Faculty of Mathematics and Computer Science

Machine learning course (ML)

Applications of Generative Adversarial Networks in images style transfer: A survey

Liviu-Ștefan Neacșu-Miclea

Department of Computer Science, Babeş-Bolyai University
1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania
E-mail: nlic3194@scs.ubbcluj.ro

Abstract

Image style transfer refers to the process of redrawing an input source image in the style of another reference image while preserving the content of the original image. Generative Adversarial Networks (GANs) have an established history in image generation, many of them being proposed as solutions for particular image-to-image tasks. In this paper, we conduct a literature survey on the performance of GANs in the visual style transfer field and the consequences their development had in improving our understanding on how these models learn data distributions and effective ways to control this process. This paper is organized in five sections. First, we start by formally introducing the style transfer problem. Then, we explain the reasoning behind the need for such a survey. Next, we recall the basic workings of a GAN and Neural Style Transfer (NST) concepts before we cover a few GAN architectures capable of performing style transfer: the StyleGAN featuring high-quality results and architectures designed for unpaired images transfer (CycleGAN, CartoonGAN, GANILLA, Andersson and Arvidsson's GAN generating pictures in Hayao Miyazaki's style , AniGAN). Subsequently, we review some GAN specific evaluation methods and discuss each architecture's points of interest from a critical viewpoint. We conclude this study with an overall perspective over the discussed subject, stating the points of success and suggesting possible directions for improvement in this matter.

Keywords: Generative Adversarial Networks; Image to Image Translation; Style Transfer

1. Introduction

Style transfer problems focus on generating new data based on two sets of samples, such that the result would feature content belonging to the first set, presented in a style specific to the second set. This follows Gatys et al.'s revolutionary findings that the content and style of an image are separable [5]. Two main subclasses of style transfer neural network models have been proposed, considering the way their inputs are structured, whether in a paired or unpaired manner [1]. Needless to say, datasets providing paired real images with corresponding artistic interpretations

are rare and small in size, since their creation is a monumental task that requires a professional hand and countless hours of work for a single sample [4].

Recently, a lot of effort was put into engineering unsupervised learning solutions for artistic image generation, namely artistic style transfer. With the astonishing advances in generative models, researchers were able to produce Generative Adversial Networks that learn the defining look and feel features of the style set and consequently apply them to the input's content in the same way a human would imagine how a natural landscape would look like if painted by a renowned painter [15]. This survey visits a handful of interesting state-of-the-art GAN applications in image style transfer domains.

This work consists of five sections, starting with this current introductory part. The second section argues the importance of studies like this one in the field. The next section recalls the theoretical concepts of GAN and NST and briefly introduces the architectures proposed for discussion (CycleGAN, StyleGAN, CartoonGAN, GANILLA, Andersson's Miyazaki-style GAN, AniGAN), highlighting some of their original contributions or important mathematical concepts, ending with mentioning the general difficulties in evaluating generative networks and what popular solutions have been proposed for that. Section 4 attempts a comparative in-depth analysis of the first five models in terms of architecture, used datasets, experimental setup, performance and results. It was decided for AniGAN to be discussed separately because of its contrastive complexity relative to the other models. The final section concludes the findings of this study, also proposing future improvement steps for the quality and reliability of the survey.

2. Motivation

Unsupervised GANs serve as a powerful tool to automate artistic drawing where in situations where time or volume of work would become an impediment and traditional software editors would often fail to create satisfactory results [13]. The main advantage of GANs over other style transfer models is their adaptability which makes them capable of producing more sophisticated results than the traditional texture transfer approached in older studies [5] [8].

Furthermore, style transfer has given birth to valuable experiments on self-supervised representation learning (regarded as a black-box process in GANs) and feature disentangling [12]. This is a popular research topic in GANs due to the quest to understand how the model handles information in order to manipulate its outcome to the desired form [3] [10]. This can prove useful in style transfer applications which require non-trivial inter-domain morpholohical transformations [13].

