Teaching StyleGAN to Read: Improving Text-to-image Synthesis with U2C Transfer Learning

Vinicius Gomes Pereira
vpereira@inf.puc-rio.br
Jônatas Wehrmann
jonatas.wehrmann@edu.pucrs.br

Information Technology Department
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil

Abstract

Generative Adversarial Networks (GANs) are unsupervised models that can learn from an indefinitely large amount of images. On the other hand, models that generate images from language queries depend on high-quality labeled data that is scarce. Transfer learning is a known technique that alleviates the need for labeled data, though it is not trivial to turn an unconditional generative model into a text-conditioned one. This work proposes a simple, yet effective finetuning approach, called Unconditional-to-Conditional Transfer Learning (U2C transfer). It can leverage well-established pre-trained models while learning to respect the given textual condition conditions. We evaluate U2C transfer efficiency by finetuning StyleGAN2 in two of the most widely used text-to-images data sources, generating the Text-Conditioned StyleGAN2 (TC-StyleGAN2). Our models quickly achieved state-of-the-art results in the CUB-200 and Oxford-102 datasets, with FID values of 7.49 and 9.47 respectively. These values represent respective relative gains of 7% and 68% when compared to prior work. We show that our method is capable of learning fine-grained details from text queries while producing photorealistic and detailed images. Finally, we show that the models structure the intermediate space in a semantically meaningful fashion.

1 Introduction

Generating realistic images from human-written sentences is a challenging research area. In recent years, many novel deep network frameworks to deal with this task have successfully been implemented and these architectures are also rapidly evolving, which increases the potential development of applications to handle real-world problems in the area of image editing, visual effects, and design industry. To address the problem of artificial image synthesis, Generative Adversarial Networks (GANs) [9, 21] have emerged as architecture with promising results [6, 13], and recently, diffusion models [8, 12], as a generative framework that synthesizes images with high variability and trustworthiness.

Generation of images based on detailed natural language descriptions is an even more difficult task, as it inserts the representation of natural language into the problem. It is noteworthy that albeit unconditional GANs are able to learn from an indefinitely large amount

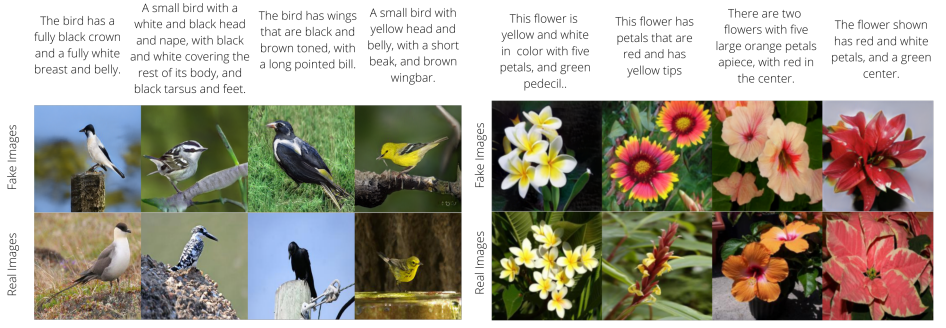

Figure 1: Images synthesized by the proposed approach and the ground truth

of images, text-to-image models depend on high-quality labeled data that is scarce. Transfer learning [25] is a known technique that alleviates the need for labeled data, though it is not trivial to turn an unconditional generative model into a text-conditioned one.

We introduce Unconditional-to-Conditional Transfer Learning (U2C transfer) which allows the use of pre-training information from an unconditional network to generate text-conditioned images. It is able to leverage pre-trained models that generate human faces and adapt them to synthesize high-quality images of Birds [24] and Flowers [18] for instance. Such images do present fine-grained details that respect the input textual query. This method not only stabilizes the training process but also makes the training convergence faster. In addition, note that it is challenging to train a text-conditioned model on these datasets given that they present a rather limited amount of images, which the discriminator easily overfits[4]. It is also complex to get proper textual representations that contain rich details regarding the data distribution of the dataset. Moreover, with few examples of captions, the space of the sentence representation is highly discontinuous, highlighting the need for a better design for such encoder.

