

Supplementary: AI Art Neural Constellation: Revealing the Collective and Contrastive State of AI-Generated and Human Art

Faizan Farooq Khan*
KAUST

Diana Kim*
KAUST

Divyansh Jha
KAUST

Youssef Mohamed
KAUST

Hanna H Chang
KAUST

Ahmed Elgammal
Rutgers University

Luba Elliott
ELLUBA

Mohamed Elhoseiny
KAUST

Supplementary Overview

This supplementary document is organized as follows

- Section **A** describes the data collection approach.
- Section **B** details style classification and time correlation experiment.
- Section **C** provides example of the distribution example of the Wölfflin's Principles.
- Section **D** provides additional analysis of the general art principles.
- Section **E** shows qualitative examples in turing test.
- Section **F** details the qualitative analysis of likability.
- Section **G** details the qualitative analysis of emotions.
- Section **H** shows art AI-generated examples with detailed explanations valued around the distribution of human art.

* indicates equal contribution.

WHow does this painting make you feel? Describe why! (Click to collapse)

- How does this painting make you **primarily** feel? (choose one button)
- Give a detailed description (**at least 8 words**) about **WHY** you feel like this, based on **SPECIFIC** details of the painting.

Examples of GOOD descriptions:

- "the sky looks gloomy and the shadows are scary"
- "the red marks on the table look like drops of blood" (we like analogies!)
- "the blue of the lake contrasts well with the orange hats of the men"

(a) Do not use **uninformative descriptions**, such as "it's fun", "nice colors", i.e. NOT explaining WHY in a specific manner.


(b) Do not start your sentence like "I feel..."


(c) Do not do more than ~30 HITs in a batch.


(d) If you feel "nothing" or "bored" you still have to explain **WHY**.


If you are not a **proficient English** speaker, please don't accept this HIT.


Thanks a lot for your hard work!


Amusement



Awe



Contentment



Excitement




Anger



Disgust


Fear


Sadness


Something Else

Art Work (below)



Q1) How much do you like this image?

1 (very much dislike) 2 (dislike) 3 (neutral) 4 (like) 5 (very much like)

Q2) Do you think this image was created by artist or generated by computer?

1 (artist) 2 (computer)

Figure 1. User interfaces of the emotion experiment (top) and likability experiment(bottom).

A. Data Collection Interface

We present the user interface utilized for collecting crucial data, including Wölfflin’s Principles, emotional responses, and Turing test data. Figure 1 and Figure 2 visually illustrate the user interface for these tasks.

The user interface design played a significant role in ensuring the effectiveness and accuracy of data collection. For the Wölfflin’s principles task (subsection 3.2.1 of the main paper), participants were prompted to identify the underlying principles depicted in the artworks after providing detailed instructions to understand the principle. The emotional responses task (subsection 3.2.5 of the main paper) involved presenting participants with various generated art pieces and human art, prompting them to select the emotions evoked by each piece. Lastly, the Turing test data collection included presenting participants with a mix of generated and human art pieces, challenging them to distinguish between the two.

We emphasize that the user interface was designed to be intuitive and user-friendly, minimizing any potential biases that could influence the data collection process. The results obtained from this comprehensive data collection methodology

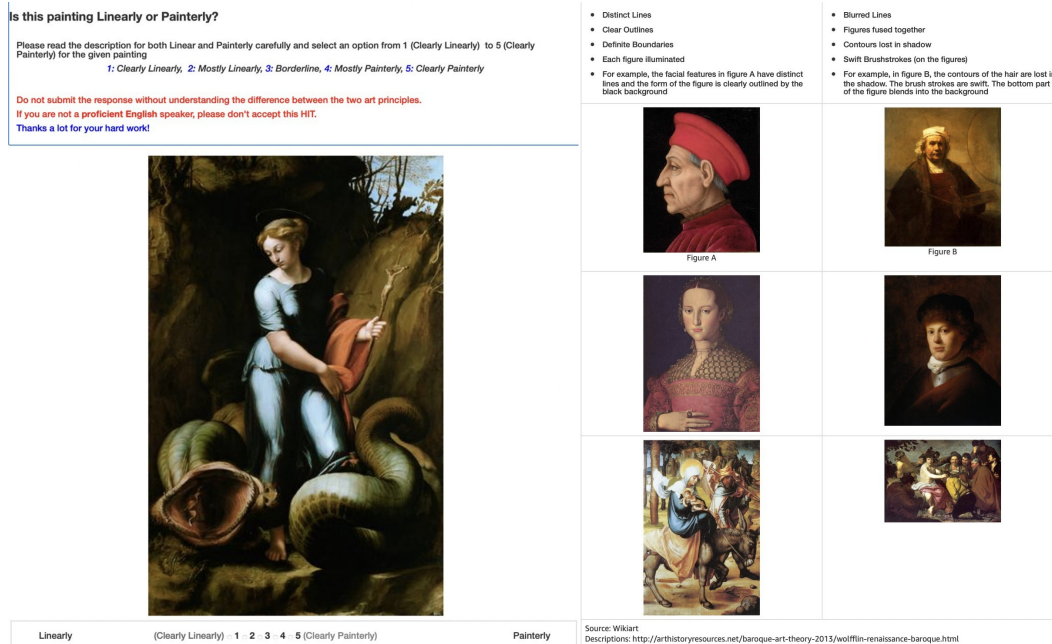


Figure 2. Data Collection Interface of Wölfflin’s principles presenting the form and examples used to train participants to classify the work as either linear or painterly.

provided valuable insights into the effectiveness and capabilities of the generative models in evoking diverse emotions and artistic principles.