3. Related Work

3.1. Generative Adversial Network

The generative adversial network (GAN) was introduced by Goodfellow et al. in 2014 as a way of modelling the generative model in the form of a cop and counterfeiter game [6]. A GAN consists of a pair of networks (G, D). A generator G is trained to produce a sample $y = G(z)$ where z is sampled from a noise distribution p_z, while a discriminator D is trained to predict the probability of its input to be a real instance of the dataset following a distribution p_x or a fake result of the generator. In principle, the generator will eventually create samples good enough to fool the discriminator into thinking they are genuine [6]. The rivality between G and D is reflected in the adversarial loss by the so called mini-max game [6]:

$$\mathcal{L}_{GAN} = E_x(log(D(x))) + E_z(log(1 - D(G(z))) \text{ [6].}$$

This loss is minimized by G and maximized by D. The authors noted that $log(1 - D(G(z))$ saturates the generator because of inducing a too low gradient, and proposed as a fix to maximize $log(D(G(z))$ instead of minimizing $log(1 - D(G(z))$ [6]. While this is not the focus of the current work, it is worth mentioning that GANs training is a difficult process that is fighting a number of issues, such as mode collapse (the model optimizing over a single sample and only generating it repeatedly), vanishing gradients, convergence problems with roots in the dynamic nature of generator and discriminator synchronization.

3.2. Neural Style Transfer

The aim in a style transfer problem is that, given two images x and z, one must generate an image y which takes the content of x depicted in the style of z [14].

Gatys et al. proved in their work that the content and the style of an image can be separated using a sufficiently capable convolutional neural network (the one used in their method is VGG-16) [5]. Let $c_1, c_2, \ldots, C_{\mathbf{L}}$ denote the layers of the convolutional network with $\mathbf{L}$ layers. We define $C^l : \mathcal{I}^{N \times N} \to \mathcal{F}_l^{N_l \times M_l}$, $C^l(u) = (c^l \circ ... \circ c^2 \circ c^1)(u)$ a function that provides the features representation of an input u after the l-th layer of the CNN, where I is the space of per-pixel input features ($I = \mathbb{R}^3$ for RGB images), $\mathcal{F}_l$ is the space of per-pixel filter activations (e.g for a CNN where the first layer is a 32-filters convolution, $\mathcal{F}_1 = \mathbb{R}^{32}$), and N_l, M_l, F_l, are the number of rows, columns, respectively channels of the l-th feature map.

Gaty's proposed approach is an iterative method that optimizes the generated y, which is initially noise, to minimize two losses: a content loss and a style loss [5]. The content loss is therefore

$$\mathcal{L}_{content}(x, y) = \frac{1}{2} \sum_{i,j} \left(C^l_{i,j}(x) - C^l_{i,j}(y) \right)^2, [5]$$

where l is the layer up to which we want to replicate the content representations [14].

The style loss uses the Gram matrix $G : \mathcal{F}_l^{N_l \times M_l \times F_l} \to \mathcal{F}_l^{N_l \times N_l \times F_l}$, $G^l(u) = C^l(u) \cdot \left(C^l(u)\right)^t$ (where the matrix operations are performed per-channel) in order to compute the feature correlations inside an image u [5]. The style loss compares at each layer l the correlations of the style source with the ones of the generated image, namely $E_l(z, y) = (2N_l M_l)^{-2} \sum_{i,j} (G^l_{i,j}(y) - G^l_{i,j}(z))^2$, and cummulates them in a weighted sum with meta-parameters vector w which gives each representation a degree of importance to the style:

$$\mathcal{L}_{style}(z, y) = \sum_{l=0}^{L} w_l E_l(z, y) \ [5].$$

The total loss $\mathcal{L}_{total}(x, z, y) = \alpha \mathcal{L}_{content}(x, y) + \beta \mathcal{L}_{style}(z, y)$ specifies to which degree y should inherit the content more than the style. Higher values of the $\frac{\alpha}{\beta}$ ratio are better in preserving the content, while a lower ratio focuses more on the style [5].

3.3. Adaptive Instance Normalization

Conventionally, normalization refers to the process of transforming a distribution in a way that it has a certain mean and variance. In the context of feature maps, instance normalization with mean γ and standard deviation β is performed in the 2D space for each feature channel and each sample independently [8]:

$$IN(x_{b,i,j,c}) = \beta \left(\frac{x_{b,i,j,c} - \mu(x_{b,*,*,c})}{\sigma(x_{b,*,*,c})} \right) + \gamma, \ \forall \ i = 1..N, j = 1..M,$$

where $\mu(x_{b,*,*,c}) = \frac{1}{NM} \sum_{i,j} x_{b,i,j,c}$, $\sigma(x_{b,*,*,c}) = \sqrt{\frac{1}{NM} \sum_{i,j} (x_{b,i,j,c} - \mu(x_{b,*,*,c}))^2 + \varepsilon}$ [8].