U2C transfer is tested in the standard StyleGAN2[13] unconditional models by finetuning them in two widely used datasets. The resulting model, a Text-Conditioned StyleGAN2 (TC-StyleGAN2 for short), effortlessly bests prior work results in a few hundred iterations. We try to follow and reuse the maximum number of pre-trained weights as possible, making only indispensable changes to the architecture. TC-StyleGAN2 also makes use of two data augmentation mechanisms to help preventing overfitting: (i) a textual augmentation technique [33], to increase smoothness and continuity of the conditional text space; and (ii) adaptive discriminator augmentation (ADA) [14] which automatically regulates the strength of the image transformations.

We quantitatively evaluate our models in terms of Fréchet Inception Distance (FID), Kernel Inception Distance (KID), and Inception Score (IS). TC-StyleGAN2 achieved FID of 7.49 and 9.47 for CUB-200 and Oxford-102 Flowers data, respectively. We also evaluate qualitatively, by demonstrating that this model learns regular structures that allow image editing via vector arithmetic operations in both condition and intermediate latent spaces.

The motivation for using the pre-trained models is that there are many more unconditional models than conditional ones, given that they are easier to train, have less complexity, and can leverage from more data. By using more data, one can have higher-quality pre-trained weights. We propose an approach that allows using such high-quality models in datasets that contain very limited amounts of data.

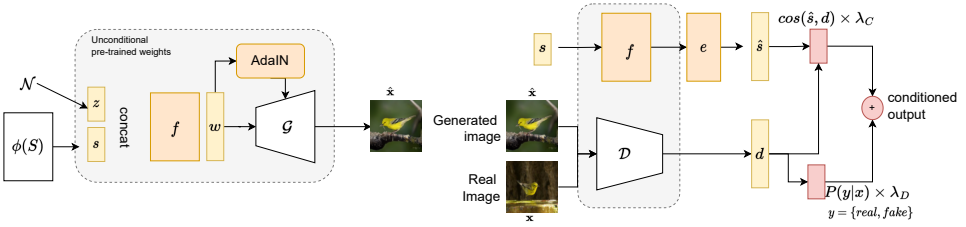

Figure 2: Overall architecture of TC-StyleGAN2.

In summary, we highlight the following contributions:

- We introduce a new transfer learning adaptation technique that involves modifications to inputs, outputs, and loss function, using an unconditional model. Our method has several advantages: it is easy to implement, trains faster and achieves state-of-the-art results in a few iterations in two widely used datasets. Therefore, by using U2C Transfer, potentially all unconditional models can be used for training text-conditional models. To the best of our knowledge, this is the first transfer-learning approach that allows that. The specific modifications we proposed were designed carefully to allow reusing the maximum amount of weights, minimize complexity and avoid training collapse.

- We demonstrate that TC-StyleGAN2 enables image editing using natural language through arithmetic operations not only in the conditional space but also in the intermediate mapping space $\mathcal{W}$, being able to modify several aspects of the image like colors, sizes, and some specific details. We showed that $\mathcal{W}$ space carries a meaningful and learned representation of the sentence.

2 TC-StyleGAN2

In this section, we describe our proposed approach, Text-Conditioned StyleGAN2 (TC-StyleGAN2) that is depicted in Figure 2. StyleGAN2 models are still considered state-of-the-art approaches for training unconditional Generative Adversarial Networks, and their results hold strong even when compared to more recent counterparts such as StyleGAN3. The usual recipe for training such kind of networks involve large amounts of data and huge computing power, specially for training in larger resolutions. Notably, they are only able to generate images in an unconditional fashion, i.e., one cannot ask the model to generate a particular kind of image using neither class information nor natural language queries. TC-StyleGAN2 aims to give StyleGAN networks the ability to generate images from textual descriptions, while leveraging high-quality pre-trained models currently available. We introduce a special kind of finetuning that we call Unconditional-to-Conditional Transfer Learning, which does allow finetuning unconditional models making them conditional.

