# 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

## 1 Supplementary Materials

In this supplementary document, we report the architectural details and additional experiments that we conducted.

### 1.1 Implementation Details

All models were optimized with Adam ($\beta_1 = 0.0$, $\beta_2 = 0.99$) and learning rate of $2 \times 10^{-3}$ for both discriminator and generator networks. We chose StyleGAN2's R1 regularization weight equal 0.8192, as suggested by guidelines of StyleGAN2 implementation [6].

### 1.2 Datasets

We evaluated our study on two popular datasets and baselines:

- **Caltech-UCSD Birds (CUB)**[10]: The dataset has 200 different categories with 11,788 images of birds in total. Each image contains 10 text descriptions of bird characteristics. CUB is split in 8,855 images of 150 categories for training and 2,933 images of 50 categories for testing.
- **Oxford-102**[5]: The dataset has 102 different categories with 8,189 images of flowers in total. Each image contains 10 text descriptions of flowers characteristics. Oxford-102 is split in 7,034 images for training and 1,154 images for testing.

### 1.3 Evaluation

This study was evaluated by using three metrics highly used in generative image models: the Inception Score (IS), the Fréchet Inception Distance (FID) [3] and the Kernel Inception Distance (KID) [2]. IS measures both objectivity and variety, so the higher the score, the better. To calculate class probabilities over generated samples, a pre-trained InceptionV3

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

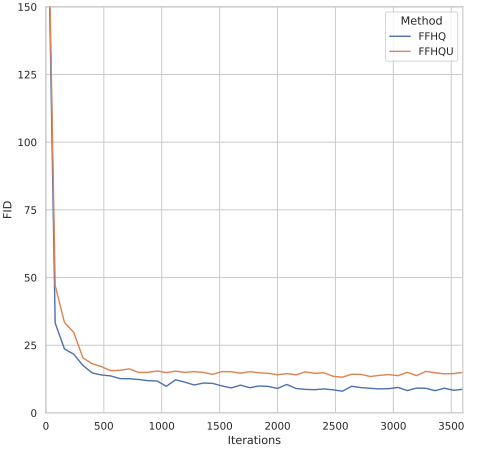


Figure 1: Training comparative of our method (TC-StyleGAN2), in terms of FID (lower is better), using different source-domains of transfer learning.

Network [8] is used. IS does not take into account the statistics available in the real image data, so we have to consider other metrics to evaluate the model.

Using statistics from the images in the training data, FID evaluates how far the statistics of the fake images are from the real ones, generated by captions from the test dataset. FID uses activation features from the InceptionV3 to extract information from images. Low FID values represent good similarity between these distributions, so the lower the score, the better. Karras et al. argue that FID is not an ideal metric for limited datasets, and Bińkowski et al.. demonstrate that FID is a biased estimator.

Recently, KID has been applied as a more suitable metric to measure the similarity between artificial and real image distribution. KID applied the Maximum Mean Discrepancy (MMD) between the InceptionV3 representation of fake and real images, through a kernel function. As with the FID score, the lower the KID value, the better.

In this paper, FID and KID were evaluated considering all the test captions and statistics from training data, as StackGAN++ study [11].

## 1.4 Additional Results

We also considered using the weights from the Stylegan2-FFHQ-U unconditioned network. FFHQ-U [1] is a modified and less aligned dataset than the FFHQ. Figure 1 demonstrates the training progress and the FID superiority of using FFHQ as a source domain.

Figure 2 depicts the training progress using the original and unconditioned StyleGAN2. It is shown a comparative, in terms of FID, between the training from scratch approach, what we considered as a baseline, and using pre-trained information from FFHQ Faces dataset. The leftmost figure the results using the flowers dataset, while the rightmost figure represents the birds dataset. The transfer-learning approach in the unconditioned case also reached a reasonable value of FID in the firsts iterations and stabilize faster than the baseline approach.

