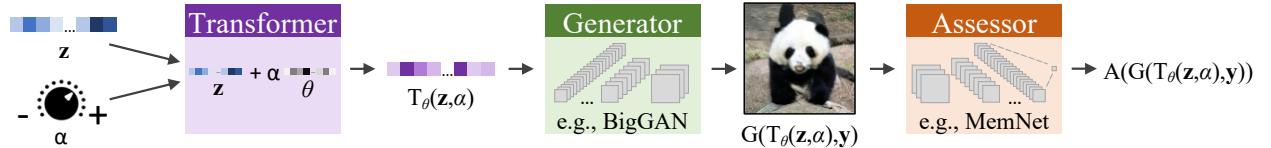


Figure 2: **Schematic of the model.** The model learns how to transform a \mathbf{z} vector such that when fed to a Generator, the resulting image's property of interest changes. The transformation is achieved by the *Transformer*, who moves the \mathbf{z} vector along a learned direction, θ , in the Generator's latent space. The property of interest (e.g., memorability) is predicted by an *Assessor module* (e.g., MemNet). Finally, α acts as knob to set the degree of change one wants to achieve in the Assessor value (e.g., MemNet score). It tells the *Transformer* how far exactly to move along θ .

3 Experiments

3.1 Model validation

Did our model learn to navigate the latent space such that it can increase (or decrease) the Assessor score of the generated image with positive (or negative) α values?

Figure 3.A suggests the model learned. The mean MemNet score of test set images increases with every increment of α . To test this formally, we fitted a linear mixed-effects regression model to the data and found a (unstandardized) slope (β) of 0.68 (95%CI = [0.66, 0.70], $p < 0.001$), confirming that the Memnet score increases significantly with α .

3.2 Emerging factors

We observe that the model can successfully change the memorability of an image, given its \mathbf{z} vector. Next, we ask which image factors it altered to achieve this. The answer to this question can provide further insight into what the Assessor has learned about the to-be-assessed image property, in this case what MemNet has learned about memorability. From a qualitative analysis of the test set (examples shown in Figures 4, S2, and S3), a number of candidate factors stand out.

First, MemNet assigns higher memorability scores when the **size of the object** (or animal) in the image is larger, as our model is in many cases zooming in further on the object with every increase of α .

Second it is **centering** the subject in the image frame.

Third, it seems to strive for **square** or **circular** shapes in classes where it is realistic to do so (e.g., snake, cheeseburger, necklace, and espresso in Figure 4).

Fourth, it is often **simplifying** the image from low to high α , by reducing the clutter and/or number of objects, such as in the cheeseburger or flamingo, or by making the background more homogeneous, as in the snake example (see Figure 4).

A fifth observation is that the subject's **eyes** sometimes become more pronounced and expressive, in particular in the dog classes (see Figure 1).

Sixth, one can also detect color changes between the different α conditions. Positive α 's often produce **brighter** and more **colorful** images, and negative α 's often produce darker images with dull colors. Finally, for those classes where multiple object hues can be considered realistic (e.g., the bell pepper and the necklace in Figure 1 and Figure 4), the model seems to prefer a **red** hue.

To verify our observations, we quantified the factors listed above for the images in the test set (except for "expressive eyes", which is more subjective and harder to quantify). Brightness was measured as the average pixel value after transforming the image to grayscale. For colorfulness, we used the metric proposed by [21], and for redness we computed the normalized number of red pixels. Finally, the entropy of the pixel intensity histogram was taken as proxy for simplicity. For the remaining three factors, a pretrained Mask R-CNN [22, 23] was used to generate an instance-level

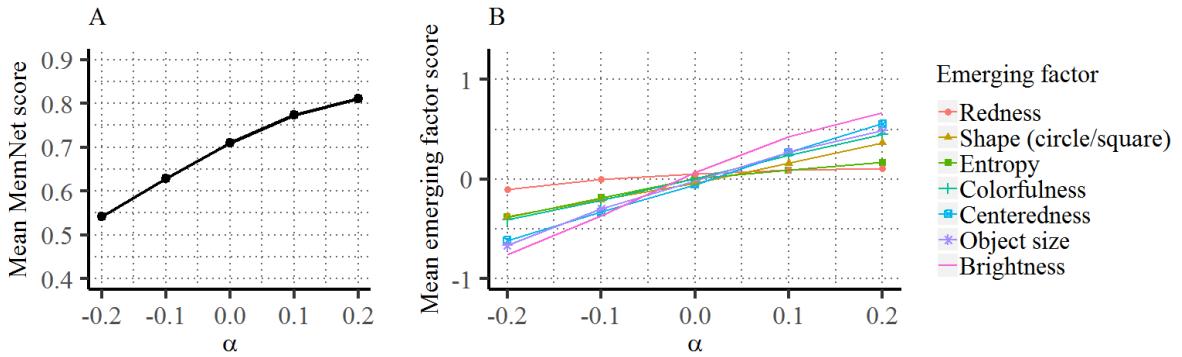


Figure 3: **Model results.** A) Graph shows the mean MemNet score across the images in every α condition. Our model successfully learned how to modify a GAN image to decrease (negative α) or increase (positive α) its MemNet score. B) List of emerging factors potentially underlying the effect observed in (A), and graph of how they change in function of α . The factors emerged from visualizations generated by the GANalyze framework (examples shown in Figures 4, S2, and S3). Emerging factor scores were first normalized and then averaged per α condition.