Our hypothesis is that by reusing pre-trained weights one can accelerate and stabilize the training convergence, and also achieve better results. Such an adaptation is not trivial since the model has to take an additional vector which dictates the natural language condition. That has to be done without causing the collapse of the pre-trained model which could be caused by randomly initialized layers for instance. For that reason, we believe that the

modifications should be minimal and added in the right places with parsimony. In addition, TC-StyleGAN2 makes use of strategies to prevent overfitting and to increase the space smoothness. These strategies are two-fold: (i) employing Conditional Augmentation (CA) on the text embedding, to allow learning a conditional smooth space using a fixed, discrete set of captions; and (ii) an Adaptive Discriminator Augmentation, that can increase image transformations when the model is being able to overfit the data. Following, we discuss each part of the proposed approach.

2.1 Overall architecture

StyleGAN networks are largely inspired by style-influencing techniques. Those techniques, such as AdaIN, allow introducing the input noise vector across multiple stages of the network allowing it to control the content and aesthetics of the synthesized images. StyleGAN model is based on a straightforward GAN framework which contains in two main networks, namely the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. Both of them make use of a mapping network $\mathcal{F}$ which helps to disentangle the representation of the noise space.

Figure 2 shows an overall view of TC-StyleGAN2. Gray boxes indicate parts that we reuse weights pre-trained from standard StyleGAN2 models. White objects denote deep networks: generator, discriminator and text encoder. Yellow boxes represent vectors. Orange ones employ (non)linear projections and transformations. Finally, red shapes are scalars. We made two main modifications to the original model in order to introduce natural language information while still being able to reuse the pre-trained weights: (i) The dimensions of the noise vector $\mathbf{z} \in \mathcal{Z}^{512}$ are split in half, and we concatenate text information in the other half of the vector; and (ii) we enforce the discriminator latent representation vector $\mathbf{d}$ to have high cosine similarity to the sentence embedding $\mathbf{s}$ used as the image synthesis condition. Such score is also used as additional information for the discriminator prediction.

Both modifications are important so the generator can synthesize a plethora of different images, due to the sampling $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, though respecting the condition fixed on the other part of the input vector. By adding the cosine constraint in the discriminator, it becomes able to penalize when generated images are not correspondent to the original sentence.

2.2 Generator

Formally, our synthesis network takes two input vectors: the noise vector $\mathbf{z} \sim \mathcal{Z}^{256}$ and the textual condition vector $\mathbf{s} \in \mathcal{S}^{256}$. Such vectors are concatenated and then mapped to an intermediate representation $w \in \mathcal{W}^{512}$ through 8 layers of a non-linear mapping network $\mathcal{F}$. Given that we adjusted and projected the input vectors into the regular dimensionality it is possible to load trained weights for the entire generator $\mathcal{G}$, including $\mathcal{F}$.

Note that z follows a certain probability density and $\mathcal{S}$ is a pre-trained embedding space (see Section 2.5), but the intermediate latent space $\mathcal{W}$ can learn a more linear, less entangled representation, since it does not have a previously defined distribution. The intermediate representation w through learnable affine transformations generates the style codes $\mathbf{t}_\gamma$ and $\mathbf{t}_\beta$ that are responsible to control adaptive instance normalization AdaIN. The AdaIN operation does perform channel-wise operations of scale and shift based on the style vectors projected from w and is used in the synthesis network $\mathcal{G}$ at each convolutional layer. The remaining of the generator architecture follows closely the original implementation.

2.3 Discriminator

In GANs the discriminator $\mathcal{D}$ network is responsible for detecting if an image is real or artificially generated. It is the responsible for generating gradient signal to the system given that we can assign discrete labels $y \in \{fake, real\}$. Therefore, discriminator goal is to estimate the probability of a given image being real or not, i.e., $\mathcal{D} = P(y_i|\mathbf{v})$. We modify the discriminator so its prediction considers also the condition vector $\mathbf{s}$.

First, the discriminator can take either a real image or a generated image. Such image is processed by several convolutional layers that output a discriminator latent representation $\mathbf{d}^{512}$. In parallel, we input the condition vector $\mathbf{s}$ into a new randomly trained mapping network that operates in $\mathbb{R}^{256}$. We then employ a linear projection layer $\mathcal{E}$ to generate $\hat{\mathbf{s}}$ which is a 512-dimensional vector. We use such linear mappings so we can get the same representation level and disentanglement from the intermediate space of the generator. We then compute a similarity score $\cos(\hat{\mathbf{s}}, \mathbf{d})$ to encourage $\mathcal{D}$ to approximate the condition distribution. The final output from the discriminator is the weighted sum (λ_C conditional weight, and λ_D unconditional weight) of a neuron and the cosine score, so both values have weight while detecting if an image is not only real, but also correspondent to the natural query.