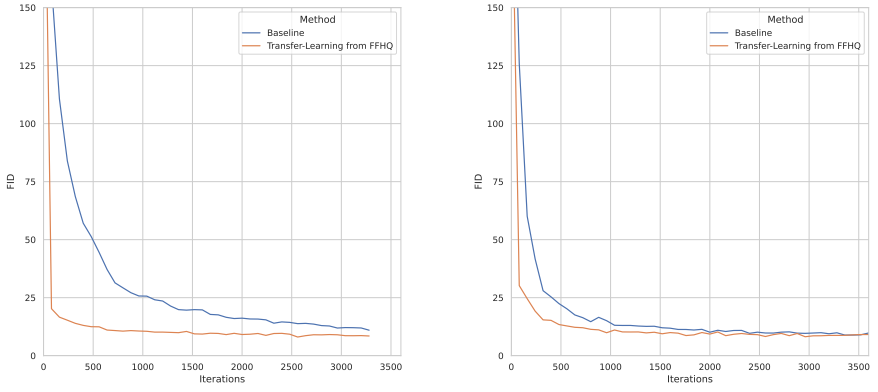


Figure 2: Training progress comparative of the original and unconditioned StyleGAN2, in terms of FID. Comparison between training from scratch (baseline) and using pre-trained weights from FFHQ. The leftmost figure shows the Oxford-102 results, and the rightmost shows the CUB-200 results.

Table 1 shows quantitative results for different generator designs. The models were trained using different configurations: from scratch (all layers randomly initialized), using ADA, and U2C transfer method. We do not considered the text augmentation mechanism. For all designs, we chose $\lambda_C = 1$ and $\lambda_D = 1$.

| | From Scratch | | | | Ours (U2C transfer) | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Using ADA | | Non-Using ADA | | Using ADA | | Non-Using ADA | |
| | FID ↓ | KID (x $10^3$) ↓ | FID ↓ | KID (x $10^3$) ↓ | FID ↓ | KID (x $10^3$) ↓ | FID ↓ | KID (x $10^3$) ↓ |
| CUB | 9,40 | 4,11 | 14,04 | 5,55 | **8,02** | **2,25** | 11,96 | 5,60 |
| Oxford-102 | 11,37 | 3,51 | 34,44 | 22,87 | **9,47** | **2,11** | 18,09 | 9,25 |

Table 1: FID and KID scores considering different training configurations (lower is better)

Table 2 shows quantitative results considering the default incarnation of TC-StyleGAN2 , which uses ADA. We considered the text-augmentation mechanism as well, with conditional parameter $\lambda_C$ varying between 0 and 1, with fixed $\lambda_D = 1$.

| Dataset | $\lambda_D$ | $\lambda_C$ | FID | KID |
|---|---|---|---|---|
| CUB | 0,125 | 1 | 7,61 | 2,20 |
| | 0,25 | 1 | **7,49** | 2,14 |
| | 0.75 | 1 | 7,83 | **1,97** |
| | 1 | 1 | 7,53 | 2,07 |
| Oxford-102 | 0,125 | 1 | 10,00 | 2,43 |
| | 0,25 | 1 | 9,95 | **2,39** |
| | 0.5 | 1 | **9,85** | 2,41 |
| | 0.75 | 1 | 10,36 | 2,64 |
| | 1 | 1 | 10,13 | 2,87 |

Table 2: FID and KID scores (lower is better) for $\lambda$ tunning

Figure 3 depicts the visualization of the sentence representation, showing some effects of the conditional augmentation module, and some properties of the mapping intermediate space $\mathcal{W}$. In b) and d), using conditional augmentation, it is possible to generate much more

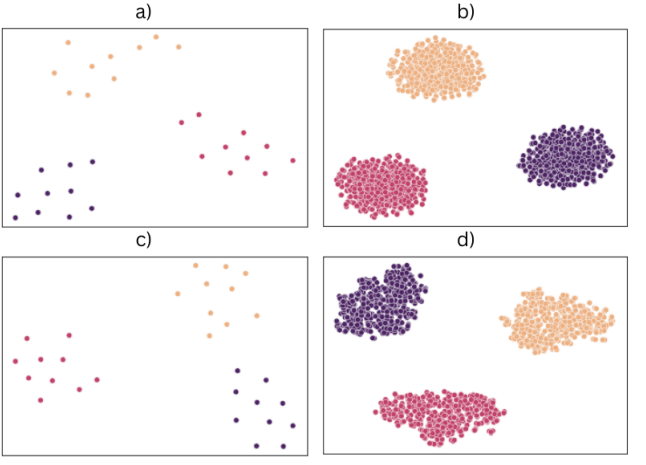


Figure 3: Visualization of sentences embbeddings. We sampled 3 random images, and applyed t-SNE [9] to reduce the original space to $\mathcal{R}^2$. In **a)** is shown 30 sentences embeddings, as each image has 10 captions, projecting $\mathcal{R}^{256}$ to $\mathcal{R}^2$. In **b)** is shown 500 sentences embeddings for each image, using the conditional-augmentation module. In **c)** is shown the intermediate representation in $\mathcal{W}$ space, using the 30 captions, projecting $\mathcal{R}^{512}$ to $\mathcal{R}^2$. In **d)** is shown the intermediate representation in $\mathcal{W}$ space of 500 sentences for each image, using the conditional-augmentation module, projecting $\mathcal{R}^{512}$ to $\mathcal{R}^2$.

sentence embeddings than each image has, allowing a continuous sampling of the textual representation. In c) and d), a learned representation of the sentence embedding is shown, in the $\mathcal{W}$ space. As each group of images is clearly separated, the $\mathcal{W}$ space carries a meaningful and learned representation of the sentence.