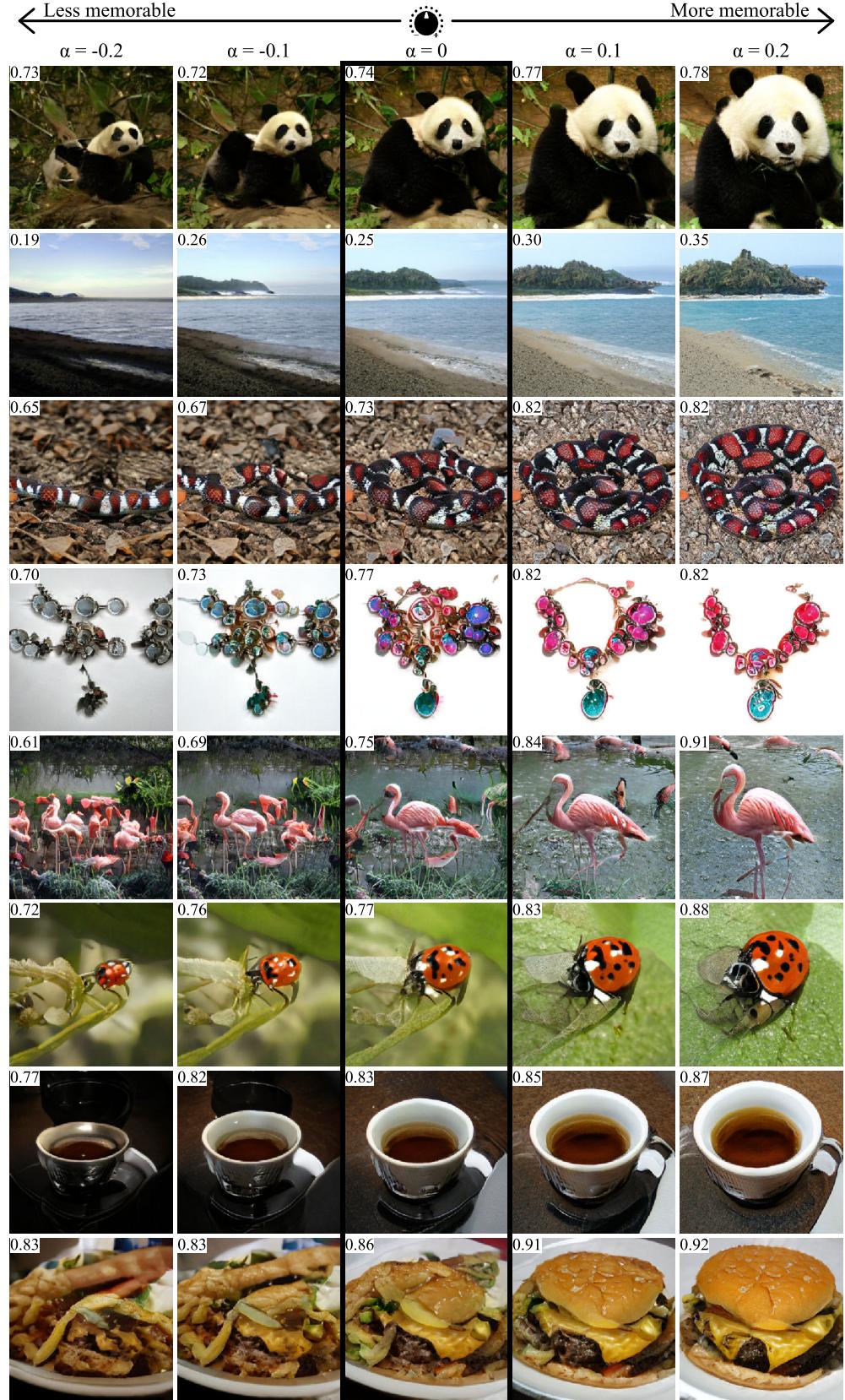


Figure 4: **Examples of generated images** along the memorability dimension. The middle column represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of clone images more or less memorable.

segmentation mask of the subject. To capture object size, we calculated the difference in the mask’s area (normalized number of pixels) as the step size α varied. To measure centeredness, we computed the deviation of the mask’s centroid from the center of the frame. Finally, we calculated the length of minor and major axes of an ellipse that has the same normalized second central moments as the mask, and used their ratio as a metric of squareness. Figure 3.B shows that the emerging factor scores increase with α .

3.3 Realness

While BigGAN achieves state-of-the-art to generate highly realistic images, there remains a certain variability in the “realness” of the generated images. How best to evaluate the realness of a set of GAN-images is still an open question. Below, we discuss two automatically computed realness measures and a human measure in relation to our data. We discuss an additional human measure, based on a different task, in the supplementary materials.

3.3.1 Automatic measures

In Figure 5.A, we plot two popular automatic measures in function of α : the Fréchet Inception Distance (FID) [24] and the Inception Score (IS) [25]. A first observation is that the FID is below 40 in all α conditions. An FID as low as 40 already corresponds to reasonably realistic images. Thus the effects of our model’s modifications on memorability are not explained by making the images unrealistic. But we do observe interesting differences in FID- and IS-differences related to α , suggesting that more memorable images have more interpretable semantics.

3.3.2 Human measure

In addition to the two automatic measures, we conducted an experiment to collect human realness scores. The experiment consisted of a two-alternative forced choice (2AFC) task, hosted on Amazon Mechanical Turk (AMT), in which workers had to discriminate GAN-images from real ones. Workers were shown a series of pairs, consisting of one GAN-image and one real image. They were presented side by side for a duration of 1.6 s. Once a pair had disappeared off the screen, workers pressed the j-key when they thought the GAN-image was shown on the right, or the f-key when they thought it was shown on the left. The position of the GAN-image was randomized across trials. The set of real images used in this experiment was constructed by randomly sampling 10 real ImageNet exemplars per GAN-image class. The set of GAN-images was the same as the one quantified on memorability in Section 3.4. A GAN-image was randomly paired with one of the 10 real images belonging to the same class. Each series consisted of 100 trials, of which 20 were vigilance trials. For the vigilance trials, we generated GAN-images from z vectors that were sampled from the tails of a normal distribution (to make them look less real). For a worker’s first series, we prepended 20 trials with feedback as practice (not included in the analyses). Workers could complete up to 17 series, but were blocked if they scored less than 65% correct on the vigilance trials. Series that failed this criterion were also excluded from the analyses. The pay rate equaled \$0.50 per completed series. On average, each of our test images was seen by 2.76 workers, meaning 4137 data points per α condition.

We did not observe differences in task performance between different α (see Figure 5.B). Indeed, a logistic mixed-effects regression fitted to the raw, binary data (correct/incorrect) did not reveal a statistically significant regression weight for α ($\beta = -0.08, 95\%CI = [-0.33, 0.18], p = 0.55$). In other words, the model’s image modifications did

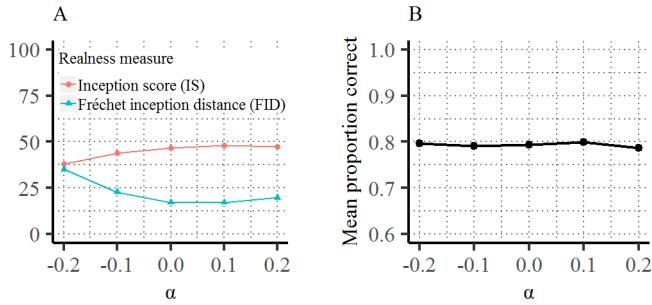


Figure 5: Realness measures as a function of α . A) Two popular automatic measures for evaluating the realness of a set of GAN images. Note that lower FID values indicate higher realness. B) Human fakeness discriminability, measured as the mean proportion correct in a 2AFC-task in which AMT workers had to discriminate GAN-images (fake) from real photographs.