B. Style Classification and Time Correlation

Architecture	trained from scratch	pre-trained & fine-tuned
ResNet50	27.80%	64.41 %
ResNet50+2	31.81%	60.00%
ResNet101	29.16%	64.13%
ResNet101+2	29.24%	64.27%
VGG16	31.48%	55.88%
VGG16+2	28.62%	56.91%
ViT-S	33.92%	60.03%
ViT-B	32.68%	55.75%
ViT-L	33.81%	56.93%

Table 1. The table displays each model’s average accuracy per class. Pretrained and fine-tuned ResNet50 architecture shows the highest art-style classification accuracy.

B.1. Style Classification

Various convolutional and transformer-based networks are trained for the task of style classification. Following the training setting in [1], we compared the two cases of (1) trained from scratch and (2) pretrained+fine-tuned from ImageNet models. For CNNs, VGG-Net16, ResNet50, ResNet101, and the other extended three ConvNet models are tested after adding the two hidden layers of 1024 and 512-dimensional nodes on the top of the original models. For transformer models, ViT-S, B, and L are considered.

These networks are trained on WikiArt. The dataset consists of more than 80,000 paintings belonging to 27-period styles ranging from the eleventh to the twenty-first century. In our experiments, the number of classes is reduced to 21 art classes by

merging similar style classes; new-realism, contemporary-realism, and realism were merged, analytical-cubism, cubism, and synthetic-cubism were merged, and action-painting was merged with abstract-expressionism. The data is split into training (80%), validation (10%), and testing sets (10%). Table 1 presents the average classification accuracy per class of the models. As expected, the pre-trained and fine-tuned networks achieved significantly better results than their counterparts trained from scratch for all networks. Pretrained and fine-tuned ResNet50 architectures achieved the highest art-style classification accuracy.

B.2. Time Correlation

The high correlation of spatial arrangements in neural nets with time is the prior knowledge [1, 3] we ground in our time analysis. As a preliminary step, we computed the PCC (Pearson’s Correlation Coefficient) of principal components of our neural net features with the year of the artwork. In Table 2, adding the two layers on all CNN models caused a decline in PCC. The reduced dimension limits the representation power for the subtle visual information in artworks, so hard to capture the smooth visual transition over time. We also noted increased correlation with discriminators of both StyleGAN-1 and StyleGAN-2 architectures with the integration of the losses from CAN and CWAN. While the discriminators in StyleGANs are trained only to follow the original distribution of human art, the creative models—CAN and CWAN—are trained to be similar to human art but also not to be close to any of human art styles. Hence, the creative loss needs to know and leverage the information of period style to find a new visual space unoccupied. This can explain the high-time correlation results in the discriminators with CAN and CWAN in Table 2. The maximum PCC is ResNet50 followed by ViT-S and ResNet101.

B.3. Time Analysis from Different Networks

As done in Section 4.4 of the main paper, we overlay the t-SNE visualization from generated images onto human images. The features used in the main paper are from ResNet50 [2], we show results on feature representations from ResNet50, ResNet101, and their corresponding +2 models(ResNet50+2 and ResNet101+2) to compare how well the +2 models separate the human art in the time dimension and how the generated art is overlaid on the human art features. For ResNet50 and 101, the high correlation with time tells that these models can separate the human artworks based on their time period, and from Figure 3, it can be seen that both these models allot generated art in the region close to modern art. The same trend is observed for the +2 models which show a slightly less correlation with the time dimension for human art.

Architecture	PCC
ResNet50	0.625
ResNet50+2	0.366
ResNet101	0.537
ResNet101+2	0.503
VGG16	0.462
VGG16+2	0.422
ViT-S	0.542
ViT-B	0.536
ViT-L	0.489
SG1 Disc.	0.184
SC1 Disc.	0.316
SG2 Disc.	0.224
SC2 Disc.	0.320
CW1 Disc.	0.385
CW2 Disc.	0.286

Table 2. The maximum absolute Pearson’s Correlation Coefficient (PCC) of the first 30 PCA components: correlation between human art’s extracted features and year of making. The maximum PCC is ResNet50 followed by ViT-S and ResNet101.

