Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Special Track on AI, the Arts and Creativity

Art Creation with Multi-Conditional StyleGANs

Konstantin Dobler[1], Florian Hübscher[1], Jan Westphal[1], Alejandro Sierra-Múnera[1], Gerard de Melo[1], Ralf Krestel[2]

[1]Hasso Plattner Institute / University of Potsdam
[2]ZBW – Leibniz Information Centre for Economics and Kiel University
{Konstantin.Dobler, Florian.Huebscher, Jan.Westphal}@student.hpi.uni-potsdam.de,
{Alejandro.Sierra, Gerard.DeMelo}@hpi.de, r.krestel@zbw.eu

Abstract

Creating art is often viewed as a uniquely human endeavor. In this paper, we introduce a multi-conditional Generative Adversarial Network (GAN) approach trained on large amounts of human paintings to synthesize realistic-looking paintings that emulate human art. Our approach is based on the StyleGAN neural network architecture, but incorporates a custom multi-conditional control mechanism that provides fine-granular control over characteristics of the generated paintings, e.g., with regard to the perceived emotion evoked in a spectator. We also investigate several evaluation techniques tailored to multi-conditional generation.

1 Introduction

The creation of art is often deemed a uniquely human endeavor. To create works of art, a human artist requires a combination of specific skills, understanding, and genuine intention. In light of this, there is a long history of endeavors to emulate this computationally, starting with early algorithmic approaches to art generation in the 1960s. Only recently, however, with the success of deep neural networks, has an automatic generation of images reached a new level, e.g., enabling us to synthesize photo-realistic faces [Karras *et al.*, 2020a].

In this paper, we investigate models that attempt to create works of art resembling human paintings. We propose techniques that encourage the model to follow a series of conditions, e.g., particular styles, motifs, evoked emotions, etc. For this, we adopt the well-known Generative Adversarial Network (GAN) framework [Goodfellow *et al.*, 2014], in particular the StyleGAN2-ADA architecture [Karras *et al.*, 2020a]. The greatest challenges have been the low resolution of generated images as well as the substantial amounts of required training data. This problem is exacerbated when there are multiple conditions, as there are even fewer training images available for each combination of conditions. We train our GAN using an enriched version of the ArtEmis dataset [Achlioptas *et al.*, 2021]. Two example images produced by our models can be seen in Figure 1. Our contributions include:

Figure 1: Example artworks produced by our StyleGAN models trained on the EnrichedArtEmis dataset (described in Section 3).

- We explore the use of StyleGAN to emulate human art, focusing in particular on the less explored conditional capabilities, to control traits such as art style, genre, and content.
- We introduce the concept of conditional center of mass in the StyleGAN architecture and explore its various applications. In particular, we propose a conditional variant of the *truncation trick* [Brock *et al.*, 2019] for the StyleGAN architecture that preserves the conditioning of samples.
- We formulate the need for *wildcard generation* in multi-conditional GANs, and propose a method to enable wildcard generation by replacing parts of a multi-condition-vector during training.

2 Related Work

Conditional GANs. Generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] are among the most well-known family of network architectures. Modern variants often adopt progressive growing to enable higher-resolution outputs [Karras *et al.*, 2018]. In recent years, several techniques have been proposed to incorporate conditions into the GAN architecture [Mirza and Osindero, 2014; Miyato and Koyama, 2018; de Vries *et al.*, 2017]. StyleGAN is a GAN architecture based on style transfer that provides control over both high-level attributes as well as finer details [Karras *et al.*, 2019]. Less attention has been given to multi-conditional GANs, where the conditioning is made up of multiple distinct categories of conditions that apply to each sample. Yildirim et al. hand-crafted loss functions for different parts of the conditioning, such as shape, color, or

Feature	Type	Size	Example
Style	Category	29	Post-Impressionism
Painter	Category	351	Vincent van Gogh
Genre	Category	30	cloudscape
Content tags	Text	768	tree, sky
Emotions	Distribution	9	40% awe, ...
Utterance	Text	768	"The sky seems ..."

Table 1: Features in the EnrichedArtEmis dataset, with example values for "The Starry Night" by Vincent van Gogh.

texture on a fashion dataset [Yildirim *et al.*, 2018]. Another study proposed a GAN conditioned on a base image and a textual *editing instruction* to generate the corresponding edited image [Park *et al.*, 2018].