Our two largest modifications are splitting the input vector from the generator, and the introduction of a new randomly initialized mapping network that is parallel to the discriminator main network flow. Such a network mainly adds a constraint to the prediction and does not affect directly all the discriminator layers. It does affect them during backpropagation given that the gradients are estimated from the loss function defined in Equation 1, whose prediction was computed from a combination of the cosine constraint and the neuron prediction. We observe that we can get away with both modifications because they do not impact strongly the main flow of the networks, though provide additional learning signal.

2.4 Loss function

We optimize the discriminator weights θ_D by minimizing the loss function of Equation 1 of the predictions for both real and generated images:

$$\Delta\theta_d \frac{1}{m} \sum_{i=1}^{m} \Big[-\mathcal{A}(-\mathcal{D}(\mathbf{v}_i, \mathbf{s}_i)) - \mathcal{A}(\mathcal{D}(\mathcal{G}(\mathbf{z}_i, \mathbf{s}_i))) \Big] \tag{1}$$

where m is the number of instances in the batch, and $\mathbf{v}_i$ is the i^{th} image drawn from the real data distribution $\mathcal{I}$, and $\mathbf{z}_i$ and $\mathbf{s}_i$ is the noise sampled and the corresponding sentence embedding from $\mathcal{Z}$ for that iteration. Besides, the activation function $\mathcal{A}(x)$ is defined by $softplus(x) = \frac{1}{\beta}\log(1 + exp(\beta * x))$, where β is a hyperparameter. Note that that $\mathcal{D}(\mathbf{v}_i, \mathbf{s}_i) = \cos(d, \hat{\mathbf{s}})\lambda_C + P(y_i|\mathbf{v}_i)\lambda_D$, where λ_C and λ_D are the weights for the conditional and unconditional prediction, respectively.

For optimizing the weights θ_g of the synthesis network, we use the opposite of the loss function for the $\mathcal{D}$ as shown in Equation 2.

$$\Delta\theta_g \frac{1}{m} \sum_{i=1}^{m} \Big[-\mathcal{A}(\mathcal{D}(\mathcal{G}(\mathbf{z}_i, \mathbf{s}_i))) \Big] \tag{2}$$

The overall optimization problem objective is then formulated as the following adversarial training framework:

$$\min_{\mathcal{D}} \max_{\mathcal{G}} \mathbb{E}_{\mathbf{v}\sim\mathcal{I}}[-\mathcal{A}(-\mathcal{D}(\mathbf{v}_i,\mathbf{s}_i))] + \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[-\mathcal{A}(\mathcal{D}(\mathcal{G}(\mathbf{z}_i,\mathbf{s}_i)))] \quad (3)$$

2.5 Text Encoder

Condition representation. A core aspect of our architecture is the design of the text encoder that will extract a vector representation from the text queries. Such encoder should be able to represent details and fine-grained information from text for the used datasets. To achieve this, text descriptions were encoded using the Deep Attentional Multimodal Similarity Model (DAMSM) from the AttnGan encoder module [30]. The idea of DAMSM module draws inspiration from multimodal alignment models [15, 29], where it learns an image-text encoding function, $\phi(I)$ and $\phi(S)$, that projects both paired representations into the same multimodal space. Such functions are trained so that the distances of related pairs is minimized, while unpaired images and texts are far from each other. It does that by training a global representation for images and text, while using a cross-attention mechanism to improve on local and fine-rained detail recognition. Note such encoder defines the condition space. Recently, Ye et al. improved AttnGan and DM-GAN [37], using a contrastive learning approach. In the pre-training stage, this technique was employed in the DAMSM module, learning a more consistent textual representation. Then, this method was utilized in the GANs training, enhancing the consistency between the generated images and their respective captions. Experimental results have shown that the quality of synthesized images has improved significantly, in terms of FID. In this paper, we used the original DAMSM text-encoder module for representing the caption embedding.