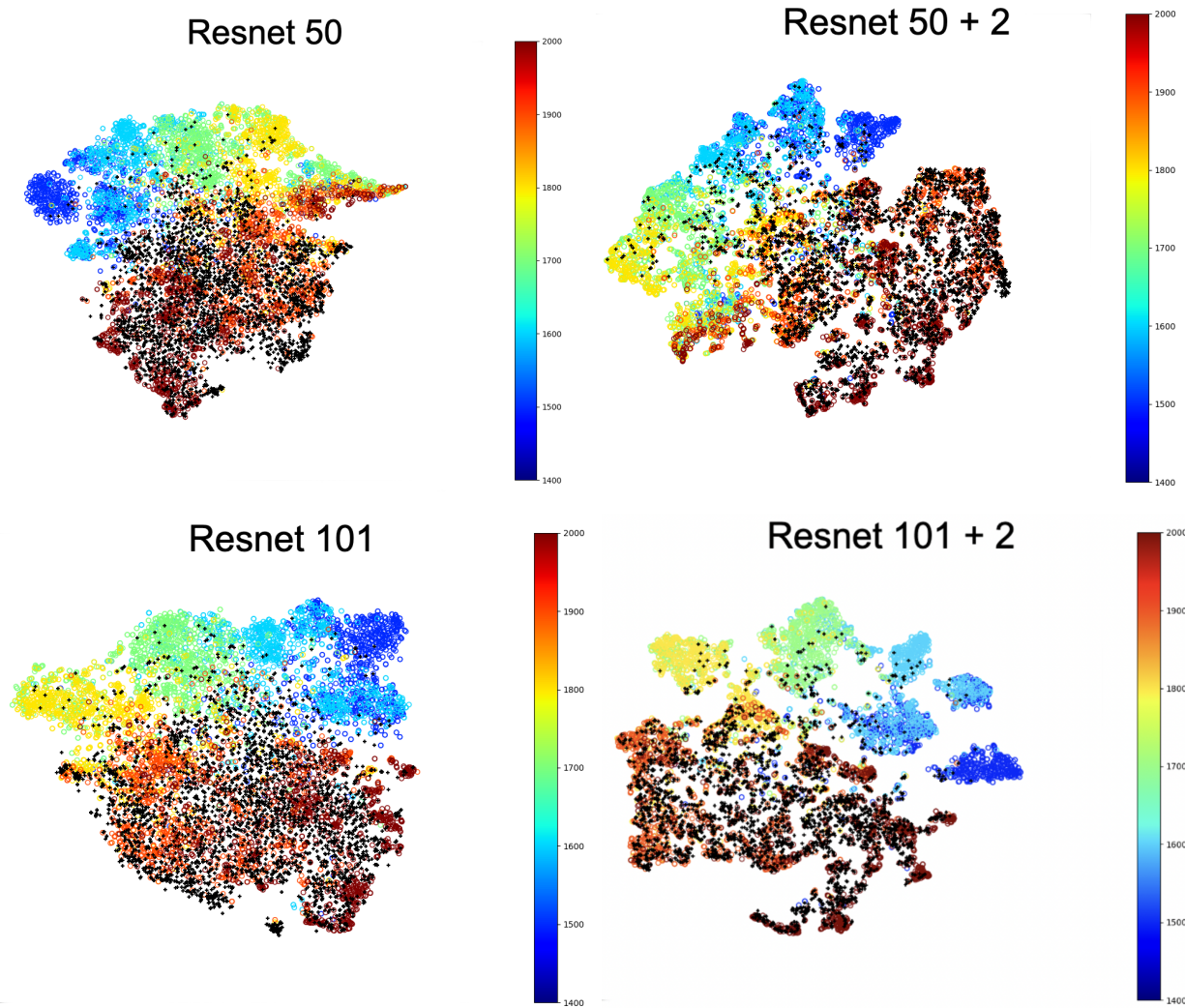


Figure 3. For ResNet50, ResNet50+2, ResNet101, and ResNet101+2, the t-SNE visualizations of features of generated art are overlaid over human art. The black markers are points corresponding to generated art.