Art with GANs. There is a long history of attempts to emulate human creativity by means of AI methods such as neural networks. Some studies focus on more practical aspects, whereas others consider philosophical questions such as whether machines are able to create artifacts that evoke human emotions in the same way as human-created art does. Further studies solicited human annotations describing how art is perceived [Mohammad and Kiritchenko, 2018; Achlioptas *et al.*, 2021]. The Creative Adversarial Network (CAN) architecture is encouraged to produce more novel forms of artistic images by deviating from style norms rather than simply reproducing the target distribution [Elgammal *et al.*, 2017]. Liu et al. proposed a new method to generate art images from sketches given a specific art style [Liu *et al.*, 2021]. Recently, vision–language models such as CLIP [Radford *et al.*, 2021] have been invoked to enable generation using natural language prompts that mainly describe the contents of the image. The focus of our work is to enable GANs to more freely generate diverse artistic images subject to more general conditions in the form of discrete attributes with fine-granular numerical values (e.g., 70% impressionism, 30% post-impressionism, 40% awe, 60% excitement). As certain paintings produced by GANs have been sold for high prices, important questions have been raised about issues such as authorship and copyrights of generated art [McCormack *et al.*, 2019].

3 Compiling an Annotated Dataset

WikiArt[1] is an online encyclopedia of visual art that catalogs both historic and more recent artworks. The service accepts community contributions and is run as a non-profit endeavor.

The ArtEmis dataset [Achlioptas *et al.*, 2021] contains roughly 80,000 artworks obtained from WikiArt, enriched with additional human-provided emotion annotations. On average, each artwork has been annotated by six different non-expert annotators with one out of nine possible emotions (*amusement*, *awe*, *contentment*, *excitement*, *disgust*, *fear*, *sadness*, *other*) along with a sentence (utterance) that explains their choice.

We enhance this dataset by adding further metadata crawled from the WikiArt website – *genre*, *style*, *painter*, and *content* tags – that serve as conditions for our model. Attribute values not provided by the corresponding WikiArt page are assigned a special UNKNOWN token. This token is also used for any categorical attribute value appearing fewer than 100 times in the data, in order to avoid conditions with low support in the training data. We refer to this enhanced version as the **EnrichedArtEmis** dataset.

A summary of the conditions present in the EnrichedArtEmis dataset is given in Table 1. The conditions *painter*, *style*, and *genre* are categorical and encoded using one-hot encoding. Emotion annotations are provided as a discrete probability distribution over the respective emotion labels, as there are multiple annotators per image. Finally, we have textual conditions, such as content tags and the annotator explanations from the ArtEmis dataset. For these, we use a pretrained TinyBERT model to obtain 768-dimensional embeddings [Jiao *et al.*, 2020].

4 Exploring Conditional StyleGAN

In this paper, we focus on the StyleGAN2-ADA variant of StyleGAN [Karras *et al.*, 2020a]. The architecture consists of a mapping network and a synthesis network. Given a latent vector $\mathbf{z}$ in the input latent space Z, the non-linear mapping network $f : Z \to W$ produces $\mathbf{w} \in W$. The mapping network is used to disentangle the latent space Z. The latent vector $\mathbf{w}$ then undergoes some modifications when fed into every layer of the synthesis network to produce the final image. This architecture improves the understanding of the generated image, as the synthesis network can distinguish between coarse and fine features.

Conditional GANs (cGANs) allow the provision of additional conditions alongside the random input vector [Mirza and Osindero, 2014]. The StyleGAN generator follows this approach but uses conditional normalization in each layer with condition-specific, learned scale and shift parameters [de Vries *et al.*, 2017; Karras *et al.*, 2020b]. With a latent code $\mathbf{z}$ from the input latent space Z and a condition $\mathbf{c}$ from the condition space C, the non-linear conditional mapping network $f_c : Z, C \to W$ produces $\mathbf{w}_c \in W$. The latent code $\mathbf{w}_c$ is then used together with conditional normalization layers in the synthesis network of the generator to produce the image. The discriminator uses a projection-based conditioning mechanism [Miyato and Koyama, 2018; Karras *et al.*, 2020b]. In the following, we study the effects of conditioning a StyleGAN.