Text Augmentation: Considering the condition space, one can see that if the amount of text queries is limited we have a discontinued space. In order to increase the continuity and smoothness of such space we employ the Condition Augmentation (CA) technique [33]. Hence, instead of considering the embedding $\phi(S_i) = \mathbf{s}_i$ of each caption for a given image $\mathbf{v}_i$, we sample a textual embedding $\hat{\mathbf{s}} \sim N(\mu(\mathbf{s}), \Sigma(\mathbf{s}))$ where $\mu(\mathbf{s})$ and $\Sigma(\mathbf{s})$ are the mean and diagonal of the covariance matrix of $\mathbf{s}_i$. Such statistics represent the textual embeddings distribution for a given image. With the aid of CA, the model is going to take far more training pairs, which is particularly important for small datasets, such as CUB and Oxford Flowers.

2.6 Adaptive Discriminator Augmentation

One of the largest challenges in training GANs is the amount of data needed for training some models, such as StyleGAN-size models. In limited datasets, it is easy for the discriminator to overfitting the data. Recent work [14] have proposed a mechanism that stabilizes training in limited data regimes, using an adaptive discriminator augmentation technique, namely ADA. The technique consists of applying 18 types of transformations to the training images with a given probability p. The probability p is adaptively incremented or decremented by a fixed value based on a overfitting level score generated by a heuristic. ADA has proved to be effective to improve transfer of learning in unconditional GANs, leading to better results in terms of FID and IS, in several benchmark datasets [1, 2, 3]. The default incarnation of TC-StyleGAN2 uses ADA, which proved to be important in unconditional-to-conditional transfer learning. It allows us to finetune large models in small datasets while maintaining generalization capabilities. We provide a complete ablation regarding its impact in Section 3.

Methods	FID ↓		IS ↑	
	CUB	Oxford-102	CUB	Oxford 102
StackGAN++ [32]	15.30	48.68	4.04 ± 0.05	3.26 ± 0.01
AttnGAN [30]	23.98	-	4.36 ± 0.03	-
DM-GAN [37]	16.09	-	4.75 ± .07	-
RATGAN [32]	13.91	-	5.36 ± 0.20	**4.09**
ET2I [23] [16]	11.17	16.47	4.23 ± 0.05	3.71 ± 0.06
Lightweight ManiGAN [16]	8.02	-	-	-
LAFITE [36]	10.48	-	5.97	
TC-StyleGAN2 (Ours)	**7.49**	**9.47**	**5.99 ± 0.20**	3.84 ± 0.15

Table 1: Comparison of TC-StyleGAN2 against state-of-the-art models.

Methods	CUB		Oxford-102	
	FID ↓	KID (x 10^3) ↓	FID ↓	KID (x 10^3) ↓
From Scratch	14,04	5,55	34,44	22,87
+ ADA	9,40	4,11	11,37	3,51
+ U2C transfer	8,02	2,25	**9,47**	**2,11**
+ Conditional Augmentation	7,53	**2,07**	10,13	2,87
+ λ_D Tuning	**7,49**	2,14	9,85	2,41

Table 2: FID and KID for various generator designs (lower is better)

2.7 Unconditional-to-Conditional Transfer Learning

One of the main goals of this work is to understand the use of pre-training information from StylenGAN2, trained in an unconditional paradigm in larger datasets. We expect that by taming such unconditional models into conditional ones, we should be able to generate authentic real-looking images coherent with their respective textual descriptions. This would not only accelerate the training convergence process, but also improve the overall quality of the images. For most of our experiments, we use the StyleGAN2 pre-trained weights on FFHQ [1] data. FFHQ is a far more diverse dataset than CUB and Flowers-102. During training, there must be a domain shift between the source and target images. Notably, we have added extra level of complexity to the model so it can take condition vectors. Hence, we reuse all weights from StyleGAN2-FFHQ to initialize $\mathcal{D}$ and $\mathcal{G}$ weights in the new architecture (see modules inside the gray boxes in Figure 2). Some layers had to be randomly initialized, such as the mapping network and linear projection of the condition vector employed in parallel to the discriminator. In Section 3 we show that such modifications benefit text-to-image synthesis by using generator and discriminator pre-trained weights from unconditional models.