## 1.5 More qualitative Results

Figures 4 and 5 depict images produced by our method, with their respective textual descriptions, compared to other architectures. Our generated images are highly photo-realistic, with fine-grained details, presenting a coherent semantic correspondence with the captions. Our method produces more realistic backgrounds between the architectures and is more coherent concerning the input query. For example, our method generates all colors in different parts of the petals considering the whole query *"The petals of the flowers are in various colors such as red, green, and purple"* but the other models synthesized images considering only part of this input.

### 1.5.1 Training Progress

Figures 6 and 7 represent images generated by different designs at the beginning of the training. Each image is generated by the model's checkpoint after seeing 0 and every 80k images. Using U2C transfer method, the model is capable of generating birds and flowers just at the beginning of the training, while the other model's approach is still generating blurring things yet. However, it is noticeable that images are being formed conditioned-to-text, even in the blurred ones.

Figure 4: Images synthesized by StackGan++ [11], HDGAN [12], ET2I [7] and our method.

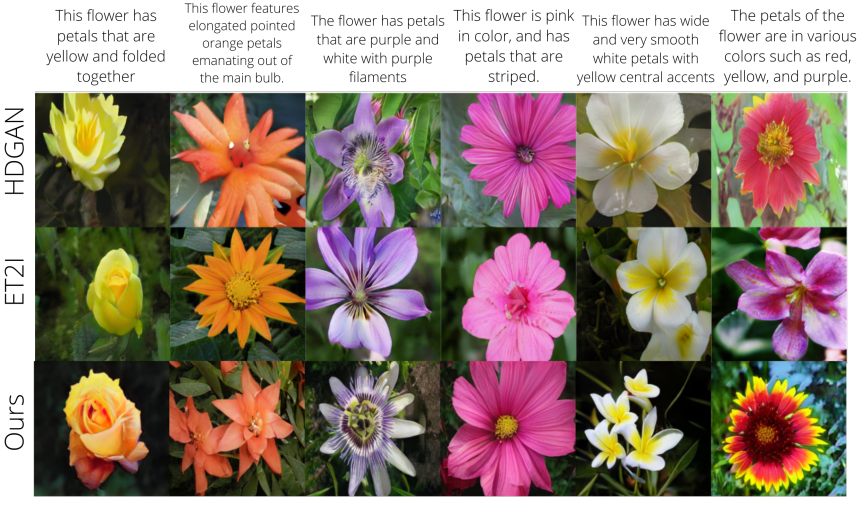


Figure 5: Images synthesized by HDGAN [12], ET2I [7] and our method.

Figure 6: Fake Images generated by the query "This flower is pink and white in color, with petals that are connected" every 80k images seen in the training.

Figure 7: Fake Images generated by the query "This is a red bird." every 80k images seen in the training.

### 1.5.2 Sentence Interpolation and Image Editing

The leftmost illustration of Figure 8 depicts the interpolation in the latent space of the encodings in 4 different directions: $\phi$("this is a brown bird") (upper left), $\phi$( "this is a red bird") (upper right"), $\phi$("this is a black bird") (bottom left) and $\phi$("this is a yellow bird) (bottom right). The intermediate images have mixed characteristics of the interpolation direction, while, in general, the environment and the color characteristic of the bird are preserved. Similarly, the rightmost illustration of Figure 8 shows a image interpolation, in Oxford-102 data, where the corners are images representing "A pink flower" (upper left), "A yellow flower"(upper right), "An orange flower" (bottom left) and "A violet flower" (bottom right).