In contrast, instead of depending on fixed parameters, the Adaptive Instance Normalization (AdaIN) brings the data x to the mean and variation of a reference sample y:

$$AdaIN(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y) \ [8].$$

The AdaIN has proved its efficiency in style transfer. It has been used by Huang et al. [8] to mix feature maps extracted from content and style image sources before using the result as a latent space for a decoder which generates the transfered image — see Fig. 3.

3.4. StyleGAN

The StyleGAN is a substantial improvement over the classical GAN architecture in the form of a direct application of AdaIN contraptions to blend the style encoded in the latent space into the generated partial outputs on an interative synthesis process of stacked convolutions [12]. This model allows for "style mixing", which means swapping the latent

code at some point in the network in order for the synthesis network to take information from multiple sources [12]. The method also includes stochastic variation, which means injecting noise before every AdaIN operation which alter local features (e.g. hair and freckles distribution) without changing the general picture (e.g. pose, facial expression) [12].

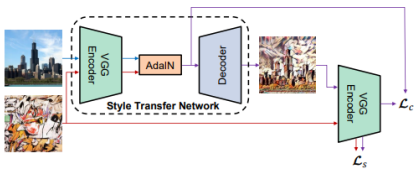

Fig. 1: Architecture of AdaIN style transfer architecture [8]

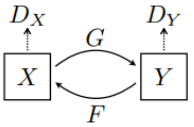

Fig. 2: CycleGAN architecture [15]

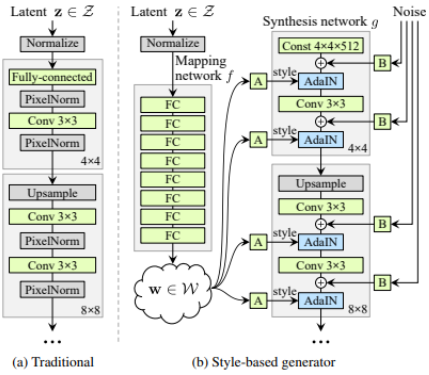

Fig. 3: Side by side comparison of GAN and StyleGAN architectures[12]

3.5. CycleGAN

The Zhu et al's CycleGAN addresses the problem of unpaired image-to-image translation arising in situations when for an image in a certain domain there is no direct correspondent (or is hard to find one) in other domain [15]. The working principle relies on the fact that, is there is a translation $G : X \rightarrow Y$ from an image domain X to another Y, there should exist the inverse mapping $F : Y \rightarrow X$ such that $F(G(x)) \approx x$ and $G(F(y)) \approx y$ [15] (see Figure 2). The CycleGAN trains two GANs (G, D_G) and (F, D_F) and, besides the usual adversarial loss, a cycle-consistency loss is introduces to penalize arbitrary mappings between domains:

$$\mathcal{L}_{cyc}(G, F) = E_x(\|F(G(x)) - x\|_1) + E_y(\|G(F(y)) - y\|_1) \text{ [15].}$$

3.6. CartoonGAN

The CartoonGAN is a proposed image-to-image GAN architecture that aims for cartoonization of real world pictures by enhancing the simplicity and the edgy look of such art styles [4]. It was designed to provide a better alternative to the cyclic strutcure of CycleGAN [7]. The work features training the discriminator with patches of both real cartoon images $c \in C$ and an edge-blurred version of them $e \in \mathcal{E}$, as well as the set of real pictures $p \in \mathcal{P}$:

$$\mathcal{L}_{edge\_promoting\_adv} = E_c(log(D(c)) + E_e(log(D(e))) + E_p(log(1 - D(G(p)))) \text{ [4].}$$

Moreover, the total loss $\mathcal{L}_{total} = \mathcal{L}_{edge\_promoting\_adv} + \omega\mathcal{L}_{content}$ includes a weighted content loss employed on the deep activations of the VGG network applied on real and generated data to ensure semantic preservation:

$$\mathcal{L}_{content} = E_p(\|VGG(G(p)) - VGG(p)\|_1) \text{ [4].}$$

3.7. GANILLA

GANILLA is a GAN designed for illustration unpaired style transfer which tackles to address the weak points of other NST and generative methods in minimizing the loss of both content and style [7]. GANILLA uses a ResNet-based generator features skip-connections on the down-up-sampling stage and a PatchGAN 70x70 discriminator, while the training process is similar to the CycleGAN one [7].

