
ATML Report

Giorgia Adorni
giorgia.adorni@usi.ch

Felix Boelter
felix.boelter@usi.ch

Stefano Carlo Lambertenghi
stefano.carlo.lambertenghi@usi.ch

Abstract

1 Introduction

This objective of this project is to investigate the reliability and reproducibility of a paper accepted for publication in a top machine learning conference. The models have been implemented using code and information provided by the authors. Actually we are not participating formally to the *Fall Edition of the ML Reproducibility Challenge 2021*.

With this work, we are going to verify the empirical results and claims of the paper “*Generative adversarial transformers*” by Hudson and Zitnick [1], by reproducing three of the computational experiments performed by the authors: (1) the **StyleGAN2** by Karras et al. [2, 3], a GAN network which uses one global latent style vector to modulate the features of each layer, hence to control the style of all image features globally, (2) the **GANformer** with **Simplex Attention** by Hudson and Zitnick [1], a generalisation of the StyleGAN design with k latent vectors that cooperate through attention, allowing for a spatially finer control over the generation process since multiple style vectors impact different regions in the image concurrently, in particular permitting communication in one direction, in the generative context – from the latents to the image features, and (3) the **GANformer** with **Duplex Attention** by Hudson and Zitnick [1], which is based on the same principles as the previous but propagating information both from global latents to local image features, enabling both top-down and bottom-up reasoning to occur simultaneously.

The first model is used as baseline, while the remaining are the architectures introduced by the authors. They consider the GANformer has “a novel and efficient type of transformer” which demonstrates its strength and robustness over a range of task of visual generative modelling — simulated multi-object environments (real-world indoor and out-door scenes) — achieving state-of-the-art results in terms of both image quality and diversity, while benefiting from fast learning and better data-efficiency.

2 Related works/Background

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [4], are deep-learning-based generative models which learn to determine whether a sample is from the model or the data distribution.

The classic architecture is composed by two main neural networks: the *generator* $G(z)$, which take as input a sample from the latent space, or a vector drawn randomly from a Gaussian distribution, and use it to generate new plausible examples in the problem domain — images in our case —, and the *discriminator* $D(x)$, which takes an example from the domain as input, and predicts a binary class label which classify the examples as real (coming from the training dataset) or fake (generated by G).

The two models are trained together: the discriminator D estimates the probability that sample x is generated by G or is a real sample, and aims at maximising the probability of assigning the correct label to both real and fake samples, while the generator G is trained to maximise the probability of the discriminator D making a mistake, so it aims to minimise $\log(1 - D(G(z)))$.

Combining the two objectives for $G(z)$ and $D(x)$ we get the *GAN min-max game* with the value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_*(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

GANs typically work with images, as in our case, for this reason, both the generator and discriminator models use Convolutional Neural Networks (CNNs).

2.2 StyleGAN2

StyleGAN are a re-design of the GANs generator architecture, aimed at control the image synthesis process [3]. The StyleGAN approach departs from the design of a common generator network, consisting in a multi-layer CNN that receives a randomly sampled vector z and transforms it into an image, and, instead, it introduces a feed-forward mapping network that outputs an intermediate vector w , which in turn interacts directly with each convolution through the synthesis network, globally controlling the feature maps’ statistics at every layer.

As mentioned by Karras et al. [2], «this generator starts from a learned constant input and adjusts the “style” of the image at each convolution layer based on the latent code, therefore directly controlling the strength of image features at different scales. Combined with noise injected directly into the network, this architectural change leads to automatic, unsupervised separation of high-level attributes (e.g., pose, identity) from stochastic variation (e.g., freckles, hair) in the generated images, and enables intuitive scale-specific mixing and interpolation operations».

The generator embeds the input latent code into an intermediate latent space. In particular, the input latent space must follow the probability density of the training data, while the intermediate latent space is free from that restriction and is therefore allowed to be disentangled.

Traditionally, the latent code is provided to the generator through an input layer of a feed-forward network. Karras et al. [2] depart from this design by omitting the input layer altogether and starting from a learned constant instead. Given a latent code z in the input latent space Z , a non-linear mapping network $f : Z \rightarrow W$ first produces $w \in W$. Learned affine transformations then specialize w to styles $y = (y_s, y_b)$ that control adaptive instance normalization (AdaIN) [?] operations after each convolution layer of the synthesis network g . The AdaIN operation is defined as:

$$\text{AdaIN}(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}, \quad (2)$$

where each feature map x_i is normalized separately, and then scaled and biased using the corresponding scalar components from style y . Thus the dimensionality of y is twice the number of feature maps on that layer.