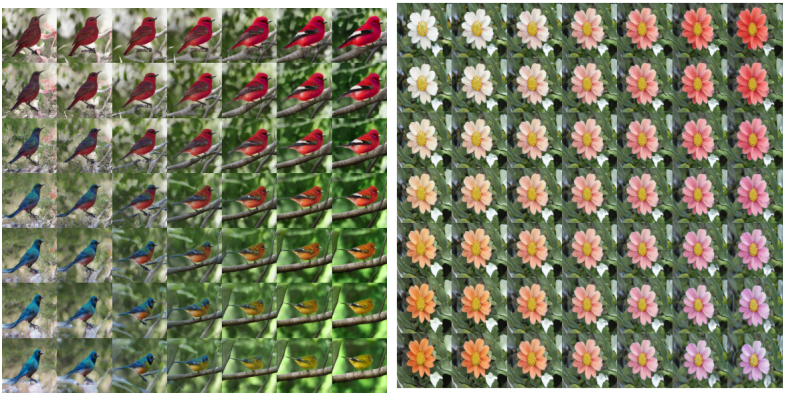

Figure 8: Image interpolation in four directions of the text-embedding space of DAMSM text-encoder

Figure 9 shows different cases of arithmetic in textual space, with addition and subtraction of characteristics that vary in color, sizes, and fine-grained details such as beak size and wing color. Figure 10 shows the interpolation of intermediate mapping space $\mathcal{W}$. 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 learned to respect the condition during Unconditional-to-Conditional Transfer Learning. Even environment details were added smoothly in a semantically meaningful fashion.

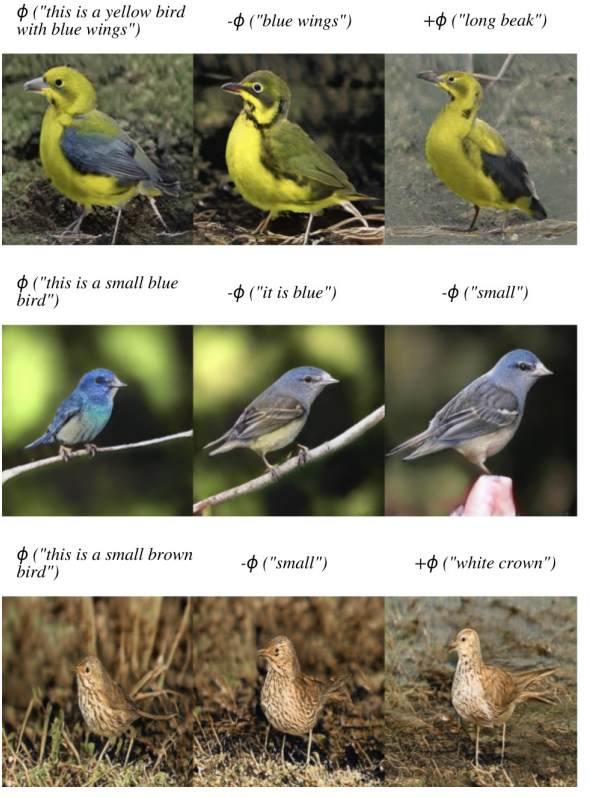


Figure 9: Arithmetic in the text-embedding space of DAMSM text-encoder, enabling image-editing.

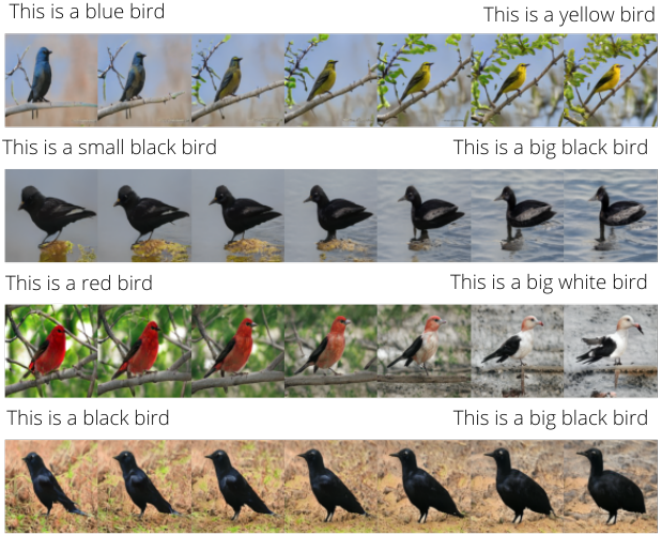


Figure 10: Images generated by linear interpolation in the intermediate space $\mathcal{W}$ (left to right)

# References


[1] Stylegan3 pretrained models: Nvidia ngc. URL https://catalog.ngc.nvidia.com/orgs/nvidia/teams/research/models/stylegan3.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] NVlabs. Stylegan3/configs.md at main · nvlabs/stylegan3, Oct 2021. URL https://github.com/NVlabs/stylegan3/blob/main/docs/configs.md.

[7] 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.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

[9] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

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

[11] 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.

[12] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018.