3.8. MiyazakiGAN

Andersson's and Arvidsson's GAN architecture, which will be hereafter reffered by as MiyazakiGAN due to lacking an official name, is a direct application of CartoonGAN in the task of generating cartoons as drawn by Hayao Miyazaki, an artist at Studio Ghibli [1]. The work brings minor modifications to the CartoonGAN model, but stands

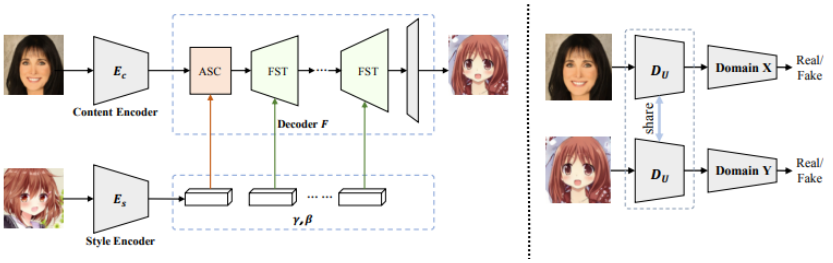

Fig. 4: Anigan generator and double branch discriminator architecture [13]

out because of the larger scale of the dataset and confirms the successful applicability of CartoonGAN in other cartoonization processes [1].

3.9. AniGAN

The face to anime stylization task encounters the problem of needing to transfer local shapes along with chromatic and textural information [13], something which above mentioned GAN architectures fail to do so, as they try to preserve global structure of the content and present it in another style - remark the CycleGAN's failure cases, like dog vs. cat or zebrification [15]. AniGAN's design (Figure 4) addresses this issue, moreover, while the other architectures perform generic inter-domain $X \rightarrow Y$ translation, the latter explicitly ensures that the output should feature the style of a specifically pinpointed donor instance $y \in Y$ [13].

The network introduces a handful of new high level layers like Adaptive Stack Convolutional (ASC) and Fine-Grained Style Transfer (FST) blocks for style injetion based on a so-called point-wise layer instance normalization (PoLIN and its adaptive version AdaPoLIN), defined by a convolution between layer and instance normalization maps. The discriminator takes advantage of the fact that that real and drawn faces should share similar characteristics, hence the idea for a double branch discriminator with shared weights followed by domain-specific extractors.

3.10. GAN Evaluation methods

Numerical measuring standards for generative methods and style transfer performance is yet to have reached a common agreement point among researchers [11] [1] [9]. The original GAN study uses a comparison between original and generated distributions by Parzen-window-based log-likelyhood metric, with the remark that the high variance it produces makes it unfit for high dimensional data [6]. Researchers often resort to qualitative surveys where humans are asked to distinguish between real and fake images [9] or score the presence of other style attributes [1]. As an automated quantitative metric, the FCN-score was introduced in the context of segmentation mask to image generation, under the intuition that if the generated images mimic reality well enough, a pre-trained classifier would be able to correctly identify semantics in the pictures [9]. The accuracy of such a classifier is defined as the FCN-score [9]. Other CNN variants imply perceptual analysis that quantifies content and style losses [7] [13]. Fréchet inception distance (FID) is another adopted metric, used in StyleGAN [12]. Other encountered approaches imply using image-specific metrics from the frequency analysis domain, such as SSIM and PSNR [11].

3.11. Experimental Results

This sections reports the models performances as stated in the literature. CycleGAN obtained 24-29.8%, respectively 18.8-26.6% real-vs-fake user test accuracy on map to/from aerial dataset, as opposed to at most 3% of other traditional GAN models [15]. StyleGAN reports 4.40 FID on their own faces dataset [12], while other study [2] provides 21.7 FID and 5.02 IS for StyleGAN, while also testing CycleGAN on Horse2Zebra and Summer2Winter tasks with an average of 3.67 IS and 71.8 FID [2]. CartoonGAN paper only provides visual analysis of results [4]. While GANILLA outperforms Cycle and CartoonGAN in their own perceptual metrics. More details are included in the next section, which shows and simultaneously compares the evaluations of these GAN models.