Moreover, they provide the generator with a direct means to generate stochastic detail by introducing explicit noise inputs.

The StyleGAN2 architecture makes it possible to control the image synthesis via scale-specific modifications to the styles. In particular, this approach attains layer-wise decomposition of visual properties, allowing StyleGAN to control global aspects of the picture such as pose, lighting conditions or colour schemes, in a coherent manner over the entire image.

But while StyleGAN successfully disentangles global properties, it is more limited in its ability to perform spatial decomposition, as it provides no direct means to control the style of a localized regions within the generated image.

2.3 Transformers

Transformers are models which avoid using *sequence-aligned RNNs* or *convolution* and instead rely entirely on an *attention mechanism* to compute representations of its input and output [5].

The architecture is composed by an encoder-decoder structure where, the *encoder* maps an input sequence of symbol representations (x_1, \dots, x_n) to a sequence of continuous representations $z = (z_1, \dots, z_n)$, while the *decoder*, given z , generates an output sequence (y_1, \dots, y_m) of symbols one element at a time.

At each step the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next. In addition, each transformer module (encoder and decoder) contains $N = 6$ identical layers, each with two sub-layers: (1) a *multi-head self-attention mechanism* working in parallel, and (2) a *position-wise fully connected feed-forward neural network*, with residual connections followed by normalisation. Moreover, the decoder self-attention sub-layer has a mask which prevents positions from attending to subsequent positions (so positional encoding provided as additional input to bottom layer). This, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on the known outputs at positions less than i . In addition, the decoder has a third sub-layer, which performs *multi-head attention* over the output of the encoder stack.

An *attention function* [5], can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

The input consists of queries and keys of dimension d_k , and values of dimension d_v . The dot products of the query with all keys is computed, each divided by $\sqrt{d_k}$ and then a softmax function is applied to obtain the weights on the values. In practice, the attention function is computed on a set of queries simultaneously, packed together into a matrix Q . The keys and values are also packed together into matrices K and V .

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Instead of performing a single attention function with d_{model} -dimensional keys, values and queries, we found it beneficial to linearly project the queries, keys and values h times with different, learned linear projections to d_k , d_k and d_v dimensions, respectively. On each of these projected versions of queries, keys and values we then perform the attention function in parallel, yielding d_v -dimensional output values. These are concatenated and once again projected, resulting in the final values. *Multi-head attention* allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1 \dots, \text{head}_h)W^O, \quad (4)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$. Vaswani et al. [5] employed $h = 8$ parallel attention layers, or heads. For each of these, $d_k = d_v = d_{\text{model}}/h = 64$. Due to the reduced dimension of each head, the total computational cost is similar to that of single-head attention with full dimensionality.

The transformer uses multi-head attention in three different ways:

1. In "encoder-decoder attention" layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence.
2. The encoder contains self-attention layers, in which all of the keys, values and queries come from the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder.
3. Similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position. We need to prevent leftward information flow in the decoder to preserve the auto-regressive property. We implement this inside of scaled dot-product attention by masking out (setting to $-\infty$) all values in the input of the softmax which correspond to illegal connections.

3 Methodology

3.1 Generative Adversarial Transformers

The Generative Adversarial Transformer (GANformer), as presented in [1], is a type of GAN, which involves two networks: the *generator* (G), which maps a sample z from the latent space to the output space (e.g., producing an image x), and the *discriminator* (D) which receives images originated from the dataset or the generator and seeks to discern between real and fake samples.

As shown in Section 2.1, the two networks compete with each other through a minimax game until reaching an equilibrium.

In the generator network, the latent z has a shape `[batch_size, component_num, latent_dim]`. The latent components are defined by splitting z along the second dimension to obtain z_1, \dots, z_k latent components.

The generator likewise is composed of two parts:

1. The *mapping network*, which consists in a series of feed-forward layers, whose objective is to convert the sampled latents from a normal distribution z to the intermediate space w . The k latent components either are mapped independently or interact with each other through self-attention.
2. The *synthesis network* is composed by multiple layers of convolution where the images features begin from a small constant/sampled grid and up-sampling until reaching the desirable resolution. After each convolution, the image features are modulated by the intermediate latent vectors w , meaning that their variance and bias are controlled. While in StyleGAN2 a single global w vector controls all the features equally, the GANformer uses attention so that the k latent components specialise to control different regions in the image to create it cooperatively, and therefore perform better especially in generating images depicting multi-object scenes, allowing also for a flexible and dynamic style modulation at the region level.