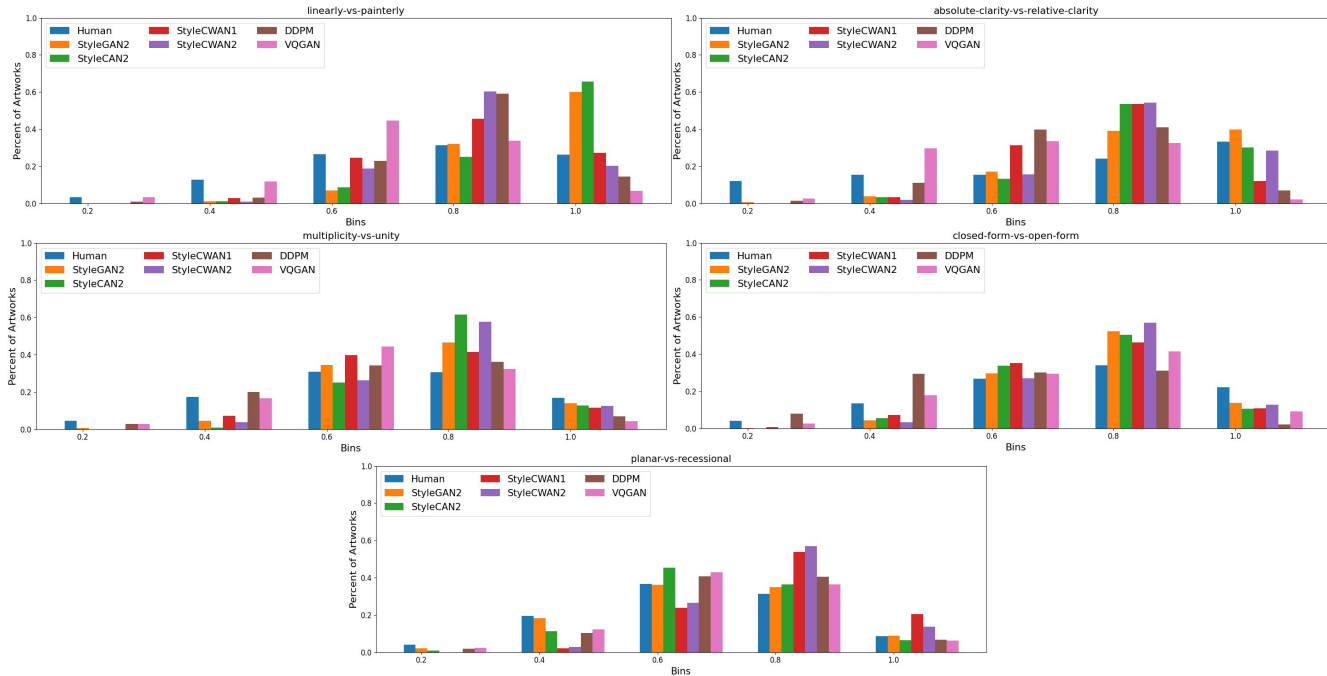


Figure 4. The distribution of all five Wölfflin’s principles values: the bins represent the values from the left value to the right value. For example, the 0.2 bin represents the percentage of artworks that have values between 0 and 0.2. In this figure, AI-generated art is less uniform across the principle concepts than human art. Machine artworks are highly populated on the right side concepts more in general.

C. Wölfflin’s Principles

In Figure 4, we provide a detailed version of Figure 2 from the main paper for the analysis done in sub-section 4.1 of the main paper. We also present illustrative examples in Figure 5 showcasing the highest values obtained for specific Wölfflin’s principles across different generative models. These examples serve as visual representations of generative art’s ability to embody and emphasize particular artistic principles.

Each example in Figure 5 highlights a distinct Wölfflin’s principle, effectively demonstrating how various generative models excel in incorporating and expressing different artistic attributes. By presenting these examples, we aim to provide a clear visual understanding of how each generative model can capture and manifest specific artistic characteristics.