3 Experiments

Setup. We employ the standard datasets for evaluating text-to-image synthesis, namely CUB [24] and Oxford-102 [18]. They are used in most of our baselines. In Section 3.1, models are evaluated with the standard GAN evaluation metrics, namely FID [11], KID, [5] and IS [22]. We also showcase qualitatively studies in 3.2. For the U2C transfer we employ weights trained on Flickr-Faces-HQ dataset [1]. It has a diversity of 70,000 high-quality 1024×1024 images.

Baselines. We compare TC-StyleGAN2 to state-of-the-art approaches, such as RATGAN [32], Lightweight ManiGAN [16], ET2I [23] and LAFITE [36], and a baseline that is the same architecture of TC-StyleGAN2 but without Unconditional-to-Conditional Transfer Learning. To evaluate the efficiency of our training and transfer-learning strategies, we trained

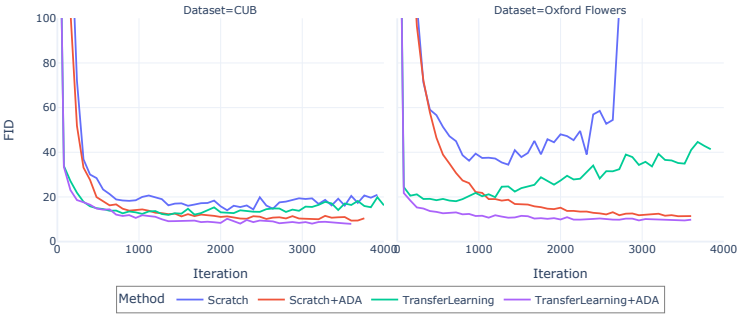

Figure 3: FID for various generator designs (lower is better)

our models for a maximum of 24 hours using a single v100 GPU per run. We used $\{\lambda_D = 0.25, \lambda_D = 1\}$ for CUB, and $\{\lambda_D = 1, \lambda_D = 1\}$ for Oxford-102. Additional hyperparameter tuning information and training details can be found in Supplementary Material.

3.1 Quantitative Analysis

Table 1 shows quantitative results for our main approach, TC-StyleGAN2, as well as results from prior work. Notably, our approach outperformed all past work by large margins. We achieve 7.48 FID versus 8.02 for Lightweight ManiGAN in CUB dataset. The third best result is 10.48 FID from LAFITE. Note that TC-StyleGAN2 results present a relative improvement of roughly 40% when compared to LAFITE. Compared to the previously reported FID results for Oxford-102, once again results of TC-StyleGAN2 are second to none.

For the sake of completeness we also report Inception Scores (IS), which do help to confirm that our approach improves over the prior art for CUB dataset. We highlight that that IS values are not as reliable as FID ones. Results clearly show that IS are not as efficient to measure progress in the field as FID ones. For instance, when we compare StackGAN++ (older model that produces lower quality samples) to our approach the relative improvement of FID for Oxford-102 is 650%; while IS values show only a 17% improvement, and most of the remaining models actually fall in the standard variation range.[1]

Recall that standard TC-StyleGAN2 incarnation is build on top of the StyleGAN2 architecture, and employs unconditional-to-conditional transfer learning, Condition Augmentation, and ADA. Table 2 provides an ablation study that shows the impact of each component in the generator design of TC-StyleGAN2. The models trained from scratch (all layers randomly initialized) had the worst results, and sometimes the training procedure diverged. ADA clearly causes strong reductions of **33.05%** and **66.99%** in the FID scores. It is very clear as well that Unconditional-to-Conditional Transfer Learning does bring important improvements to the results. Textual augmentation improved performance on the CUB dataset, while for Oxford-102 Flowers dataset results decreased marginally.

In Figure 3, training from scratch results in more training instability after a few iterations, and may cause divergence. Hence, each proposed component in TC-StyleGAN2 is quite important and helpful not only for improving generalization but also for accelerate and stabilize

[1] More details in Supplementary Material.