Attention can be used in several ways:

1. Simplex Attention: when attention is applied in one direction only from the latents to the image features (top-down).
2. Duplex Attention: when attention is applied in the two directions: latents to image features (top-down) and then image features back to latents (bottom-up), so that each representation informs the other iteratively.
3. Self Attention between latents: can also be used so to each direct interactions between the latents.
4. Self Attention between image features (SAGAN model): prior approaches used attention directly between the image features, but this method does not scale well due to the quadratic number of features which becomes very high for high-resolutions.

The discriminator model performs multiple layers of convolution downsampling on the image, reducing the representation's resolution gradually until making final prediction. Optionally, attention can be incorporated into the discriminator

as well where it has multiple k aggregator variables, that use attention to adaptively collect information from the image while being processed. We observe small improvements in model performance when attention is used in the discriminator, although note that most of the gain in using attention based on our observations arises from the generator.

The main difference between the GANformer and a GAN is that, instead of relying on multiple layers of convolution, a *bipartite transformer* structure is used to compute *soft attention*, that iteratively aggregates and disseminates information between the generated image features and a compact set of *latent variables* enable bidirectional interaction between these dual representations.

The *transformer network* corresponds to the *multi-layer bidirectional transformer encoder* (BERT) described by Devlin et al. [6], which interleaves *multi-head self-attention* and *feed-forward layers*. Each pair of self-attention and feed-forward layers is intended as a *transformer layer*, hence, a transformer is a stack of several such layers. The *self-attention layer* considers all pairwise relations among the input elements, so to update each single element by attending to all the others. The *bipartite transformer* generalises this formulation, featuring instead a bipartite graph between two groups of variables — in the GAN case, latents and image features. There are two attention operations that could be computed over the bipartite graph, depending on the direction in which information propagates, (1) *simplex attention* permits communication either in one way only, in the generative context, from the latents to the image features, and (2) *duplex attention* which enables it both top-down and bottom-up ways.

3.1.1 Simplex attention

As already mentioned, simplex attention distributes information in a single direction over the bipartite transformer graph.

Formally, let $X^{n \times d}$ denote an input set of n vectors of dimension d — where, for the image case, $n = W \times H$ — and $Y^{m \times d}$ denote a set of m aggregator variables — the latents, in the generative case. Specifically, the attention is computed over the derived bipartite graph between these two groups of elements, as in Equation (3), moreover:

$$a(X, Y) = \text{Attention}(q(X), k(Y), v(Y)), \quad (5)$$

where $q(\cdot), k(\cdot), v(\cdot)$ are functions that respectively map elements into queries, keys, and values, all maintaining dimensionality d . The mappings are provided with positional encodings to reflect the distinct position of each element (e.g. in the image). This bipartite attention is a generalisation of self-attention, where $Y = X$.

Standard transformers implement an additive update rule of the form:

$$u^a(X, Y) = \text{LayerNorm}(X + a(X, Y)), \quad (6)$$

however, [1] used the retrieved information to control both the scale as well as the bias of the elements in X , in line with the practice promoted by the StyleGAN model [3]:

$$u^s(X, Y) = \gamma(a(X, Y)) \odot \omega(X) + \beta(a(X, Y)), \quad (7)$$

where $\gamma(\cdot), \beta(\cdot)$ are mappings that compute multiplicative and additive styles (gain and bias), maintaining dimensionality d , and $\omega(X) = X - \mu(X)$ normalises each element with $\sigma(X)$ respect to the other features. By normalizing X (image features), and then letting Y (latents) control the statistical tendencies of X , the information propagation from Y to X is enabled, allowing the latents to control the visual generation of spatial attended regions within the image, so as to guide the synthesis of objects or entities. The multiplicative integration permits significant gains in the model performance.