We train a StyleGAN on the paintings in the EnrichedArtEmis dataset, which contains around 80,000 paintings from 29 art styles, such as impressionism, cubism or expressionism. We condition the StyleGAN on these art styles to obtain a conditional StyleGAN. Examples of generated images can be seen in Figure 2.

4.1 Conditional Truncation

The *truncation trick* [Brock *et al.*, 2019] is a method to adjust the trade-off between the fidelity (to the training distribution) and diversity of generated images by truncating the space from which latent vectors are sampled. For the StyleGAN architecture, the truncation trick works by first comput-

[1] https://www.wikiart.org/

Impressionism Cubism Expressionism

Figure 2: Images generated with StyleGAN conditioned on Style using identical random noise $\mathbf{z}$, yielding a similar color palette.

ing the *global* center of mass in W as

$$\bar{\mathbf{w}} = \mathbb{E}_{\mathbf{z} \sim P(\mathbf{z})}[f(\mathbf{z})]. \quad (1)$$

A sampled vector $\mathbf{w}$ in W is then moved towards $\bar{\mathbf{w}}$ with

$$\mathbf{w}' = \bar{\mathbf{w}} + \psi(\mathbf{w} - \bar{\mathbf{w}}), \text{ where } \psi < 1. \quad (2)$$

Moving towards a global center of mass has two disadvantages: Firstly, the *condition retention* problem, where the conditioning of an image is lost progressively the more we apply the truncation trick. This is because the global center of mass in W does not adhere to any given condition and hence the more we move towards it, the more the generated samples will deviate from their originally specified condition.

Secondly, when dealing with datasets with structurally diverse samples, such as EnrichedArtEmis, the global center of mass itself is unlikely to correspond to a high-fidelity image. For the Flickr-Faces-HQ (FFHQ) dataset [Karras *et al.*, 2019], the global center of mass produces a "typical", high-fidelity face. The FFHQ dataset contains centered, aligned and cropped images of faces and therefore has low structural diversity. On EnrichedArtEmis however, the global center of mass does not yield a high-fidelity painting, because the dataset is extremely diverse. Hence, applying the truncation trick is counterproductive with regard to the originally sought tradeoff between fidelity and the diversity.

Instead, we propose the *conditional* truncation trick, based on the intuition that different conditions are bound to have different centers of mass in W. The mean of a set of randomly sampled $\mathbf{w}$ vectors of flower paintings is going to be different than the mean of randomly sampled $\mathbf{w}$ vectors of landscape paintings. Thus, we compute a separate *conditional* center of mass $\bar{\mathbf{w}}_c$ for each condition c:

$$\bar{\mathbf{w}}_c = \mathbb{E}_{\mathbf{z} \sim P(\mathbf{z})}[f_{\mathrm{c}}(\mathbf{z}, c)]. \quad (3)$$

The computation of $\bar{\mathbf{w}}_c$ involves only the *mapping network* and not the bigger synthesis network. This enables an "on-the-fly" computation of $\bar{\mathbf{w}}_c$ at inference time for a given condition c. Moving a given vector $\mathbf{w}$ towards a conditional center of mass is done analogously to Equation 2:

$$\mathbf{w}' = \bar{\mathbf{w}}_c + \psi(\mathbf{w} - \bar{\mathbf{w}}_c) \quad (4)$$

We find that the introduction of a conditional center of mass is able to alleviate both the condition retention problem as well as the problem of low-fidelity centers of mass. Naturally, the conditional center of mass for a given condition will adhere to that specified condition. This effect can be seen in Figure 3, where the flower painting condition is reinforced the closer we move towards the conditional center of mass.

Furthermore, for datasets with low intra-class diversity, samples for a given condition have a lower degree of structural diversity. Although there are no *universally applicable* structural patterns for art paintings, there certainly are *conditionally applicable* patterns. For example, flower paintings usually exhibit flower petals. On diverse datasets that nevertheless exhibit low intra-class diversity, a conditional center of mass is therefore more likely to correspond to a high-fidelity image than the global center of mass. This effect can be observed in Figure 3 when considering the centers of mass with $\psi = 0$.

4.2 Condition-Based Vector Arithmetic

Given a trained conditional model, we can steer the image generation process in a specific direction. However, this can be taken even further. As we have a latent vector $\mathbf{w}$ in W for each generated image, we can apply transformations to $\mathbf{w}$ to alter the resulting image. One such transformation is vector arithmetic based on conditions.