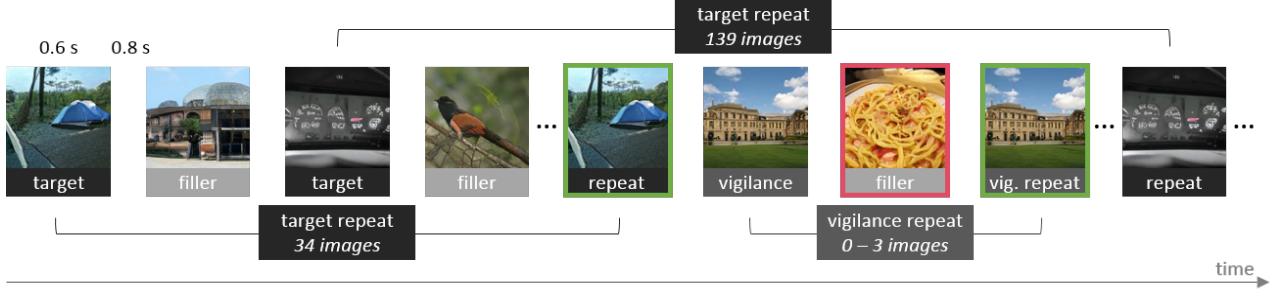


Figure 6: **Schematic of the visual memory game.** Each image is shown for 600 ms, with a blank interstimulus interval of 800 ms. Workers are asked to respond whenever they recognize a repeat of a previously shown image. For a correct response, the frame around the image briefly turns green. A red frame, on the other hand, indicates a mistake.

not affect workers’ ability to correctly identify the fake image, indicating that perceptually, the image clones of a seed image did not differ in realness.

3.4 Do our changes causally affect memory?

In addition to the MemNet scores, is our model also successful at changing the probability of an image being recognized by participants in an actual memory experiment?

We tested people’s memory for the images of a test set (see Section 2.2) using a repeat-detection visual memory game, which was hosted on AMT (see Figure 6). [4, 5]. AMT workers watched a series of one image at the time and had to press a key whenever they saw a repeat of a previously shown image. Each series consisted of 215 images, shown each for 600 ms with a blank interstimulus interval of 800 ms. Sixty images were targets, sampled from our test set, and repeated after 34 to 139 intervening images. The remaining images were either filler or vigilance images and were sampled from a separate set. This set was created with 10 \mathbf{z} vectors per class and the same five α values as the test set: $[-0.2, -0.1, 0, 0.1, 0.2]$, making a total of 37.5K images. Filler images were only presented once and ensured spacing between a target and its repeat. Vigilance images were presented twice, with 0 to 3 intervening images in-between the two presentations. The vigilance repeats constituted easy trials to keep workers attentive. Care was taken to ensure that a worker never saw more than one $G(T_\theta(\mathbf{z}, \alpha), \mathbf{y})$ for a given \mathbf{z} . Workers could complete up to 25 series, but were blocked if they missed more than 55% of the vigilance repeats in a series or made more than 30% false positives. Series that failed this were excluded from the analyses. The pay rate was \$0.50 per completed series. On average, a test image was seen by 3.16 workers, with 4740 data points per α condition.

Workers could either recognize a repeated test image (hit, 1), or miss it (miss, 0). Figure 7.A shows the hit rate across all images and workers. The hit rate increases with every step of α . Fitting a logistic mixed-effects regression model to the raw, binary data (hit/miss), we found that the predicted log odds of image being recognized increase with 0.19 for an increase in α of 0.01 ($\beta = 1.92, 95\%CI = [1.71 - 2.12], p < 0.001$). This shows that our model can successfully navigate the BigGAN latent space in order to make an image more (or less) memorable to humans.

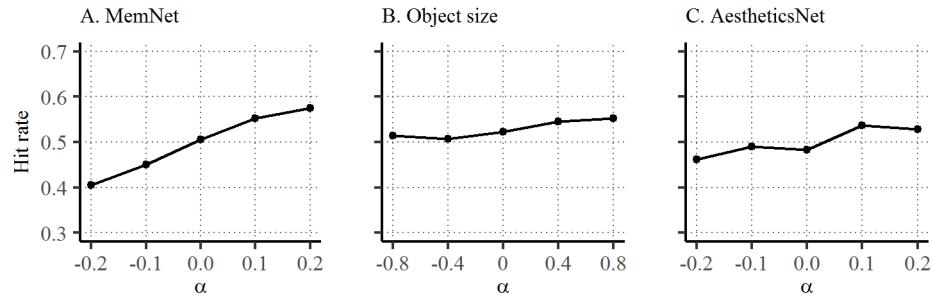


Figure 7: **Human memory performance** for images modified according to different Assessors: A) MemNet, B) Object size and C) AestheticsNet. Performance is measured as the hit rate across all images and workers in the memory game for each property.

Factor	Log Odds	CI	p	Tjur's D
Brightness	0.28	[0.24, 0.32]	< 0.001	0.066
Centeredness	0.24	[0.19, 0.29]	< 0.001	0.059
Colorfulness	0.17	[0.14, 0.21]	< 0.001	0.054
Entropy	0.03	[-0.04, 0.10]	0.441	0.062
Redness	0.06	[0.00, 0.12]	0.042	0.055
Shape	0.19	[0.14, 0.24]	< 0.001	0.060
Object size	0.32	[0.27, 0.37]	< 0.001	0.050
α	1.92	[1.71, 2.12]	< 0.001	0.074

Table 1: **Relation between emerging factors and human memory performance.** We show the output of logistic mixed-effects regressions. From left to right: the regression weight, the confidence interval (CI) for that weight, the p -value for statistical significance, and Tjur's coefficient of discrimination (D), being the regression model's goodness of fit [26]. The emerging factor values were normalized before running the regression models.

3.4.1 Emerging factors

Given human memory data for images modified for memorability, we evaluate how the images' emerging factor scores relate to their likelihood of being recognized. We fitted mixed-effects logistic regression models, each with a different emerging factor as the predictor, see Table 3.4.1. Except for entropy, all the emerging factors show a significant, positive relation to the likelihood of a hit in the memory game, but none fit the data as well as the model's α . This indicates that a single emerging factor is not enough to fully explain the effect observed in Figure 7.A. Note that the emerging factor results are correlational and the factors are intercorrelated. This makes it hard to draw conclusions about which individual factors truly causally affect human memory performance. As an example of how this can be addressed within the GANalyze framework, we conducted an experiment focusing on the effect of one salient emerging factor: **object size**. As seen in Figure 4, more memorable images tend to center and enlarge the object class.

We trained a version of our model with an Object size Assessor, instead of the MemNet Assessor. This is the same Object size Assessor used to quantify the object size in the images modified according to MemNet (e.g., for the results in Figure 3.B), now teaching the Transformer to perform "enlarging" modifications. After training with 161750 \mathbf{z} vectors, we generated a test set as described in Section 2.2, except that we used a different set of α 's: $[-0.8, -0.4, 0, 0.4, 0.8]$. We chose these values to qualitatively match the degree of object size changes achieved by the MemNet version of the model. Figure 8.A visualizes the results achieved on the test set. The model successfully enlarges the object with increasing alpha's, as confirmed by a linear mixed-effects regression analysis ($\beta = 0.07, 95\%CI = [0.06, 0.07], p < 0.001$). Figure 10 shows example images generated by that model, after having been trained with 161750 \mathbf{z} vectors. A comparison with images modified according to MemNet suggests that the latter model was doing more than just enlarging the object.