3.1.2 Duplex attention

Duplex attention can be explained by taking into account the variables Y to set their own key-value structure: $Y = (K^{n \times d}, V^{n \times d})$, where the values store the content of the Y variables, as before (e.g. the randomly sampled latent vectors in the case of GANs) while the keys track the centroids K of the attention-based assignments between Y and X , which can be computed as $K = a(Y, X)$ — namely, the weighted averages of the X elements using the bipartite attention distribution derived through comparing it to Y . Consequently, the new update rule is defined as follows:

$$u^d(X, Y) = \gamma(A(X, K, V)) \odot \omega(X) + \beta(A(X, K, V)), \quad (8)$$

where, two attention operations are compound on top of each other: first compute the *soft attention* assignments between X and Y , by $K = a(Y, X)$, and then refine the assignments by considering their centroids, by $A(X, K, V)$. This is analogous to the *k-means algorithm* and works more effectively than the simpler update u^a defined above in Equation (7).

Finally, to support bidirectional interaction between X and Y (the image and the latents), two reciprocal simplex attentions are chain from X to Y and from Y to X , obtaining the duplex attention, which alternates computing $Y := u^a(Y, X)$ and $X := u^d(X, Y)$, such that each representation is refined in light of its interaction with the other, integrating together bottom-up and top-down interactions.

3.1.3 Vision-specific adaptations

Some adaptations, that will be later described in Section 4.2, have been applied to the structure of the GANformer in order to foster an interesting communication flow. Rather than densely modelling interactions among all the pairs of pixels in the images, instead it supports *adaptive long-range interaction* between far away pixels in a moderated manner, passing through a compact and global latent bottleneck that selectively gathers information from the entire input and distributes it back to the relevant regions.

3.1.4 The Generator and Discriminator Networks

The StyleGAN2 model [2, 3], presented in Section 2.2, is used as a starting point for the GAN design.

The bipartite transformer offers a solution to the StyleGAN limitation in its ability to perform spatial decomposition which led to the impossibility of controlling the style of a localised regions within the generated image.

Instead of controlling the style of all features globally, the new attention layer is used to perform *adaptive region-wise modulation*. The latent vector z is split into k components, $z = [z_1, \dots, z_k]$ and, as in StyleGAN [3], pass each of them through a shared mapping network, obtaining a corresponding set of intermediate latent variables $Y = [y_1, \dots, y_k]$.

During synthesis, after each CNN layer in the generator, the feature map X and latents Y play the roles of the two element groups, mediating their interaction through our new attention layer (either simplex or duplex). This setting thus allows for a flexible and dynamic style modulation at the region level.

Since soft attention tends to group elements based on their proximity and content similarity, the transformer architecture naturally fits into the generative task and proves useful in the visual domain, allowing the model to exercise finer control in modulating local semantic regions. This capability turns to be especially useful in modelling highly-structured scenes.

For the discriminator, attention is applied after every convolution using trained embeddings to initialise the aggregator variables Y , which may intuitively represent background knowledge the model learns about the task. At the last layer are concatenated these variables Y to the final feature map X to make a prediction about the identity of the image source. This structure empowers the discriminator with the capacity to likewise model long-range dependencies, which can aid it in its assessment of the image fidelity, allowing to acquire a more holistic understanding of the visual modality.

- Compositional Latent Space with multiple variables that coordinate through attention to produce the image cooperatively, in a manner that matches the inherent compositionality of natural scenes.
- Bipartite Structure that balances between expressiveness and efficiency, modelling long-range dependencies while maintaining linear computational costs.
- Bidirectional Interaction between the latents and the visual features, which allows the refinement and interpretation of each in light of the other.
- Multiplicative Integration rule to impact the features' visual style more flexibly, akin to StyleGAN but in contrast to the transformer network.

4 Implementation

The code has been implemented using code and information provided by the authors.

4.1 Datasets

The original paper [1] explored the GANformer model on four datasets for images and scenes: CLEVR [7], LSUN-Bedrooms [8], Cityscapes [9] and FFHQ [3].

Initially, we tried to use the Cityscapes dataset, since it is the smaller among the four: it contains 24998 images with 256x256 resolution. However, the time to complete the training was too high on this dataset (TODO: how much?), even if using computational resources like GPU.

For this reason, we switch to another dataset, the Google Cartoon Set [10]¹, containing 10k 2D cartoon avatar images with 64x64 resolution, composed of 16 components that vary in 10 artwork attributes, 4 colour attributes, and 4 proportion attributes (see in Table 1).

¹<https://google.github.io/cartoonset>

Table 1: Attributes of the Cartoon Set <https://google.github.io/cartoonset/download>.