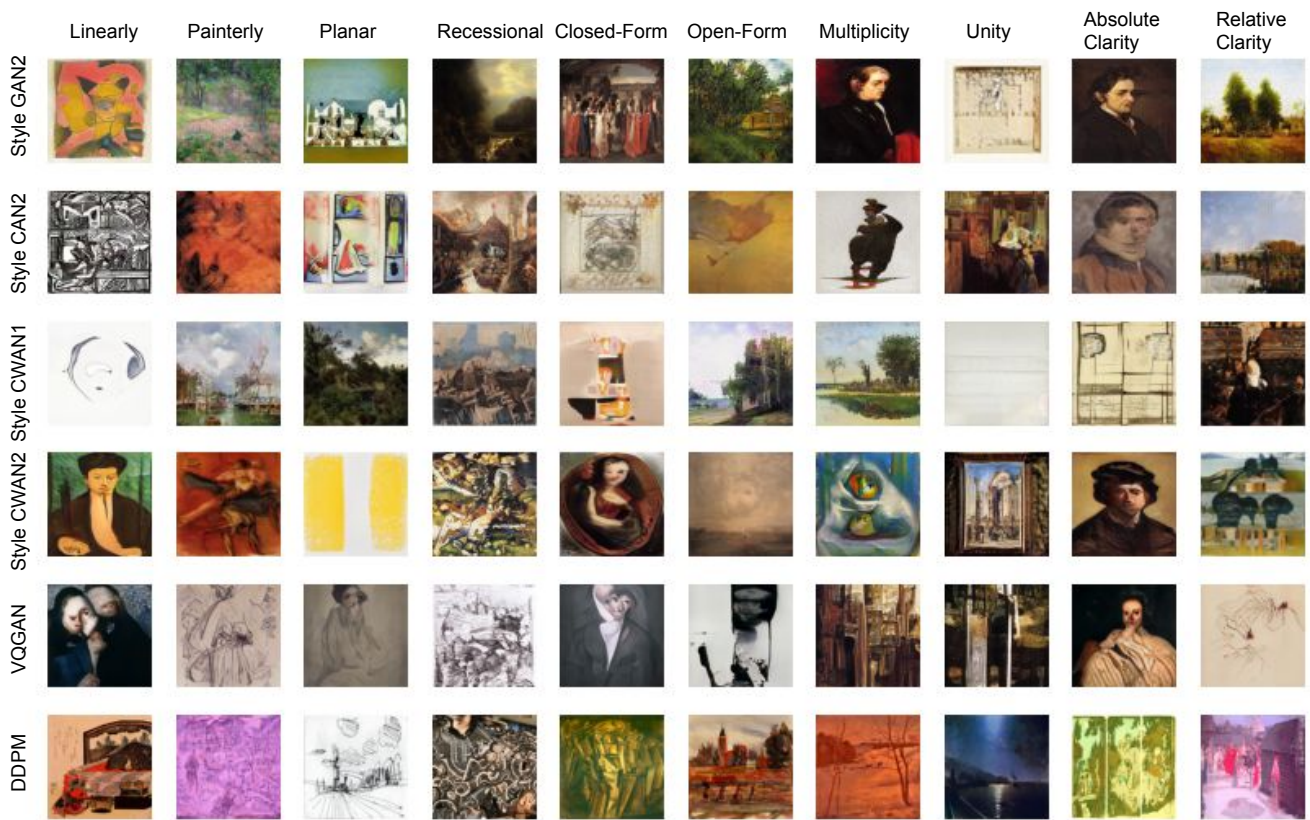


Figure 5. Generated art using StyleGAN2, StyleCAN2, StyleCWAN1, StyleCWAN2, VQGAN, and DDPM architectures for the different Wölfflin's principles.

D. General Art Principles

D.1. Experimental Procedures

This section explains in detail normalization and hypothesis testing in statistical analysis on proxy space for comparison of population means between human and generated art. Lastly, we will introduce OODness-proxy (**O**ut-**O**f-**D**istribution) to measure the significance of a generated sample value compared to human art.

D.1.1 Normalization by Human Art Statistics

For proxy analysis, two proxy embeddings $R \in \mathbb{R}^{m \times 15}$ and $G \in \mathbb{R}^{n \times 15}$ are collected respectively as human and generated samples; the numbers of samples are m and n . To have unified scales for comparison, the two embeddings are normalized by the mean and standard deviation of R : $\mu_R \in \mathbb{R}^{15}$ and $\sigma_R \in \mathbb{R}^{15}$. All analyses on proxy space started from the normalized embeddings: $R' = \frac{R - \mu_R}{\sigma_R}$ and $G' = \frac{G - \mu_R}{\sigma_R}$, assuming the repetitive expansions of μ_R and σ_R over the m and n samples. By normalization, the 15-means over the m samples of R' got exactly aligned with zero for all the 15 visual concepts, while the generated centers varied around zero depending on the relation of generated art to human art for different visual concepts.

D.1.2 Hypothesis Testing

To compare the centers of human and generated arts after normalization, a hypothetical test is set as below.

$$\begin{aligned} H_0 : \mu_R &= \mu_G \\ H_a : \mu_R &\neq \mu_G \end{aligned}$$

The null hypothesis is rejected when the test statistic: $|Z| > 3.0$ and p -value: 0.0026. The test statistic is $Z = \frac{\bar{X}_G - \bar{X}_R}{\frac{S_G}{\sqrt{n}}}$, where \bar{X}_G and \bar{X}_R are the sample means for generated and human artworks, S_g is the sample standard deviation of generated samples, and n is the number of the samples. [4] are referenced to set up our experiments.

D.1.3 OODness-proxy

The smaller tail probability of a generated sample on the normalized space by human statistics implies a significant visual deviation from general human art. Let p be the tail probability, and x be an arbitrary instance value in G' for i -th sample and j -th concept, and $n(x)$ is a normal distribution with $\mu = 0$ and $\sigma = 1$: $n(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2}$. Then, OODness-proxy is computed as below. It measures how a generated sample is outlied visually from the center of human art on proxy space.

$$\begin{aligned} p &= \min\left(\int_{-\infty}^x n(x) dx, \int_x^{-\infty} n(x) dx\right) \\ \text{OODness} &= -\log p \end{aligned}$$

D.2. Hypothesis Testing Results for Individual Generative Models

Table 3. For all generative models, we computed Z -statistics. Even though different models had different significant visual concepts, the signs of the significant statistics were consistent across the different models except for “linear” on VQ-GAN, and were matched to the direction of the visual changes from classical to modern art. For example, modern paintings are more “non-representational” and “geometric”, so positive statistics while less “representational” and “perspective” than classical human art, so got negative statistics on the concepts.