Let $\mathbf{w}_{c_1}$ be a latent vector in W produced by the mapping network. The inputs are the specified condition $c_1 \in C$ and a random noise vector $\mathbf{z}$. Furthermore, let $\mathbf{w}_{c_2}$ be another latent vector in W produced by the same noise vector but with a different condition $c_2 \neq c_1$. We seek a transformation vector $\mathbf{t}_{c_1,c_2}$ such that $\mathbf{w}_{c_1} + \mathbf{t}_{c_1,c_2} \approx \mathbf{w}_{c_2}$. For better generalizability, we attempt to find the *average* difference between the conditions c_1 and c_2 in the W space.

Specifically, we sample $\mathbf{w}_{c_1}$ and $\mathbf{w}_{c_2}$ as described above with the same random noise vector $\mathbf{z}$ but different conditions and compute their difference. We repeat this process for a large number of randomly sampled $\mathbf{z}$ and compute the mean difference, which serves as our transformation vector $\mathbf{t}_{c_1,c_2}$. This is equivalent to computing the difference between the conditional centers of mass of the respective conditions:

$$\begin{aligned} \mathbf{t}_{c_1,c_2} &= \mathbb{E}_{\mathbf{z} \sim P(\mathbf{z})}[f_{\mathrm{c}}(\mathbf{z}, c_2) - f_{\mathrm{c}}(\mathbf{z}, c_1)] \\ &= \mathbb{E}_{\mathbf{z} \sim P(\mathbf{z})}[f_{\mathrm{c}}(\mathbf{z}, c_2)] - \mathbb{E}_{\mathbf{z} \sim P(\mathbf{z})}[f_{\mathrm{c}}(\mathbf{z}, c_1)] \\ &= \bar{\mathbf{w}}_{c_2} - \bar{\mathbf{w}}_{c_1}. \end{aligned} \quad (5)$$

Obviously, when we swap c_1 and c_2, the resulting transformation vector is negated as $\mathbf{t}_{c_1,c_2} = -\mathbf{t}_{c_2,c_1}$. Simple *conditional interpolation* is the interpolation between two vectors in W that were produced with the same $\mathbf{z}$ but different conditions. In contrast to conditional interpolation, our translation vector can be applied even to vectors in W for which we do not know the corresponding $\mathbf{z}$ or condition. This is the case in GAN inversion, where the $\mathbf{w}$ vector corresponding to a real-world image is iteratively computed. One such example can be seen in Figure 4, where the GAN inversion process is applied to the original Mona Lisa painting. For the GAN inversion, we used additive ramped-down noise [Karras *et al.*, 2020b]. To improve the low reconstruction quality, we optimized for the P^+ space [Zhu *et al.*, 2020]. The resulting approximation of the Mona Lisa is clearly distinct from the original painting, which we attribute to the fact that human proportions in general are hard to learn for our network.