		# Variants	Description
Artwork	chin_length	3	Length of chin (below mouth region)
	eye_angle	3	Tilt of the eye inwards or outwards
	eye_lashes	2	Whether or not eyelashes are visible
	eye_lid	2	Appearance of the eyelids
	eyebrow_shape	14	Shape of eyebrows
	eyebrow_weight	2	Line weight of eyebrows
	face_shape	7	Overall shape of the face
	facial_hair	15	Type of facial hair (type 14 is no hair)
	glasses	12	Type of glasses (type 11 is no glasses)
	hair	111	Type of head hair
Colors	eye_color	5	Color of the eye irises
	face_color	11	Color of the face skin
	glasses_color	7	Color of the glasses, if present
	hair_color	10	Color of the hair, facial hair, and eyebrows
Proportions	eye_eyebrow_distance	3	Distance between the eye and eyebrows
	eye_slant	3	Similar to eye_angle, but rotates the eye and does not change artwork
	eyebrow_thickness	4	Vertical scaling of the eyebrows
	eyebrow_width	3	Horizontal scaling of the eyebrows

4.2 Hyper-parameters

Note that in the code provided by the author [1], the hyper-parameters are not the same mentioned in the article.

A *kernel size* of $k = 3$ is used after each application of the attention, together with a *Leaky ReLU non-linearity* after each convolution and then upsample or downsample the features X , as part of the generator or discriminator respectively, as in e.g. StyleGAN2 [2]. To account for the features location within the image, we use a sinusoidal positional encoding along the horizontal and vertical dimensions for the visual features X , and trained positional embeddings for the set of latent variables Y . Overall, the bipartite transformer is thus composed of a stack that alternates attention (simplex or duplex), convolution, and upsampling layers, starting from a 4×4 grid up to the desirable resolution.

Both the simplex and the duplex attention operations enjoy a bilinear efficiency of $\mathcal{O}(mn)$ thanks to the network’s bipartite structure that considers all pairs of corresponding elements from X and Y . Since, as we see below, we maintain Y to be of a fairly small size, choosing m in the range of 8–32, this compares favourably to the prohibitive $\mathcal{O}(n^2)$ complexity of self-attention, which impedes its applicability to high-resolution images.

As to the loss function, optimization and training configurations, we adopt the settings and techniques used in StyleGAN2 [2], including in particular style mixing, stochastic variation, exponential moving average for weights, and a non-saturating logistic loss with a lazy R1 regularization.

4.3 Experimental setup

The source code of our work is available at the following GitHub repository: <https://github.com/GiorgiaAuroraAdorni/gansformer-reproducibility-challenge>.

The approaches proposed in both the original paper codebase by Karras et al. [2] and by Hudson and Zitnick [1] have been implemented in Python using TensorFlow [11], so, according to that, we used the same setup. We created a Jupyter Notebook which runs all the experiments in Google Colaboratory, which allows us to write and execute Python in the browser.

All the models have been trained on a Tesla P100-PCIE-16GB (GPU) provided by Google Colab Pro.

4.4 Computational requirements

In the original paper [1], they evaluate all models under comparable conditions of training scheme, model size, and optimization details, implementing all the models within the codebase introduced by the Style-GAN authors [2]. All models have been trained with images of 256×256 resolution and for the same number of training steps, roughly spanning a week on 2 NVIDIA V100 GPUs per model (or equivalently 3-4 days using 4 GPUs).

Considered that we had available just one GPU and not enough time to reproduce this settings, we decided to resize the images from 256x256 to 64x64 resolution.

For the GANformer, we select k —the number of latent variables, from the range of 8–32.

All models have been trained for the same number of steps, which ranges between 5k to 15k king training samples.

The paper present results after training 100k, 200k, 500k, 1m, 2m, 5m and 10m samples

For the StyleGAN2 model we present results after training 100k (2h30m), 200k (6h), 500k, obtaining good results. Note that the original StyleGAN2 model has been trained by its authors [2] for up to 70k king samples, which is expected to take over 90 GPU-days for a single model.

For the GANformer, the authors [2] show impressive results, especially when using duplex attention: the model manages to learn a lot faster than competing approaches, generating astonishing images early in the training. This model is expected to take 4 GPU-days.

However, we are not able to replicate this achievements, first because this model learns significantly slower than the StyleGAN2, which is instead able to produce high-quality images in approximately x -times less training steps than the GANformer.

(in the paper they reach better results with the GANformer with 3-times less training steps than the StyleGAN2, but they don't specify the time required for this training steps...)