models	Z- statistics for 15 Visual Concepts														
	non-representational	representational	geometric	abstract	planar	closed	open	rough	perspective	broken	thin	flat	distorted	linear	all
StyleGAN-1	5.78	-5.12	4.34	6.92	5.65	-4.17	-0.3	6.43	-6.95	6.27	-6.12	6.84	6.25	3.73	6.45
StyleGAN-2	-0.5	0.79	-1.39	-1.11	-0.37	-1.25	1.3	0.7	0.35	-0.57	1.59	-0.15	-0.22	-0.74	-0.96
StyleCAN1	2.53	-2.16	2.15	3.6	3.44	-3.14	-0.59	4.85	-4.29	4.14	-1.55	4.86	4.92	3.81	3.04
StyleCAN2	4.38	-4.24	3.29	3.93	3.79	-2.39	1.17	3.01	-3.66	2.04	-2.77	3.4	2.69	1.47	4.48
StyleCWAN1	1.38	-1.27	1.18	2.63	3.26	-3.01	-1.68	4.72	-2.96	3.72	-0.91	3.92	3.97	2.62	2.14
StyleCWAN2	2.26	-1.97	1.13	3.05	2.3	-6.04	3.11	5.66	-3.94	4.03	-2.05	3.7	2.87	1.3	2.74
VQ-GAN	0.08	1.07	-2.75	0.07	-0.47	-1.59	1.48	2.24	-0.8	-0.61	-3.15	0.15	1.15	-3.66	0.17
DDPM	19.0	-18.29	17.71	18.89	17.5	-11.03	8.17	14.76	-18.8	11.51	-9.41	17.78	5.89	8.69	19.63

Table 4. After dividing human art samples into two groups by the year of made: before and after 1800, we computed Z -statistics to see how AI paintings are different from the time groups. We observed that AI art is visually similar to modern art while they are significantly different from classical human art.

years	Z- statistics for 15 Visual Concepts														
	non-representational	representational	geometric	abstract	planar	closed	open	rough	perspective	broken	thin	flat	distorted	linear	ambiguous
-1800	19.90	-18.39	19.07	25.99	19.60	-28.70	12.34	34.75	-31.49	28.35	-6.52	29.58	21.50	17.96	24.01
1800-	-1.21	1.39	-2.66	-2.27	-1.37	1.24	0.22	-2.24	2.78	-3.02	-0.79	-2.38	-1.86	-3.0	-1.94

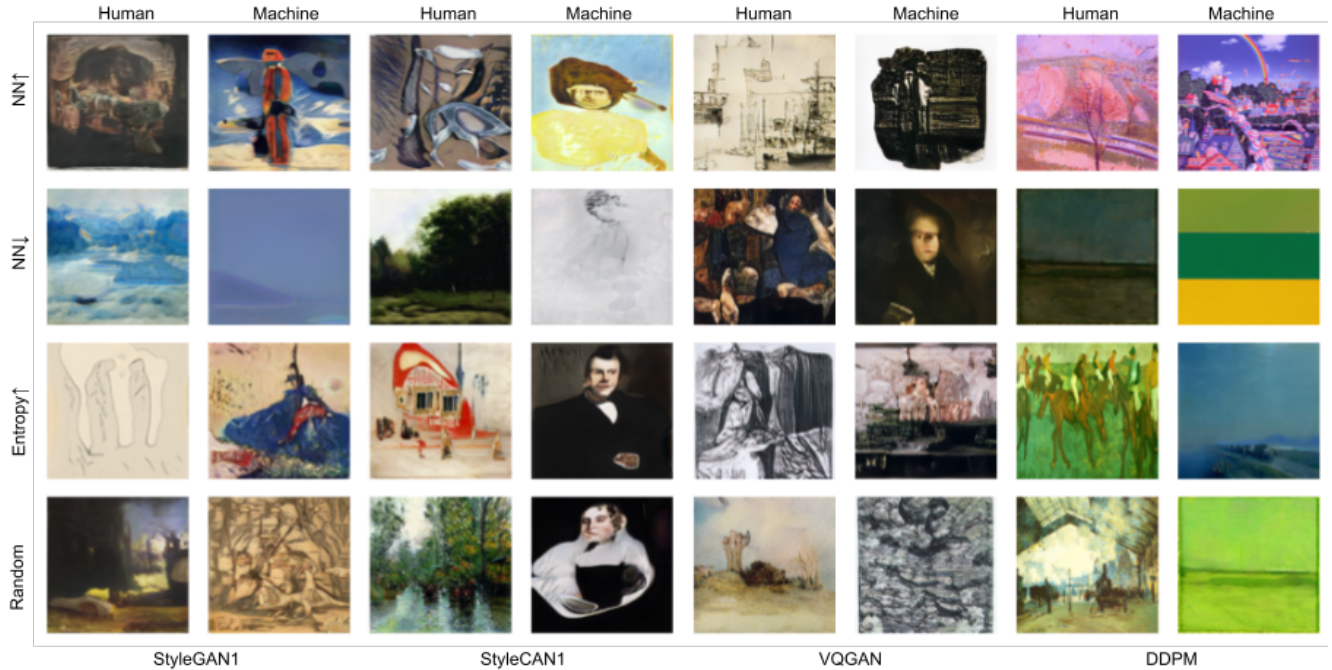


Figure 6. AI-generated art using StyleGAN-1, StyleCAN1, VQ-GAN, and DDPM architectures (from left to right respectively), which is most likely believed by users to be human-made and which are believed to be generated by a machine.