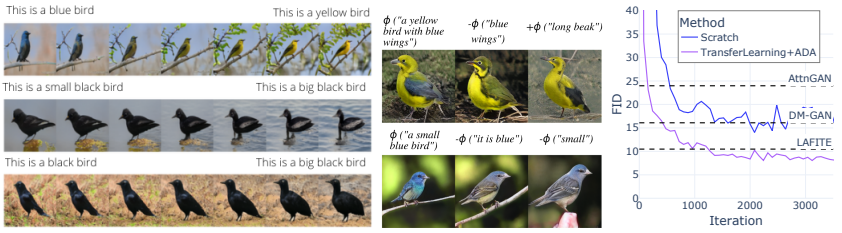

Figure 4: Leftmost illustration showcases images generated by linear interpolation in the intermediate space $\mathcal{W}$ (left to right). Center image depicts arithmetic in the text-embedding space which enables natural language-based image-editing. Rightmost image shows that TC-StyleGAN2 outperforms most of prior work in a few hundred iterations.

the training procedure. Using ADA the FID will take longer to improve but model gets far more robust to overfitting and unstability. Using the complete approach (ADA+Finetuning), the transfer-learning results in state-of-the-art FID values very quickly.

3.2 Qualitative Analysis

Image Editing: similarly to [23], our model is also capable of editing images in textual space, while preserving structural and main features of the birds and also the environment. The middle illustration in Figure 4 shows different examples, with addition and subtraction of characteristics that vary color, sizes, and fine-grained details such as beak size and wing color. The sentence encoded by $\phi($*"this is a yellow bird with a blue wings"*$)$ when subtracted from the embedding $\phi($"blue wings"$)$, produces an image that preserves its structure, but with the semantic characteristic was removed.

Interpolation: we can visualize if the learned $\mathcal{W}$ has structural regularity by interpolating between two vectors in that space. Leftmost part of Figure 4 shows the interpolation between distinct input condition embeddings, but with the same noise. We can observe a gradual merging of features between the generated images as requested by the prompted text query. It clearly shows that our models were in fact learn to respect the condition during Unconditional-to-Conditional Transfer Learning. Even environment details were added smoothly in a semantically meaningful fashion. In the second row, the water background are gradually added to the image to match the bird species living environment. This elucidates how the intermediate latent space has regions that are far less entangled than the latent spaces of noise and embedding. Though, we leave for future work to further explore how to isolate better background modifications when those are not present in the natural queries.

4 Related Work

Some transfer-learning methods in GANs have been used in conditional generation. Wang et al. proposed the initialization the weights of WGAN-GP [10] pre-trained on a diverse dataset and then finetuned this model on small datasets. The pre-trained initialization rather the random one improved the results of the model, achieving better results on fewer iterations. Then, Noguchi and Harada studied a method to reduce the number of weights to be trained by focusing only on learnable batch statistics parameters of the hidden layers of a

pre-trained generator. Wang et al. introduced a transfer method based on extracting knowledge from multiple pre-trained GANs through a trainable miner network. Freezing some layers of the generator or the discriminator has been studied in [17, 35]. Mo et al. proposed to freeze lower layers of the discriminator and only finetune the upper layers. Similarly, Zhao et al. showed that low-level layers of the generator and discriminator trained on large-scale datasets can be transferred to facilitate generation in distinct and small targets domains.

Karras et al. argue that transfer learning often gives better results than from scratch training, but it depends on the diversity of the source dataset, instead of the similarity between the domains. Hence, diverse datasets can be used as source domains to generate more specific ones. The use of such strategies in unconditional GANs showed promising results in limited datasets, often achieving better results than training from scratch. The aforementioned techniques are easily extended to conditional tasks, but not to the text-conditioned ones. The use of such strategies in unconditional GANs showed promising results in limited datasets, often achieving better results than training from scratch. The aforementioned techniques are easily adapted for use in conditional GANs, but not to the text-conditioned ones.

Wang et al. proposed a unified transfer learning method, which can be used for various kinds of image synthesis tasks, like text-to-image, audio-to-image, and image-to-image, using style mixing data triplets computed from pre-trained and unconditional style GANs. Then, the style mixing triplets are used in several image synthesis architectures, like SPADE [20] and StarGanv2 [7], distilling the knowledge from the pre-trained teacher GAN. This technique improved image quality results in different conditional image synthesis tasks. Our transfer-learning approach has the advantage of being simpler while incarnating all the weights of non-conditional StyleGAN2 architecture which notably achieved better results in text-to-image tasks.