To study how the new size modifications affect memorability, we generated a new set of images (7.5K targets, 37.5K fillers) with α 's $[-0.8, -0.4, 0, 0.4, 0.8]$. The new images were then quantified using the visual memory game (on average 2.36 data points per image and 3540 per α condition). Figure 7.B shows the results. Memory performance increases with α , as confirmed by a logistic mixed-effects analysis ($\beta = 0.11, 95\%CI = [0.06, 0.18], p < 0.001$, although mostly for positive α values).

4 Other properties

As mentioned in Section 2.2, the proposed method can be applied to other image properties, simply by substituting the Assessor module. To show our framework can generalize, we trained a model for aesthetics, using Kong et al's [27] CNN (hereinafter referred to as AestheticsNet) as the Assessor. In addition, we also trained a model for emotional valence. Emotional valence refers to the extent to which the emotions evoked by an image are experienced as positive (or negative). For this property, we trained our own Assessor by fine-tuning a ResNet50 model [28], pretrained on the Moments database [29], to the Cornell Emotion6 Image Database [30]. We refer to this Assessor as EmoNet. Finally, we generated a test set for each of the two new models, like we did for the memorability model.

Figure 8.B shows the average AestheticsNet scores per α condition. The scores significantly increase with α , as evidenced by the results of a linear mixed-effects regression ($\beta = 0.72, 95\%CI = [0.70, 0.74], p < 0.001$). We can successfully train the model to increase (or decrease) an image's aesthetic score as shown in Figure 9 (left) and Figure S4. Similarly, Figure 8.C shows the average EmoNet scores per α condition. Here too, the scores significantly

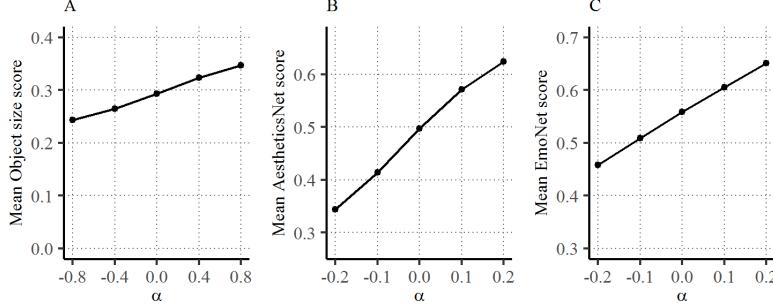


Figure 8: **Model results for additional Assessors.** A) Graph shows the mean Object size Assessor score across images in every α condition. A different set of α values was chosen to qualitatively match the degree object size changes achieved by the model trained with the MemNet Assessor. B) Graph shows the mean AestheticsNet score across the images in every α condition. C) Graph shows the mean EmoNet score across the images in every α condition

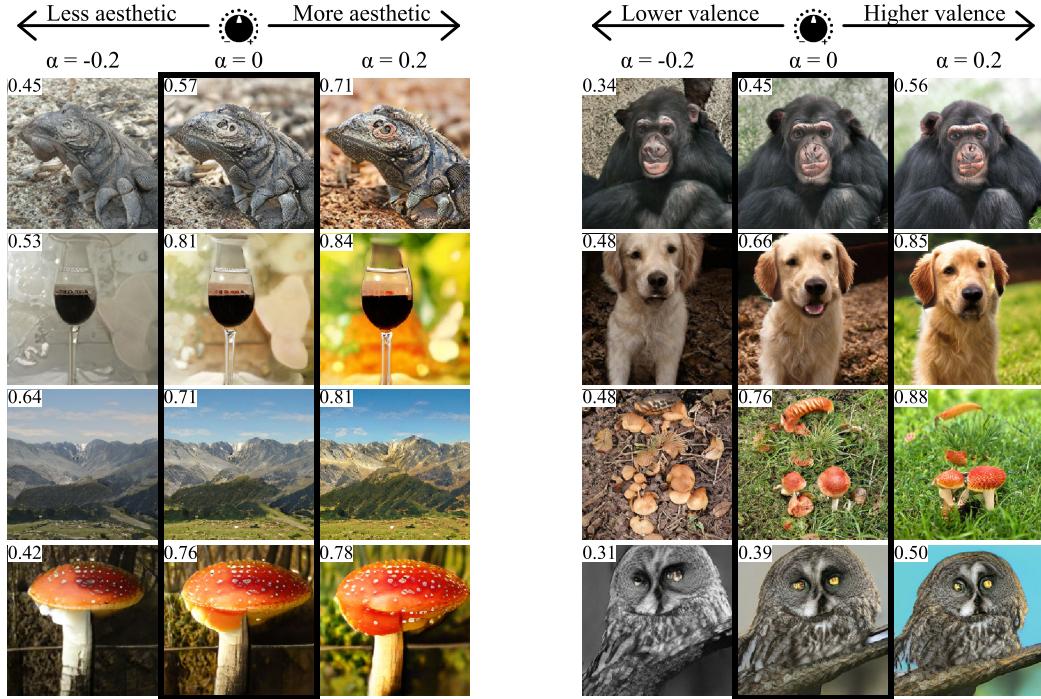


Figure 9: **Examples of generated images** along the aesthetics dimension (left) and the emotional valence dimension (right). Each middle column represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of clone images scoring higher or lower on the respective dimension according to the Assessor. The images' Assessor scores are presented in their top left corner.

increase with α ($\beta = 0.44$, $95\%CI = [0.43, 0.45]$, $p < 0.001$). Example visualizations generated by this model are presented in Figure 9 (right) and Figure S5.

Based on a qualitative inspection of such visualizations, we observed that the aesthetics model is modifying factors like depth of field, color palette, and lighting, suggesting that the AestheticsNet is sensitive to those factors. Indeed, the architecture of the AestheticsNet includes attribute-adaptive layers to predict these factors, now highlighted by our visualizations. The emotional valence model often averts the subject's gaze away from the "camera" when decreasing valence. To increase valence, it often makes images more colorful, introduces bokeh, and makes the skies more blue in landscape images. Finally, the teddy bear in Figure 1 (right) seems to smile more. Interestingly, the model makes different modifications for every property (see Figure 10), suggesting that what makes an image memorable is different from what makes it aesthetically pleasing or more positive in its emotional valence.