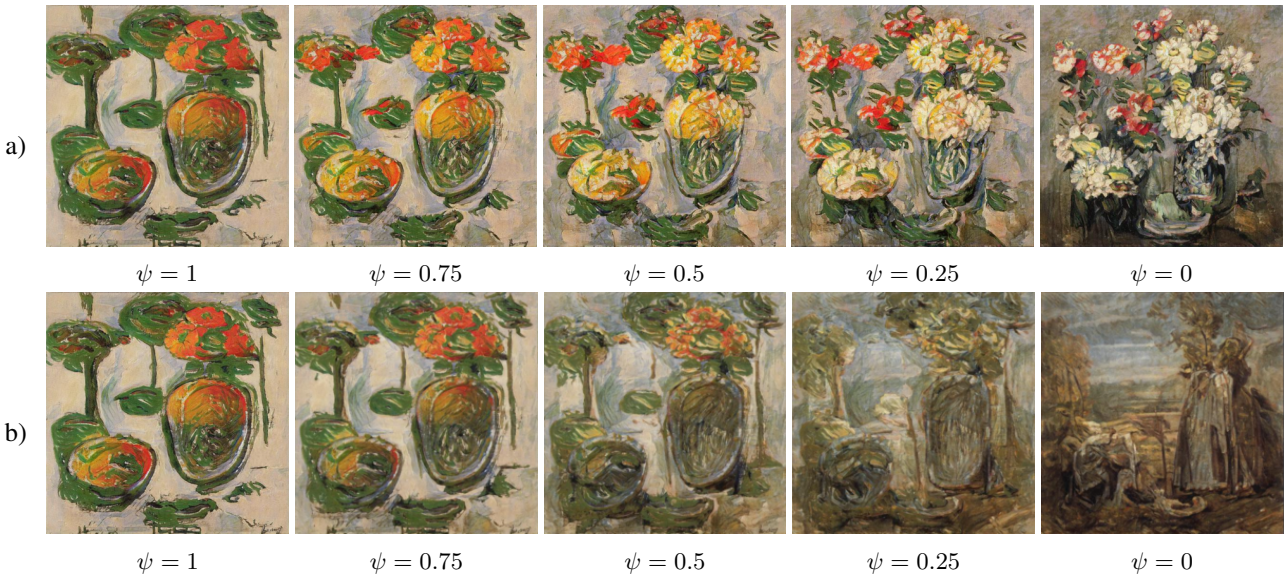
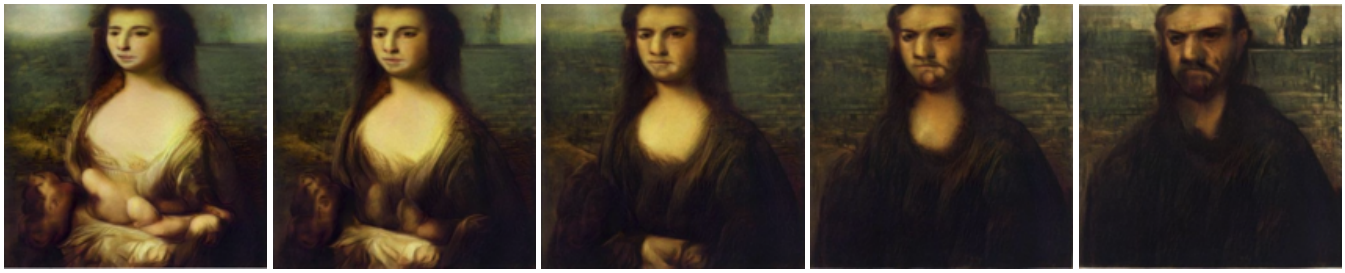


Figure 3: Visualization of the a) conditional (top) and b) conventional (bottom) truncation trick with the condition *flower paintings*. As $\psi \rightarrow 0$ and $\mathbf{w}_c$ is moved towards the *local* center of mass, the condition is retained (top), whereas when moving it to the *global* center (bottom), the *flower painting* condition is increasingly lost. Moreover, the *global* center of mass at $\psi = 0$ yields a low-fidelity image.



towards emotion: **awe** Inversion on Mona Lisa towards emotion: **fear**

Figure 4: The image at the center is the result of a GAN inversion process for the original *La Gioconda* (Mona Lisa) painting. Then we apply condition-based vector arithmetic between the emotions *awe* and *fear*. Note that the network has acquired several biases: Moving towards *fear*, the woman turns into a grim-looking man. Towards *awe*, the appearance becomes noticeably lighter.

Emotion: **anger**, Genre: **portrait**, Style: **Baroque**, Painter: **Rembrandt**, Cont.: **gentleman**

Emotion: **awe**, Genre: **landscape**, Style: **Impressionism**, Painter: **Monet**, Content: **trees**

Figure 5: Multi-conditional StyleGAN model trained with conditions *emotion*, *genre*, *style*, *painter*, and *content* tags.

5 Exploring Multi-Conditional StyleGAN

With data for multiple conditions at our disposal, we of course want to be able to use all of them simultaneously to guide the image generation. This could be skin, hair, and eye color for faces, or art style, emotion, and painter for EnrichedArtEmis. Let S be the set of unique conditions. We define a multi-condition ζ as being comprised of multiple sub-conditions c_s for $s \in S$.

5.1 Creating a Multi-Conditional Condition Vector

To use a multi-condition ζ during training, we need to find a vector representation that can be fed into the network alongside the random noise vector. We achieve this by first obtaining a vector representation for each sub-condition c_s, as described in Section 3. Then we concatenate these individual representations.