4. Discussion

4.1. Design and training specifications overview

	CycleGAN [2]	**StyleGAN** [2]	**CartoonGAN**	**GANILLA**	**MiyazakiGAN**
Generator	Johnson et al.	style injection	residual-blocks based	ResNet-18 with skip connections upsampling	CartoonGAN's
Discriminator	PatchGAN	ProGAN's	PatchGAN		
Losses	cycle consistency, identity, adversarial	adversarial, style mixing,perceptual	"edge-promoting" adversarial, perceptual	minimax, cycle consistency	CartoonGAN's

Table 1: Design summaries per model

Table 1 provides technical information about the mentioned GAN architectures. The models adopting the cycle-consistency approach (CycleGAN and GANILLA) feature two distinct pairs of generator and discriminator. CartoonGAN and its direct application MiyazakiGAN follow the original GAN route with a generator design that best fits the studied problem, while StyleGAN completely turns away from the established generator model in image-to-image tasks (when the input is usually directly a noise distribution or an image source) in favor of the ability to control style and stochasticity at any depth of the generation. Karras et al. even conduct an impressive study on feature entanglement and linear separability in the feature space, which guides and supports their style swap and mixing idea [12].

	CycleGAN [2]	**StyleGAN** [2]	**CartoonGAN**	**GANILLA**	**MiyazakiGAN**
Dataset(content)	Horse2Zebra, Summer2Winter	FFHQ*, LSUN	Flickr scraps*	CycleGAN's	Flickr30K
Dataset(style)			M. Shinkai & M. Hosoda art*	illustrators' work*	H. Miyazaki art*
Epochs	200	1000	200	200	60
Learning rate	$2 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$2 \cdot 10^{-4}$ **2	$2 \cdot 10^{-4}$	$1 \cdot 10^{-3}$ **1
Batch size	1	8	16 **2	1	11 **1
GPUs (NVIDIA)	2x Tesla V100	4x Tesla V100	Titan XP	Tesla V100	RTX 2080
No.pms.[7]	11.4mil	26.2mil[12]	11.1mil	7.2mil	N/A
Train.time[7]	1347s	1 week[12]	1400s	887s	144h[1]

Table 2: Experimental summaries per model
* dataset created by the authors
** unspecified in the paper, found in public implementations:
[1] https://github.com/FilipAndersson245/cartoon-gan/blob/master/experiment.ipynb
[2] https://tobiassunderdiek.github.io/cartoon-gan/#tc6

Table 2 provides a side by side view of experimental setups used to train the five GANs. CycleGAN aims to general usage by being trained on a remarkable amount of unpaired datasets (real vs. Monet/Van Gogh, apple to oranges, dog to cat, horse to zebra, photo to DSLR). The cartoon/illustration style transfer GANs expectedly use real world images as content source and a collection target artists' works as style reference. As visible on the project's website, CycleGAN also receives a lot of attention from the community, seeing applications in grayscale photo colorizing, bears to pandas or comical face to ramen translations [15]. On the other hand, StyleGAN seems to have less diverse applications

at a brief lookup into the matter. The authors focus intensively on face generation with undoubtedly hard to equal results, while also trying their architecture on LSUN cards and bedrooms [12]. In the case of StyleGAN, as opposed to the other models, the border between content and style providers is volatile, since the mapping network responsible for mixing styles works like a black-box mechanism, hence the images used in training are sampled from the same domain. The complexity of the solution also explains the higher number of epochs and training time over the span of multiple days, the latter being comparable only with MiyazakiGAN which, however, might be the consequence of using a weaker GPU. Moreover, CartoonGAN's authors noted that pretraining the generator on real images in order to help with adapting the weight to replicate the contents makes for a remarkable boost to the model's convergence time [4]. On the discriminator side, PatchGANs are the most popular option since they are an easy and fast way to evaluate high resolution images, except for the StyleGAN which is built upon a ProgressiveGAN-like framework.

4.2. Evaluation, performance analysis and comparison

	CycleGAN	**StyleGAN**	**CartoonGAN**	**GANILLA**	**MiyazakiGAN**
User Study	Illus.[7], M2A[15], O/CST[11]	-	Illus.[7], F2S/H*[4], MiyaST[1]	Illus.[7], O/CST[11], MiyaST[1]	MiyaST[1]
Style/Cont. CNN	Illus.[7]	-	Illus.[7]	Illus.[7]	-
FCN-score [9]	cityscapes[15]	-	-	-	-
FID	H2Z,S2W[2]	FFHQ[2],[12]	-	-	-
IS	H2Z,S2W[2]	FFHQ[2]	-	-	-
SSIM	O/CST[11]	-	O/CST[11]	O/CST[11]	-
PSNR	O/CST[11]	-	O/CST[11]	O/CST[11]	-