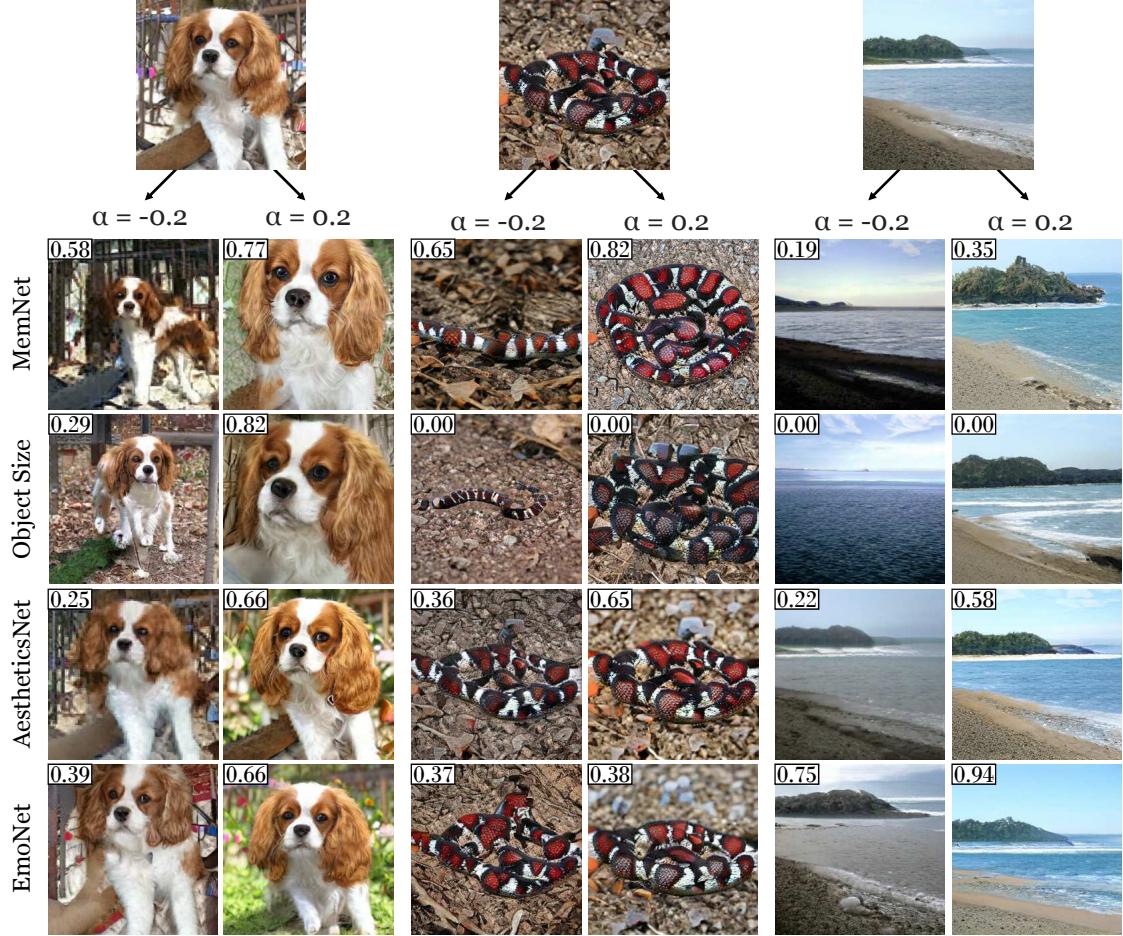


Figure 10: **Comparison of examples generated according to different Assessors.** The top row represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of images with a higher or lower Assessor value. The respective Assessor values are indicated in the top left corner. Note that for object size, we used a different α range: $\{-0.8, 0.8\}$.

A final question we asked is whether an image modified to become more (less) aesthetic also becomes more (less) memorable? To test this, we quantified the images of the aesthetic test set on memorability by presenting them to workers in the visual memory game (we collected 1.54 data points per image and 2306 data points per α condition). Figure 7.C shows the human memory performance in function of an α that is tuning aesthetics. A logistic mixed-effects regression revealed that with an 0.1 increase in the aesthetics α , the predicted log odds of an image being recognized increase with 0.07 ($\beta = 0.72$, 95%CI = [0.44, 1.00], $p < 0.001$). While modifying an image to make it more aesthetic does increase its memorability, the effect is rather small, suggesting that memorability is more than only aesthetics and that our model was right to modify memorability and aesthetics in different ways.

5 Conclusion

We introduce GANalyze, a framework that shows how a GAN-based model can be used to visualize what another model (i.e. CNN as an Assessor) has learned about its target image property. Here we applied it to memorability, yielding a kind of “visual definition” of this high-level cognitive property, where we visualize what it looks like for an image to become more or less memorable. These visualizations surface multiple candidate features that may help explain why we remember what we do. Importantly, our framework can also be generalized to other image properties, such as aesthetics or emotional valence: by replacing the Assessor module, the framework allows us to explore the visual definition for any property we can model as a differentiable function of the image. We validated that our model successfully modified GAN images to become more (or less) memorable via a behavioral human memory experiment on manipulated images.

GANalyze's intended use is to contribute to the scientific understanding of otherwise hard to define cognitive properties. Note that this was achieved by modifying images for which the encoding into the latent space of the GAN was given. In other words, it is currently only possible to modify seed images that are GAN-images themselves, not user-supplied, real images. However, should advances in the field lead to an encoder network, this would become possible and it would open applications in graphics and education, for example, where selected images can be made more memorable. One should also be wary, though, of potential misuse, especially when applied to images of people or faces. Note that the BigGAN [7] generator used here was trained on ImageNet categories [20] which only occasionally include people, and that it does not allow to render realistically looking people. Nevertheless, with generative models yielding ever more realistic output, an increasingly important challenge in the field is to develop powerful detection methods to allow us to reliably distinguish generated, fake images from real ones [31][32][33].

6 Acknowledgments

This work was partly funded by NSF award 1532591 in Neural and Cognitive Systems (to A.O), by a fellowship (Grant 1108116N) and a travel grant (Grant V4.085.18N) awarded to Lore Goetschalckx by the Research Foundation - Flanders (FWO).

References

- [1] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.
- [2] Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25(2):207–222, 1973.
- [3] Tony W Buchanan and Ralph Adolphs. The role of the human amygdala in emotional modulation of long-term declarative memory. *Advances in Consciousness Research*, 44:9–34, 2002.
- [4] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, jul 2014.
- [5] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018.
- [8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.
- [9] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
- [10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.
- [11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [12] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.
- [13] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. *CoRR*, abs/1709.07857, 2017.
- [14] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [15] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

- [16] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR)*, May 2015.
- [17] Aditya Khosla, Wilma A. Bainbridge, Antonio Torralba, and Aude Oliva. Modifying the memorability of face photographs. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.
- [18] Oleksii Sidorov. Changing the image memorability: From basic photo editing to gans. *CoRR*, abs/1811.03825, 2018.
- [19] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memorable?: A deep style transfer approach. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, pages 322–329, 2017.
- [20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, December 2015.
- [21] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–96. International Society for Optics and Photonics, 2003.
- [22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [23] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. <https://github.com/facebookresearch/maskrcnn-benchmark>, 2018. Accessed: Mar 2019.
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- [25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.
- [26] Tue Tjur. Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, 63(4):366–372, 2009.
- [27] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 662–679, Cham, 2016. Springer International Publishing.
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [29] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [30] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 860–868, 2015.
- [31] Drew Harwell. Top ai researchers race to detect ‘deepfake’ videos: ‘we are outgunned’. *The Washington Post*, Jun 2019.
- [32] Karen Hao. Deepfakes have got congress panicking. this is what it needs to do. *MIT Technology Review*, Jun 2019.
- [33] Drew Harwell. Top ai researchers race to detect ‘deepfake’ videos: ‘we are outgunned’. *The Washington Post*, Jun 2019.