E. Qualitative Examples in Turing Test

We show some examples in Figure 6 from AI-generated art that is selected by users as most likely to be drawn by humans and some that are most likely AI-generated.

F. Qualitative Analysis of Likability

HighestNN: the generated artworks with the highest likability in HighestNN are similar to modern abstract art with contrasting solid colors. Paintings by StyleGAN-1 tended to have a more disorderly composition and unidentifiable figures, whereas StyleCAN-1 demonstrated identifiable abstract figures in a more orderly manner. The StyleCAN-2 and StyleCWAN-2 with a high mean likability (rated 4 to 4.4) displayed artistic characteristics reminiscent of Salvador Dali's surreal paintings. The balance of surreal figures and orderly arrangement of the overall composition make the artwork interesting and comforting for the viewer. In contrast, StyleCWAN-1 artwork in this category had architectural characteristics, sometimes with disorderly compositions. StyleGAN-2 artwork had architectural characteristics and geometric shapes. The artworks reminded the viewers of Kandinsky's *Composition* painting series. In HighestNN by VQ-GAN, we found many landscape-like abstract paintings to have vivid, thin, and patterned strokes in vertical, horizontal, or sometimes circular directions. It reminds us of Claude Monet's thin, dense strokes and loose swirls often appearing in his late artworks.

LowestNN: this group showed consistent characteristics in terms of style and subject matters in general. The most liked images exhibited an impressionist style marked by vibrant colors and rapid and distinct brush strokes. The subject matter of the liked artworks is mostly natural landscapes. These art pieces exhibited similarities to Claude Monet's iconic impressionist paintings, such as *Impression Sunrise* and *The Garden of Monet at Argenteuil*. This group is collected for the close nearest neighbor distance from WikiArt in ResNet50, so portrait genres in human art are often shown but disliked by subjects when the figures are contorted. However, for VQ-GAN, LowestNN presents no figuration images, but mostly abstract shapes, and opaque and heavy brush touches are common like the VQ-GAN LowestNN examples in Figure ??, it was hard to feel visual speed from the VQ-GAN samples.

Highest Shape Entropy: most paintings in the Highest shape entropy group tend to depict complex compositions and dense arrangements of figures. The complex nature of these paintings may have caused difficulties for the classifier to categorize them. The lively colors and complicated subject matter demonstrate similarities to Hieronymus Bosch's *The Garden Of Earthly Delights*. The clarity of the paintings determined by the distinct depictions of figures affects the mean likability and artist ratings. Paintings with blurred and overly distorted figures tend to receive lower mean likability and artist ratings.

G. Qualitative Analysis of Emotion

Here, we provide detailed analyses of two positive and negative emotions.

Anger: paintings that contain unfamiliar abstract figures tend to evoke confusion and emotions of anger within the audience. As shown in one of the "fear" responses, "Another kind of painting that makes me ask to myself what does it mean?" and "The painting looks confusing and shows no representation.", some puzzling and ambiguous depictions of machine art caused discomfort and annoyance from participants. The use of colors and artistic composition affects emotional response, too. Some vivid coloration, especially red, was usually associated with the emotion of anger, as stated by the participant, "The red color used seems like a man bleeding with anger." We also observed how the color can be associated with the concept of violence; "This makes me feel anger because the black marks between the rocks look like violent movements." Participants also felt anger in a painting representing a situation of disorder. The disorderly arrangement of figures creates a sense of unease and constructs anger.

Awe: Artwork that portrays ordinary subject matter, such as a brown coat, a man, a plant, or the sky, tends to construct emotions of awe within the audience, as demonstrated in the participant's statement; "The plant looks realistic." Many participants referred to the use of complementary colors and a soft color scheme as the grounds for their emotions of awe. In addition, artwork depicting natural landscapes also tends to create calming effects and raise awe for natural scenes, as stated by the participants: "Colors are vibrant.", "The play of colors in the sky of this painting is magnificent.", "The trees stand tall in the open air and the sky shines brightly.", and "Trees are often associated with life, growth, and vitality. In a desert landscape where life may be scarce, the presence of trees can symbolize endurance, strength, and the tenacity of living

organisms. This symbolism can evoke a sense of awe as it highlights the ability of nature to thrive in challenging conditions.”

Contentment: Floral color arrangement constructed emotions of contentment according to the answers from the participants: “The light of roses.”, “The green color invokes a sense of fertility and contentment.”, or “Mix of colors between green and yellow reminds the changing of seasons, brings peace and tranquility.” In addition to color, participants underlined the effect of depth, layers, and orderly composition in their feeling of contentment, or sometimes they associated the paintings with their own experience or memory from the past. These are the example statements: “This painting makes me feel relaxed because the items are well-ordered and displayed coherently.”, “The traces of the image remind us of happy everyday things.”, “I liked the image I felt pleasure because reminder my childhood.”, and “It reminds me of my grandfather’s place where he had ponds and lots of frogs.”