Table 3: Metrics and datasets used to evaluate the models, found in literature
Datasets: M2A=map2aerial, H2Z=horse2zebra, S2W=summer2winter,
Illus.=GANILLA's illustrations dataset, F2S/H=Flickr to Shinkai/Hayao style,
O/CST = oil and cartoon style transfer datasets [11]
* = visual comparison, no numerical data; - = no data

In terms of performance evaluation, the reported metrics and their associated datasets are aggregated in Table 3. As it is observed, half of the measurements denote qualitative analysis performed manually, either by visually assessing the generated result [4], or by writing questionaries where people were asked to distinguish between real and fake samples (the CycleGAN's AMT process [15]), grade pictures' particularities (aesthetic/cartooniness [1], yes/no approval of style transfer [7][11] and ranking of content/appeal [7], aestetic/similarity [11]). Figure 5 shows the percentual results of user studies conducted by different experiments. For MiyazakiGAN's aestetichs and correctness, which are computed as a mean of four ranking user scores (1,4), the formula $percentual\_score = (1 - score/4) \cdot 100$ was used in order to bring the values to the same range as the others. The subjective nature of such evaluation results in some metrics placing the models in different orders. An example would be the aestetic factor used by Jiang et al. [11] and Andersson et al. [1] on GANILLA and CartoonGAN. The latter admitted that humans preferred the aestetics of GANILLA instead of their own model, MiyazakiGAN, not due to a performance fault, but simply because the human eye found the less cartoonish images more pleasing than an accurate cartoonization, hence the lack of common views for the meaning of aestetic altered the interpretability of the evaluation [1].

The quantitative analysis often sees utilization of distribution comparing metrics like Fréchet inception distance (FID) or the inception score (IS) (see Figure 6), or other established image quality metrics, such as the Peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM). The involvement of pretrained perceptual models is featured in a few measurements. For example, CycleGAN uses a fully convolutional network (FCN) to assess the fidelity of the generated image to a provided input mask (in the case of label to cityscapes generation), in idea that a good segmentation network applied on this image would predict approximately the same labels map as the one that

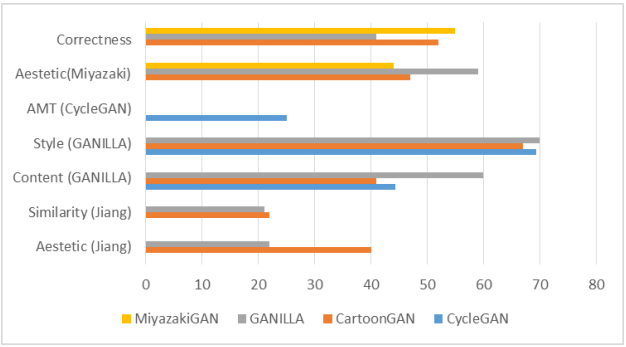

Fig. 5: User studies conducted on the GAN architectures

was used as a generator input in the first place [15]. An approach more in resonance with the style transfer literature is GANILLA's style and content CNN classifiers [7]. Figure 7 depicts the reported metrics from the the illustrators dataset introduced with GANILLA [7]. It can be deducted that both style and content can not be simultaneously preserved, and GANILLA finds the sweet spot just enough to slightly have better average performance (with a subunitary difference) than the already almost tightly competing CycleGAN and DualGAN.

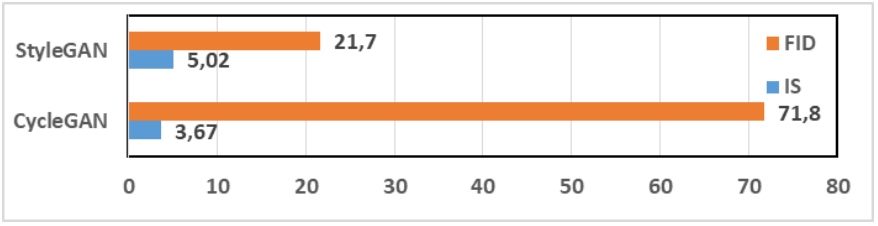

Fig. 6: Quantitative comparison between StyleGAN and CycleGAN [2]

Figure 6 shows numerical results of a comparison study between CycleGAN and StyleGAN [2]. The two models were not tested the same dataset, but the ones used in each of the corresponding networks' original papers: high resolution faces (FFHQ) were used with StyleGAN, while the metrics for CycleGAN were obtained by averaging the measurements on Horse2Zebra and Summer2Winter. The SSIM and PSNR metrics were computed by a side study on oil painting and cartoonization style transfer using ghost module which compares CartoonGAN and GANILLA with their approach, outperforming both of them on an average evaluation of SSIM and PSNR performed on 5 images [11].