With this setup, multi-conditional training and image generation with StyleGAN is possible. In Figure 5, we can see

Style: **Post-Impressionism**,
Painter: **Vincent van Gogh**

Style: **Impressionism**,
Painter: **Claude Monet**

Figure 6: Comparison of paintings produced by a multi-conditional StyleGAN model for the painters *Monet* and *van Gogh*. Both paintings are produced using the same random noise $\mathbf{z}$ and with *genre*: **landscape**, *emotion*: **contentment**, *content* tag: **water**.

paintings produced by this multi-conditional generation process. It is worth noting that some conditions are more subjective than others. The emotions a painting evoke in a viewer are highly subjective and may even vary depending on external factors such as mood or stress level. Nevertheless, we observe that most sub-conditions are reflected rather well in the samples. In Figure 6, we compare our network's renditions of Vincent van Gogh and Claude Monet conditions.

5.2 Wildcard Generation

A multi-conditional StyleGAN model allows us to exert a high degree of influence over the generated samples. For example, when using a model trained on the sub-conditions *emotion*, *art style*, *painter*, *genre*, and *content* tags, we can attempt to generate "awe-inspiring, impressionistic landscape paintings with trees by Monet". However, this degree of influence can also become a burden, as we *always* have to specify a value for every sub-condition that the model was trained on.

Therefore, we propose *wildcard generation*: For a multi-condition ζ, we wish to be able to replace arbitrary sub-conditions c_s with a *wildcard mask* and still obtain samples that adhere to the remaining parts of ζ. As our wildcard mask, we choose replacement by a zero-vector: Any sub-condition c_s within ζ that is not specified is replaced by a zero-vector of the same length.

To ensure that the model is able to handle such ζ, we also integrate this into the training process with a *stochastic condition masking* regime. Whenever a sample is drawn from the dataset, k sub-conditions are randomly chosen from the entire set of sub-conditions. Each of the chosen sub-conditions is masked by a zero-vector with a probability p.

Figure 7 shows results of such wildcard generation. All paintings match the condition of "landscape painting with mountains", while other conditions vary. Still, there is a degree of *structural similarity* between the samples, with similar subject matter depicted in the same places across all of them. This may be a weakness of multi-conditional StyleGANs especially for rare combinations of sub-conditions.

Figure 7: Paintings produced by a multi-conditional StyleGAN model with conditions *genre*: **landscape**, *content* tag: **mountains**, and *style*, *painter*, *emotion* replaced by a *wildcard* zero-vector.

6 Evaluation

Evaluating generated art is a difficult endeavor. To make the evaluation more tangible, we do not attempt to evaluate artistic value but, more generally, the quality of the generated images and to what extent they adhere to the provided conditions. Although we meet the main requirements to produce pleasing computer-generated images [Baluja *et al.*, 1994], the question remains whether our generated artworks are of sufficiently high quality.

Models and Data. One of our GANs has been trained only on the *content tag* condition, which we denote as GAN_{T}. The GAN_{ESG} model is trained on *emotion*, *style*, and *genre*. Finally, GAN_{ESGPT} includes the conditions of both GAN_{T} and GAN_{ESG} in addition to *painter*. All GANs are trained with default parameters on the EnrichedArtEmis dataset described in Section 3, using a standardized 512×512 resolution obtained via resizing and optional cropping. The conditional truncation trick was not used during the evaluation.

6.1 Metrics

Manual Evaluation. We conducted a manual analysis checking to what extent the models consider the specified conditions. Given a sample S, where each entry $s \in S$ consists of the image s_{img} and the condition vector s_c, we summarize the overall correctness as $e_{\mathrm{qual}}(S)$, defined as:

$$e_{\mathrm{qual}}(S) = \frac{1}{|S|} \sum_{s \in S} \frac{1}{d} \sum_{i=1}^{d} b(s_{\mathrm{img}}, s_{c_i}) \tag{6}$$

Here, $b(i, c) = 1$ if image i matches condition c in a manual assessment, and 0 otherwise. The sample size for S is 76 for GAN_{T}, following previous work [Bohanec *et al.*, 1992], and 100 for the other models.

Fréchet Inception Distance (FID). In the literature on GANs, a number of metrics have been found to correlate well with image quality and hence have gained widespread adoption [Szegedy *et al.*, 2016; Devries *et al.*, 2019; Bińkowski *et al.*, 2018]. The FID estimates the quality of a collection of generated images based on its proximity to real data. Specifically, it considers the Fréchet distance between the multivariate Gaussian distributions of the generated data and real human data, using the embedding space of the pretrained InceptionV3 model. A lower score represents a closer proximity to the original dataset.