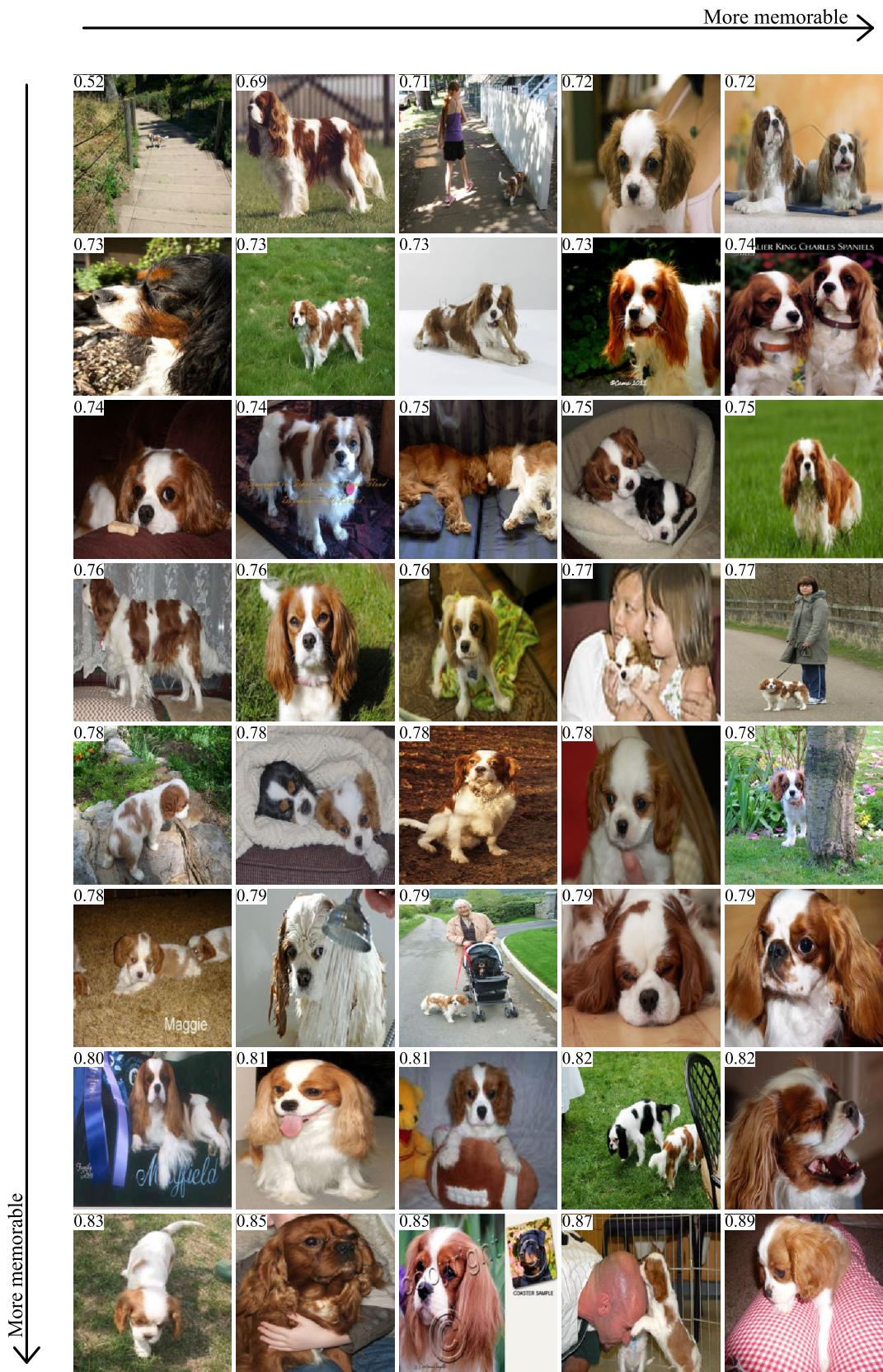


Figure S1: **Real images sorted on their MemNet score.** As opposed to GANalyze, this is a non-parametric way of visualizing what it looks like for images to become more memorable.