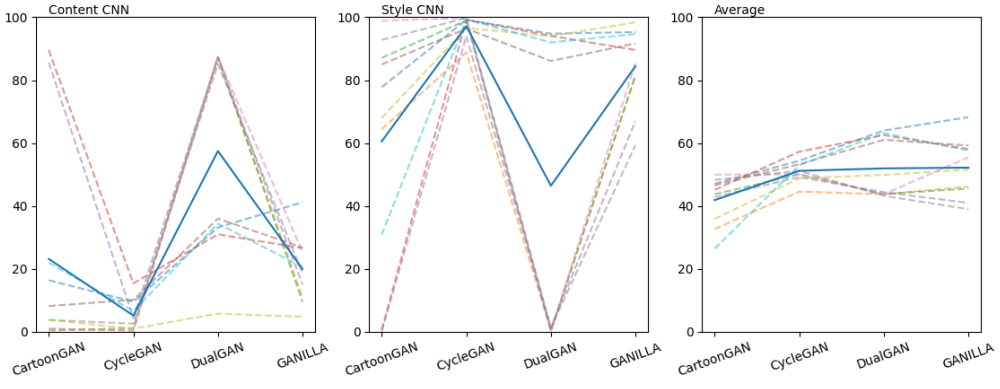

Fig. 7: CNN metrics comparison between different GAN architectures [7]
dashed lines = measure per illustrator style; continuous line = average measures per entire dataset

Fig. 8: Failure cases extracted from [15], [7], [1]

While a numerical support is crucial in evaluating and especially comparing two models, the most objective viewpoint also includes explicit vizualization of generation samples. While an image generation might be consistent enough to score optimally in the most popular metrics, blatant structural issues that cannot fool the human eye might pass through the filters and statistics of any measuring method. Examples of failures are shown in figure 8. In the case of CycleGAN, while the shape of reality is good enough, the generated content may look unplausible (best seen in the zebrification phenomenon), which happens in the lack of the network ability to translate geometries along with textures [15]. The anomalies in cartoonization, like unnatural colors or facial elements inference in untipical places, are signs of possible limitations in the model's receptivity to the level of detail (for example, large almost empty panoramic landscapes lead to better results than crowdy surfaces of any type) [7].

4.3. AniGAN summary

The geometry style transform problem is approached in the design of AniGAN generator, which is a network architecture in a class of its own that clearly separates from the ones analyzed so far. For this reason, it was chosen to discuss it independently the others. The training is performed on the unpaired selfie2anime dataset, and additionally the authors' custom dataset face2anime [13]. The adversarial loss resembles the cycle consistency pattern, but with the same generator for both face and anime domains [13]. Alongside it, there is a feature matching loss defined in a perceptual manner over the common (shallow) part of the discriminator, which helps the model match features from various scales [13]. In addition, a reconstruction loss is present and guides the model into approximating the identity function when the same input is used as both the source and target images [13]. AniGAN is a lightweight model, taking almost half a day to train for 100,000 epochs With a batch size of 4 and a TeslaV100 GPU [13].

For state-of-the-art comparison, AniGAN is tested against other generative models, among which the only generative adversial networks are CycleGAN and a certain StarGAN-v2 [13]. The paper highlights the structural complexity of face2anime dataset distributions compared to the horse2zebra used by the CycleGAN network [13].

Both qualitative and quantitative analysis are performed. The qualitative part focuses on plain visual comparison between outputs from different models. This reveals that while some of the models can generate high quality samples, they do not reflect the content of the real image, while the CycleGAN, altgough managing to find the usual equilibrium between content and style, might struggle with the general features that define an anime face [13]. The only numerical supplement is a statistic of "preference percentanges" from 20 surveyed people, which shows that more than 80% of styled transforms from both selfie and face datasets are on the subjects' liking (the highest from the compared methods) [13]. The quantitative metrics used are FID and the Learned Perceptual Image Patch Similarity (LPIPS), a kind of perceptual distance with a standard implementation [13]. Visual bar diagrams of the metrics can be consulted in Figure 9. As expected, the specialized architectural quircks of AniGAN boosts its ability to replicate the style it was designed for.