5 Results

Sample images from different points in training are based on the same sampled latent vectors, thereby showing how the image evolves during the training.

For CLEVR and Cityscapes, we present results after training to generate 100k, 200k, 500k, 1m, and 2m samples. For the Bedroom case, we present results after 500k, 1m, 2m, 5m and 10m generated samples during training

These results show how the GANformer, especially when using duplex attention, manages to learn a lot faster than competing approaches, generating impressive images early in the training

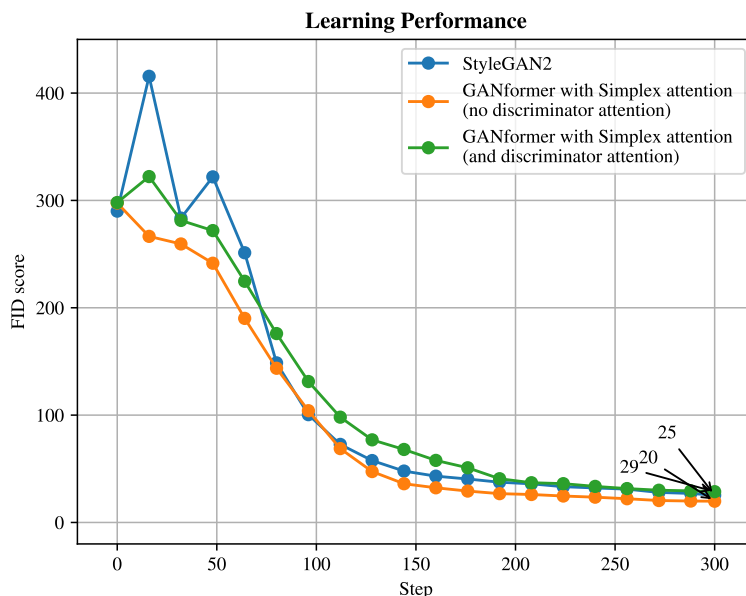


Figure 1: **Comparison between the StyleGAN2 and the GANformer models.** We evaluate the models according to FID score along 300k image samples. The score is computed every sixteen checkpoints.

The Frechet Inception Distance (FID) is one of the most popular metrics for evaluating GANs, which provides stable and reliable indications of *image fidelity* and *diversity*. It is a measure of similarity between curves that takes into

account the location and ordering of the points along the curves. For this specific application, FID is used to measure the feature distance between the real and the generated images, but it can be used for measuring the distance between two distributions as well.



(a) StyleGAN2.

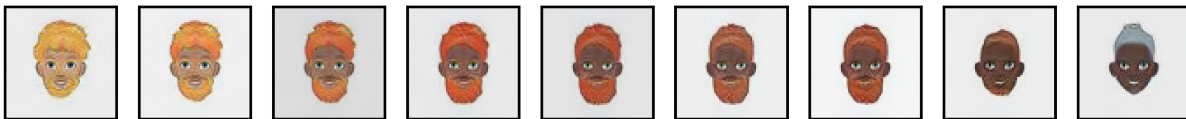


(b) GANformer model with Simplex attention and discriminator without attention.

Figure 2: **Visualisation of 9 images generated with the various models.**



(a) StyleGAN2.



(b) GANformer model with Simplex attention and discriminator without attention.

Figure 3: **Simple z interpolation using of the various models.**

Visualisation of the models' comparison throughout the training.

6 Discussion and conclusion

References

- [1] Drew A. Hudson and C. Lawrence Zitnick. Generative Adversarial Transformers. 2021. URL <http://arxiv.org/abs/2103.01209>.
- [2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. URL <http://arxiv.org/abs/1912.04958>.
- [3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. URL <http://arxiv.org/abs/1812.04948>.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, page 2672–2680. MIT Press, 2014. URL <http://arxiv.org/abs/1406.2661>.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. URL <http://arxiv.org/abs/1706.03762>.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. URL <http://arxiv.org/abs/1810.04805>.
- [7] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. URL <http://arxiv.org/abs/1612.06890>.
- [8] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 2015. URL <http://arxiv.org/abs/1506.03365>.
- [9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. URL <http://arxiv.org/abs/1604.01685v2>.
- [10] Cole Forrester, Mosseri Inbar, Krishnan Dilip, Sarna Aaron, Maschinot Aaron, Freeman Bill, and Fuman Shiraz. Cartoon set. <https://google.github.io/cartoonset/>.
- [11] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.