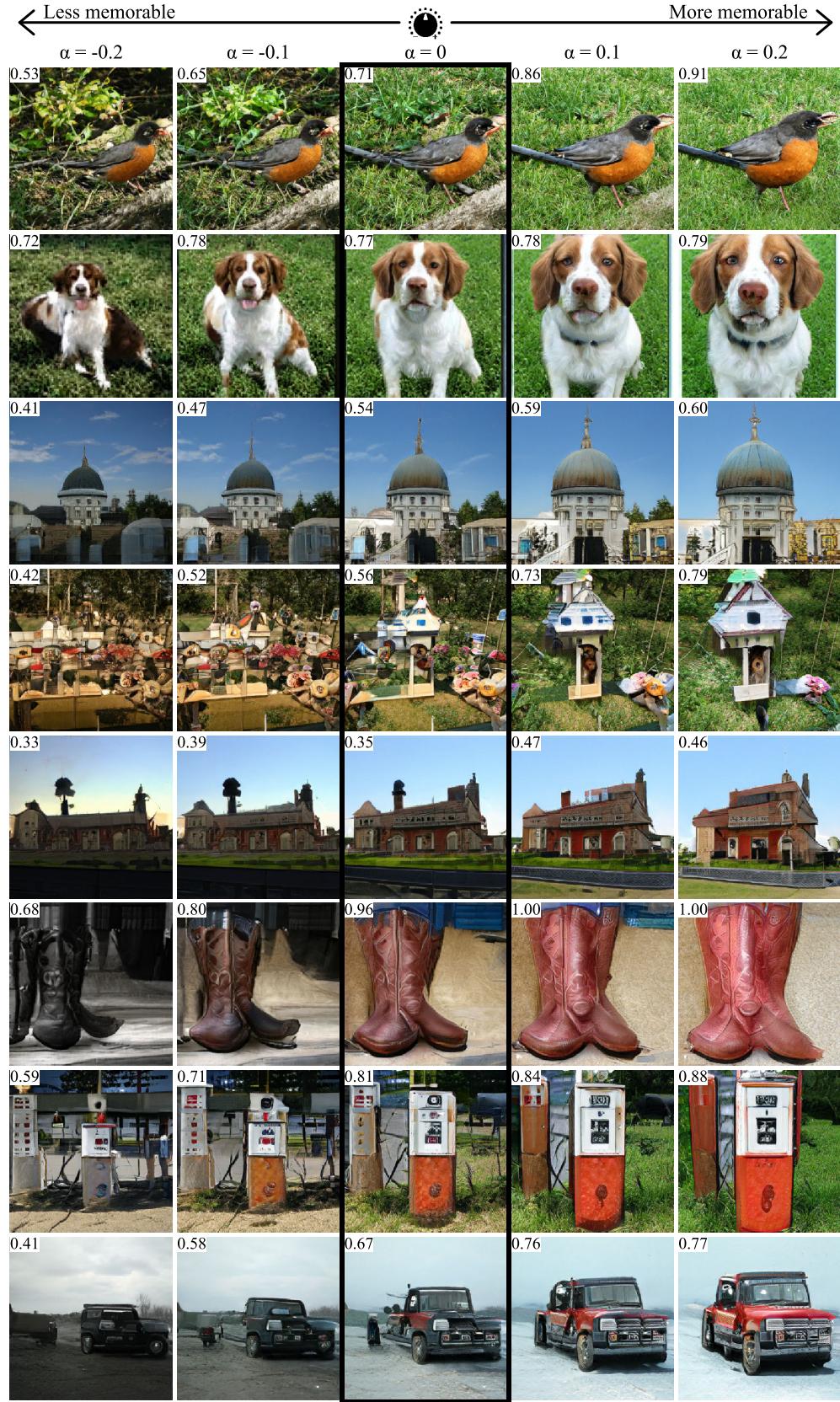


Figure S2: **More examples of generated images** along the memorability dimension. The middle column represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of clone images more or less memorable.

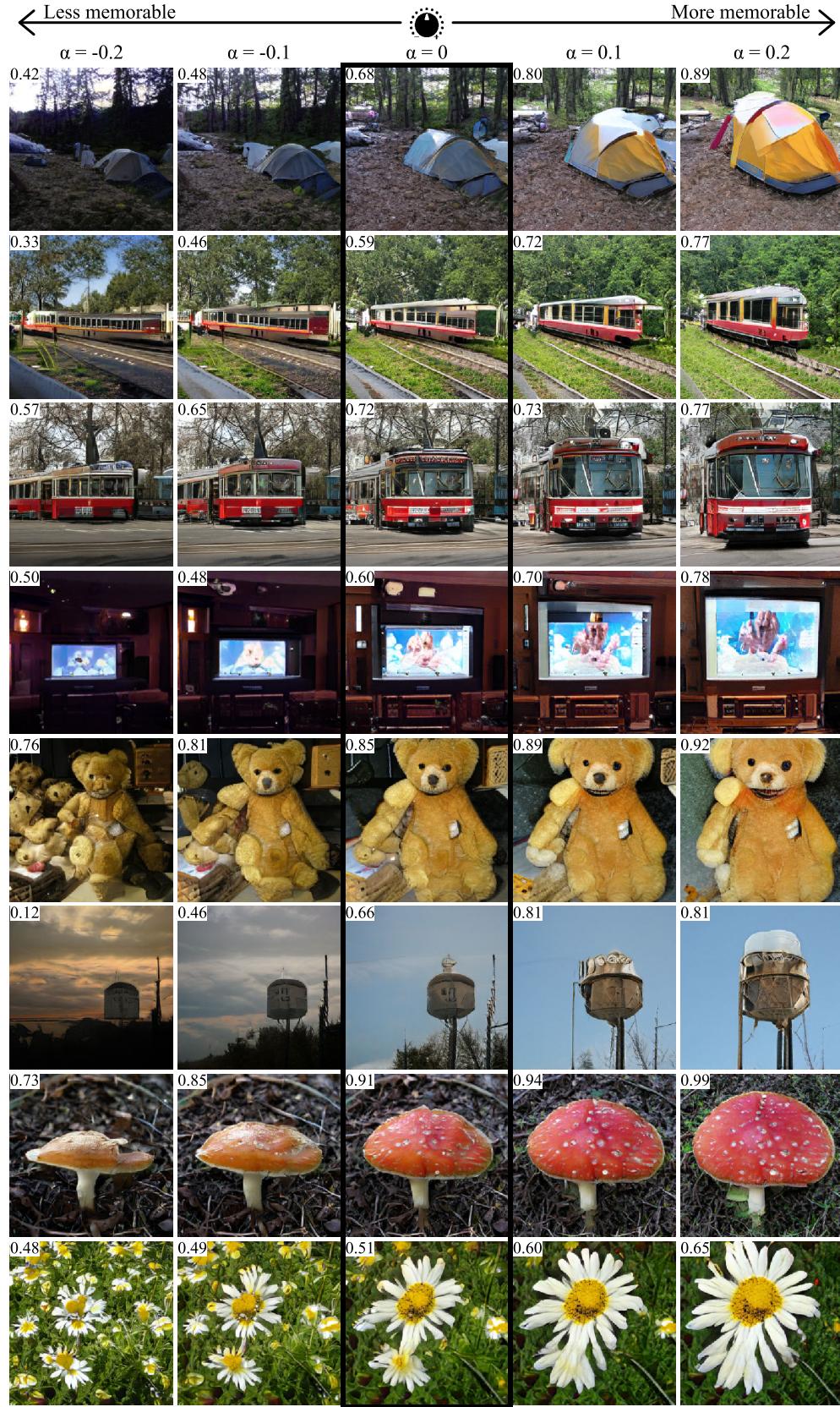


Figure S3: **More examples of generated images** along the memorability dimension. The middle column represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of clone images more or less memorable.