Metric	GAN_{T}	GAN_{ESG}	GAN_{ESGPT}
FID	5.38	5.37	4.67
Emotion Intra-FID	–	10.51	9.74
Style Intra-FID	–	9.23	7.98
Genre Intra-FID	–	8.19	7.31
Painter Intra-FID	–	–	8.65
Content-Tag Intra-FID	5.46	–	6.83
Intra-FID (average)	5.46	9.31	8.10
FJD ($\alpha = 0.5$)	9.42	9.29	8.47
Qualitative results (e_{qual})	0.91	0.88	0.83
Hybrid metric (e_{art})	8.11	10.42	9.69

Table 2: Overall evaluation using quantitative metrics as well as our proposed hybrid metric for our (multi-)conditional GANs.

Fréchet Joint Distance (FJD). A downside of FID is that it disregards the conditioning. Accounting for both conditions and the output data is possible with the Fréchet Joint Distance (FJD) [Devries *et al.*, 2019]. It involves calculating the Fréchet Distance over the joint image-conditioning embedding space, using a function $g = \left[f(\mathbf{x}^{(i)}),\ \alpha h(\mathbf{y}^{(i)})\right]$ that concatenates representations for the image vector $\mathbf{x}$ and the conditional embedding $\mathbf{y}$. The representation for the latter is obtained using an embedding function h that embeds our multi-conditions as stated in Section 5.1. A scaling factor α allows us to flexibly adjust the impact of the conditioning embedding compared to the vanilla FID score.

Intra-Fréchet Inception Distance (I-FID). We adapt the Intra-Fréchet Inception Distance (I-FID) [Miyato and Koyama, 2018] to be able to properly *compare* the impact of different conditions, while still taking image quality, conditional consistency, and intra-class diversity into account.

For scalability to highly multi-conditional settings, we select 50% of the condition entries ce within the corresponding distribution, and for every ce generate the intra-conditional images based on S and calculate the local FID score. We can then compute the average for each condition and finally compute the condition average that represents our I-FID score.

Hybrid Evaluation Metric. We further propose a combination of qualitative and quantitative evaluation scoring for our GAN models [Bohanec *et al.*, 1992]. For this, we compute the quantitative metrics as well as the qualitative score: $e_{\mathrm{art}} = \frac{1}{2}(e_{\text{I-FID}} + e_{\text{FJD}})(2 - e_{\mathrm{qual}})$.

6.2 Results

Given a particular GAN model, we followed previous work [Szegedy *et al.*, 2016] and generated at least 50,000 multi-conditional artworks for each quantitative experiment in the evaluation. The results in Table 2 reveal that the quantitative metrics mostly match the actual results of manually checking the presence of every condition. However, with an increased number of conditions, the qualitative results start to diverge from the quantitative metrics. This validates our assumption that the quantitative metrics do not perfectly represent our perception when it comes to the evaluation of multi-conditional images. Despite the small sample size, we can conclude that our manual labeling of each condition acts as an uncertainty score for the reliability of the quantitative measurements. We conjecture that the worse results for GAN_{ESGPT} may be caused by outliers, due to the higher probability of producing rare condition combinations.

Overall, we find that we do not need an additional classifier that would require large amounts of training data to enable a reasonably accurate assessment. Instead, we can use our e_{art} metric to put the considered GAN evaluation metrics in context. All in all, somewhat unsurprisingly, the conditional GAN_{T} produces more accurate results in comparison to multi-conditional GANs with many different conditions. However, the latter provide substantial control that can be very useful when generating art.

Analysis. In order to eliminate the possibility that a model is merely replicating images from the training data, we compare a generated image to its nearest neighbors in the training data. To find these nearest neighbors, we use a *perceptual similarity* measure [Zhang *et al.*, 2018], which measures the similarity of two images embedded in a deep neural networks' intermediate feature space. Using this method, we did not find any generated image to be a near-identical copy of an image in the training dataset.