Disgust: The use of dark color and its visual effects construct emotions of disgust, as stated by the participants: “Too much darkness on the sea.”, “Looks like a dark cloud about to eat a human.”, and “The dark color scheme except for what to me could be a pair of broken glasses makes me feel frustrated.” In addition, when the artwork did not convey a message or meaning, the audience felt a lack of expression and evoked emotions of disgust. Participants stated that the lack of expression came from “A lifeless representation,” “Gray colors without expression,” and “It is only an object, so I didn’t like it, has no emotion.”


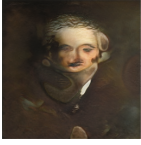




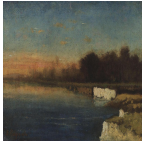

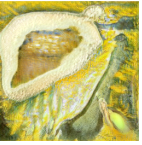






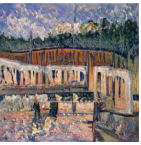
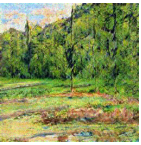


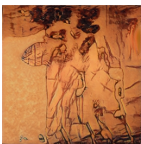

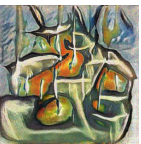
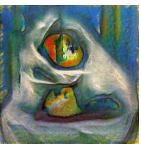





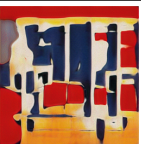



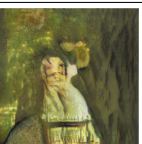


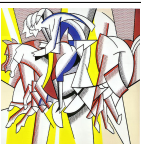



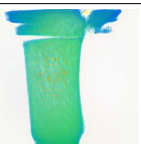
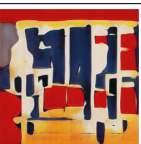


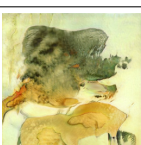


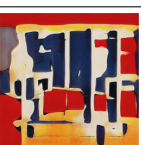

H. Significant and Insignificant AI-generated Art Instances on Proxy-Space

Continued to the next page.

References

- [1] Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone. The shape of art history in the eyes of the machine. In *AAAI*, 2018. 3, 4
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4
- [3] Diana Kim, Bingchen Liu, Ahmed Elgammal, and Marian Mazzone. Finding principal semantics of style in art. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018. 4
- [4] William M Mendenhall and Terry L Sincich. *Statistics for Engineering and the Sciences*. CRC Press, 2016. 8

Table 5. Based on standardized values by human art, the most insignificant (columns 2 and 3) and significant (columns 4 and 5) AI-generated artworks are presented with the most insignificant (column 1) and significant (column 6) human art with the two scores in brackets: OODness-proxy and normalized values. Significant machine samples are observed on the same side of the bias toward modern periods, as seen in our hypothesis test, while still less valued than the most out-liked human art for all 15 concepts. The last row samples are selected by the mean of OODness-proxy of 15 visual concepts.

	human	AI-generated				human
	most in-liked human	most in-liked AI-generated		most out-liked AI-generated		most out-liked human
Non-representational (scores)	 (0.69, 0.0)	 (0.69, 0.0)	 (0.69, 0.0)	 (13.09, 4.61)	 (13.56, 4.70)	 (24.97, 6.65)
abstract (scores)	 (0.69, 0.001)	 (0.69, 0.0)	 (0.69, 0.0)	 (9.83, 3.87)	 (11.20, 4.19)	 (15.59, 5.10)
closed (scores)	 (0.69, 0.0)	 (0.69, 0.0)	 (0.69, 0.0)	 (7.19, -3.17)	 (7.38, -3.23)	 (7.48, -3.26)
rough (scores)	 (0.69, 0.0)	 (0.69, 0.0)	 (0.69, 0.0)	 (7.44, 3.24)	 (7.95, 3.39)	 (8.60, 3.56)
perspective (scores)	 (0.69, 0.0)	 (0.69, 0.0)	 (0.69, 0.0)	 (8.69, -3.59)	 (9.60, -3.82)	 (13.0, -4.58)
distorted (scores)	 (0.69, 0.0)	 (0.69, 0.0)	 (0.69, 0.0)	 (7.97, 3.39)	 (9.52, 3.80)	 (16.46, 5.26)
ambiguous (scores)	 (0.69, 0.0)	 (0.69, 0.0)	 (0.69, 0.0)	 (10.29, 3.98)	 (10.66, 4.07)	 (17.85, 5.51)
mean over 15 concepts (OODness-proxy only)	 (0.66, -)	 (0.70, -)	 (0.71, -)	 (6.47, -)	 (7.34, -)	 (10.60, -)