5. Conclusions and future work

Despite the impressive efforts and outstanding results, general purpose style transfer GANs remain an open problem. As it was observed, every single problem has its own designated solution that tries to maximize to performance

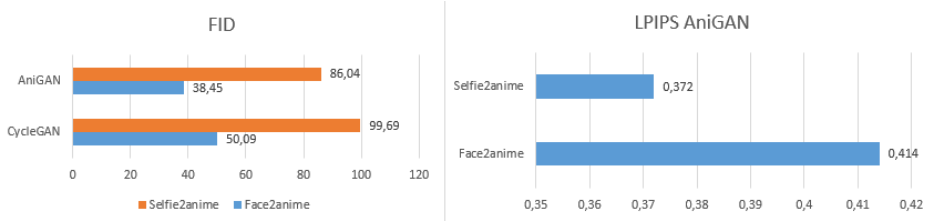

Fig. 9: AniGAN quantitative metrics

on the domain of interest. State of the art cartoonization models may perform worse when attempting to train to replicate other styles such as oil painting, book illustration or even other artistic footprints. On the other hand, some style transfer models are not robust to morphing translations and are susceptible to hallucinations when faced with diverse and highly granular data. On the optimistic side, style transfer problems have lead to a lot of progress being made in the field of feature analysis and disentanglement, and brilliant and creative ideas arise from sometimes pure intuitions based on a globally dominant aspect of a dataset, such as the sharp lines and simplicity of the cartoons. From StyleGAN's ingenuity to CartoonGAN's efficiency to AniGAN's impressive accomplishments, style transfer is a wonderful yet still of surprises and uncovered possibilities.

On a long-term thought, the quality of the current work could be substantially improved by trying to run the discussed models in order to complement the results found in literature. A standard evaluation framework should be established, in which all models performance are compared on the same datasets with the same set of metrics for a better interpretation reliability. Last but not least, studying the impact of other generative networks like Variational Autoencoders or Stable Diffusion models might help in completing the overall picture on this beautiful topic.

References

[1] Andersson, F., Arvidsson, S., 2020. Generative adversarial networks for photo to hayao miyazaki style cartoons URL: https://arxiv.org/abs/2005.07702, arXiv:2005.07702.

[2] Ashokan, A., 2024. Comparative analysis of cyclegan and stylegan in unpaired image-to- image translation and high-quality image synthesis. International Research Journal of Engineering and Technology (IRJET) 11, 274–278.

[3] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. Infogan: interpretable representation learning by information maximizing generative adversarial nets, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 2180–2188.

[4] Chen, Y., Lai, Y.K., Liu, Y.J., 2018. Cartoongan: Generative adversarial networks for photo cartoonization, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9465–9474. doi:10.1109/CVPR.2018.00986.

[5] Gatys, L., Ecker, A., Bethge, M., 2016. A neural algorithm of artistic style. Journal of Vision 16, 326–326. doi:10.1167/16.12.326.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63, 139–144. URL: https://doi.org/10.1145/3422622, doi:10.1145/3422622.

[7] Hicsonmez, S., Samet, N., Akbas, E., Duygulu, P., 2020. Ganilla: Generative adversarial networks for image to illustration translation. Image and Vision Computing 95, 103886.

[8] Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1510–1519. doi:10.1109/ICCV.2017.167.

[9] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. CVPR .

[10] Jeong, S., Liu, S., Berger, M., 2022. Interactively assessing disentanglement in gans, in: Computer Graphics Forum, Wiley Online Library. pp. 85–95.

[11] Jiang, Y., Jia, X., Zhang, L., Yuan, Y., Chen, L., Yin, G., 2021. Image-to-image style transfer based on the ghost module. Computers, Materials Continua 68, 4051–4067. doi:10.32604/cmc.2021.016481.

[12] Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4401–4410.

[13] Li, B., Zhu, Y., Wang, Y., Lin, C.W., Ghanem, B., Shen, L., 2021. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. IEEE Transactions on Multimedia 24, 4077–4091.

[14] Li, Y., Wang, N., Liu, J., Hou, X., 2017. Demystifying neural style transfer, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), IJCAI. pp. 2230–2236.

[15] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.