Discussion. The quantitative methods do not explicitly judge the visual quality of an image but rather focus on how well the images produced by a GAN match those in the original dataset, both generally and with regard to particular conditions. Hence, the *image quality* here is considered with respect to a particular dataset and model. We follow the definition of creativity of Dorin and Korb, which evaluates the probability to produce certain representations of patterns [Dorin and Korb, 2009] and extend it to the GAN architecture. Of course, historically, art has been evaluated qualitatively by experts. Such assessments, however, are typically costly to procure and are also a matter of taste and thus it is not possible to obtain a completely objective evaluation.

7 Conclusion

In this paper, we have applied the powerful StyleGAN architecture to a large art dataset and investigated techniques to enable multi-conditional control. The images that this trained network is able to produce are convincing and in many cases appear to be able to pass as human-created art. Due to the nature of GANs, the created images of course may perhaps be viewed as imitations rather than as truly novel or creative art. This stems from the objective function that is optimized during training, which encourages the model to imitate the training distribution as closely as possible. Our evaluation shows that automated quantitative metrics start diverging from human quality assessment as the number of conditions increases, especially due to the uncertainty of precisely classifying a condition. To alleviate this challenge, we also conduct a qualitative evaluation and propose a hybrid score.

Overall, our multi-conditional models deliver promising results. For future work, it might be interesting to investigate how inherent biases in the training data translate to the generated images.

References

[Achlioptas *et al.*, 2021] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. ArtEmis: Affective language for visual art. In *CVPR*, pages 11564–11574, 2021.

[Baluja *et al.*, 1994] Shumeet Baluja, Dean Pomerleau, and Todd Jochem. Towards automated artificial evolution for computer-generated images. *Connection Science*, 6(2-3):325–354, 1994.

[Bińkowski *et al.*, 2018] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

[Bohanec *et al.*, 1992] Marko Bohanec, Bozo Urh, and Vladislav Rajkovič. Evaluating options by combined qualitative and quantitative methods. *Acta Psychologica*, 80(1):67–89, 1992.

[Brock *et al.*, 2019] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.

[de Vries *et al.*, 2017] Harm de Vries, Florian Strub, Jeremie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *NeurIPS*, volume 30, pages 6594–6604, 2017.

[Devries *et al.*, 2019] Terrance Devries, Adriana Romero, L. Pineda, Graham W. Taylor, and Michal Drozdzal. On the evaluation of conditional GANs. *ArXiv*, abs/1907.08175, 2019.

[Dorin and Korb, 2009] Alan Dorin and Kevin B Korb. Improbable creativity. In *Computational Creativity: An Interdisciplinary Approach*, number 09291 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2009.

[Elgammal *et al.*, 2017] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *ICCC*, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, pages 2672–2680, 2014.

[Jiao *et al.*, 2020] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *EMNLP*, pages 4163–4174, 2020.

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4396–4405, 2019.

[Karras *et al.*, 2020a] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, volume 33, pages 12104–12114, 2020.

[Karras *et al.*, 2020b] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, pages 8107–8116, 2020.

[Liu *et al.*, 2021] Bingchen Liu, Kunpeng Song, Yizhe Zhu, and Ahmed Elgammal. Sketch-to-art: Synthesizing stylized art images from sketches. In *Computer Vision – ACCV 2020*, pages 207–222, 2021.

[McCormack *et al.*, 2019] Jon McCormack, Toby Gifford, and Patrick Hutchings. Autonomy, authenticity, authorship and intention in computer generated art. In *Computational Intelligence in Music, Sound, Art and Design*, pages 35–50, 2019.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014.

[Miyato and Koyama, 2018] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *ICLR*, 2018.

[Mohammad and Kiritchenko, 2018] Saif M. Mohammad and Svetlana Kiritchenko. An annotated dataset of emotions evoked by art. In *LREC*, volume 11, 2018.

[Park *et al.*, 2018] Hyojin Park, Youngjoon Yoo, and Nojun Kwak. MC-GAN: Multi-conditional generative adversarial network for image synthesis. In *The British Machine Vision Conference (BMVC)*, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[Yildirim *et al.*, 2018] Gökhan Yildirim, Calvin Seward, and Urs M. Bergmann. Disentangling multiple conditional inputs in GANs. *ArXiv*, abs/1806.07819, 2018.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[Zhu *et al.*, 2020] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Improved StyleGAN embedding: Where are the good latents? *ArXiv*, abs/2012.09036, 2020.