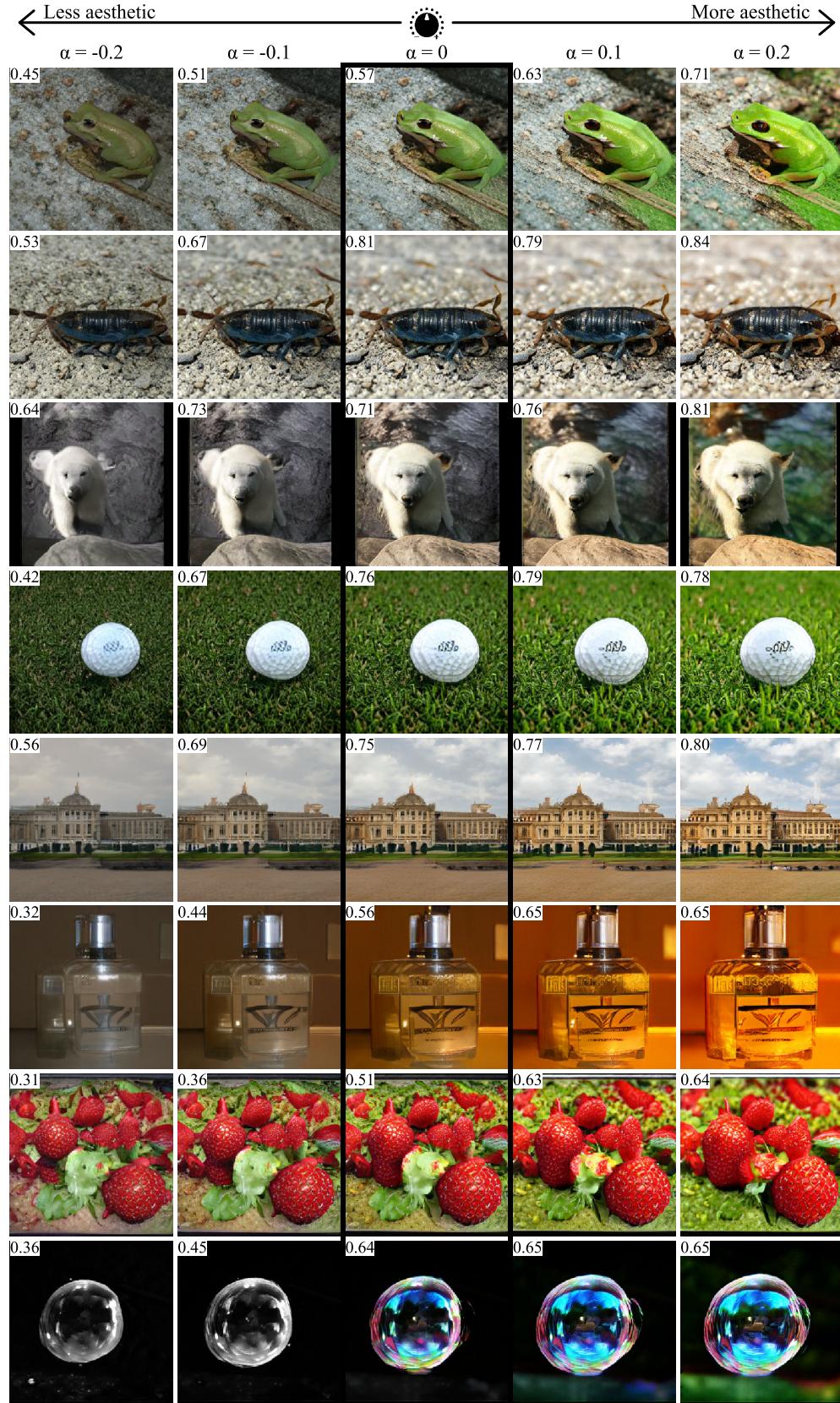


Figure S4: **More examples of generated images** along the aesthetics dimension. The middle column represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of clone images more or less aesthetic.

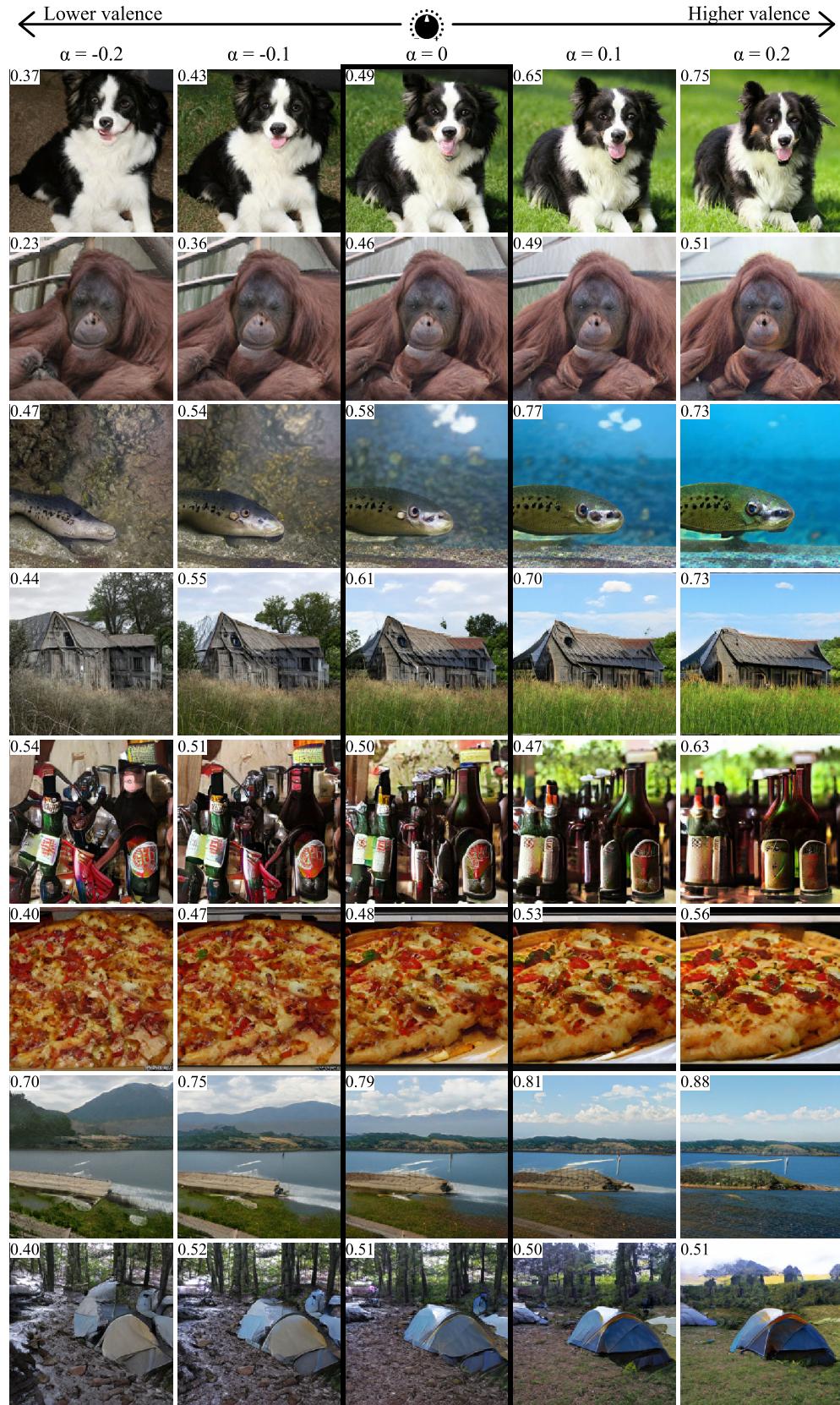


Figure S5: **More examples of generated images** along the emotional valence dimension. The middle column represents $G(\mathbf{z}, \mathbf{y})$, the generated image serving as the original seed to create a series of clone images higher or lower in emotional valence.