5 Conclusion and Future Work

In this work, we proposed a simple, yet very effective transfer-learning approach for training text-conditioned GANs, namely Unconditional-to-Conditional Transfer Learning (U2C transfer). By using such an approach we were able to modify the unconditional architecture of StyleGAN2 to allow text-conditioned image synthesis, which we called Text-Conditioned StyleGAN2 (TC-StyleGAN2). We also added stronger augmentation recipes and strategies, which allowed us to train reasonably large models in very small datasets. Such a method effortlessly outperformed previous state-of-the-art models by large margins in terms of FID in widely used benchmarks. We have shown that pre-training information of an unconditional model trained in a different and more diverse dataset is really helpful when training in smaller datasets. TC-StyleGAN2 took only a few hundred iterations to top most of the prior work. In addition, learning procedure was much more stable when used the proposed strategy. We show that our model is capable of image editing by doing arithmetic operations on the text embedding information and interpolation in the latent intermediate space of the Mapping Network. In future work, we intend to explore an adaptive mechanism of U2C transferthat, in training, adjusts the currently fixed hyperparameters, based on heuristics of how text-conditioned the model is.

References

[1] Github - nvlabs/ffhq-dataset: Flickr-faces-hq dataset (ffhq). https://github.com/NVlabs/ffhq-dataset. (Accessed on 07/02/2022).

[2] stargan-v2/readme.md at master · clovaai/stargan-v2 · github. https://github.com/clovaai/stargan-v2/blob/master/README.md#animal-faces-hq-dataset-afhq. (Accessed on 07/02/2022).

[3] Github - nvlabs/metfaces-dataset. https://github.com/NVlabs/metfaces-dataset. (Accessed on 07/02/2022).

[4] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks, 2017. URL https://arxiv.org/abs/1701.04862.

[5] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2018. URL https://arxiv.org/abs/1809.11096.

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017. URL https://arxiv.org/abs/1704.00028.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017. doi: 10.48550/ARXIV.1706.08500. URL https://arxiv.org/abs/1706.08500.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2019. URL https://arxiv.org/abs/1912.04958.

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020. URL https://arxiv.org/abs/2006.06676.

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[16] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. *Advances in Neural Information Processing Systems*, 33, 2020.

[17] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans, 2020. URL https://arxiv.org/abs/2002.10964.

[18] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

[19] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. *CoRR*, abs/1904.01774, 2019. URL http://arxiv.org/abs/1904.01774.

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[21] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. URL https://arxiv.org/abs/1606.03498.

[23] Douglas M. Souza, Jonatas Wehrmann, and Duncan D. Ruiz. Efficient neural architecture for text-to-image synthesis. *CoRR*, abs/2004.11437, 2020. URL https://arxiv.org/abs/2004.11437.

[24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[25] Kafeng Wang, Xitong Gao, Yiren Zhao, Xingjian Li, Dejing Dou, and Cheng-Zhong Xu. Pay attention to features, transfer learn faster cnns. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxyCeHtPB.

[26] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan C. Raducanu. Transferring gans: generating images from limited data. *CoRR*, abs/1805.01677, 2018. URL http://arxiv.org/abs/1805.01677.

[27] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. *CoRR*, abs/1912.05270, 2019. URL http://arxiv.org/abs/1912.05270.

[28] Yaxing Wang, Joost van de weijer, Lu Yu, and SHANGLING JUI. Distilling GANs with style-mixed triplets for x2i translation with limited data. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=QjOQkpzKbNk.

[29] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12313–12320, 2020.

[30] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017. URL http://arxiv.org/abs/1711.10485.

[31] Hui Ye, Xiulong Yang, Martin Takác, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *CoRR*, abs/2107.02423, 2021. URL https://arxiv.org/abs/2107.02423.

[32] Senmao Ye, Fei Liu, and Minkui Tan. Recurrent affine transformation for text-to-image synthesis, 2022. URL https://arxiv.org/abs/2204.10482.

[33] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2016. URL https://arxiv.org/abs/1612.03242.

[34] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017. URL http://arxiv.org/abs/1710.10916.

[35] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained GANs for generation with limited data. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11340–11351. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhao20a.html.

[36] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.

[37] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. *CoRR*, abs/1904.01310, 2019. URL http://arxiv.org/abs/1904.01310.