clip2latent: Supplementary Material

Justin N. M. Pinkney
justin@lambdal.com
Chuan Li
c@lambdal.com
Lambda, Inc
San Francisco, USA

Appendices

A Hyperparameters

A.1 Noise level

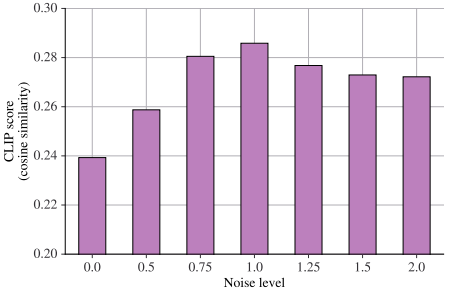

Figure 1: CLIP similarity score measured at different noise scaling factors (α) applied to the CLIP image embeddings on which the model is conditioned.

Following LAFITE[47] we apply scaled Gaussian noise to generate an augmented CLIP embeddings, e_i', as follows:

$$e_i' = \frac{y}{||y||_2}, \quad y = e_i + \alpha \frac{\varepsilon}{||\varepsilon||_2} \tag{1}$$

where e_i represents the normalised CLIP image embedding, $\varepsilon \sim \mathcal{N}(0, I)$ is a Gaussian noise sample of the same dimension as e_i, and α is the noise scaling factor.

To ascertain the optimal noise level for CLIP image embedding augmentations we perform a simple search of possible noise values. For each noise level we run training of clip2latent model using the StyleGAN2 FFHQ model for 100 million iterations. We then select the best model by generating images using a set of 16 text prompts and measuring the CLIP similarity score. We find that the optimal value for α is 1.0, see Figure 1

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

A.2 Model hyper parameters

Model and training hyper-parameters used for training both the StyleGAN2 FFHQ and StyleGAN3 LHQ clip2latent models. Parameter names follow the keys in the training configurations available in the code repository at: https://github.com/justinpinkney/clip2latent. In total our model has 48.9 million parameters.

Model parameters

Parameter	Value
timesteps	1000
beta_schedule	cosine
predict_x_start	True
cond_drop_prob	0.2
dim	512
depth	12
dim_head	64
heads	12

Training parameters

Parameter	Value
iterations	1,000,000
batch_size	512
lr	1.0e-4
weight_decay	1.0e-2
ema_beta	0.9999
ema_update_every	10
noise_scale (α)	1.0
Adam β_1, β_2	0.9, 0.999

B CLIP score prompts

Below is the list of prompts used for CLIP similarity scoring with clip2latent and comparable methods, all captiosn were prefixed by "A photograph of". An asterisk indicates the prompt could not be used with DALLE-2 due to the content policy applied by OpenAI.

1. a person with glasses
2. a person with brown hair
3. a person with curly blonde hair
4. a person with a hat
5. a person with bushy eyebrows and a small mouth
6. a person smiling
7. a person who is angry
8. a person looking up at the sky
9. a person with their eyes closed
10. a person talking
11. a man with a beard
12. a happy man with a moustache
13. a young man
14. an old man
15. a middle aged man
16. a youthful man with a bored expression
17. a woman with a hat
18. a happy woman with glasses
19. a young woman
20. an old woman
21. a middle aged woman
22. a baby crying in a red bouncer
23. a child with blue eyes and straight brown hair in the sunshine
24. an old woman with large sunglasses and ear rings
25. a young man with a bald head who is wearing necklace in the city at night
26. a youthful woman with a bored expression
27. President Xi Jinping *
28. Prime Minister Boris Johnson *
29. President Joe Biden *
30. President Barack Obama *
31. Chancellor Angela Merkel *
32. President Emmanuel Macron *
33. Prime Minister Shinzo Abe *
34. Robert De Niro
35. Danny Devito
36. Denzel Washington
37. Meryl Streep *
38. Cate Blanchett *
39. Morgan Freeman *
40. Whoopi Goldberg *
41. Usain Bolt *
42. Muhammad Ali *
43. Serena Williams *
44. Roger Federer *
45. Martina Navratilova *
46. Jessica Ennis-Hill *
47. Cathy Freeman *
48. Christiano Ronaldo *
49. Elsa from Frozen
50. Eric Cartman from South Park
51. Chihiro from Spirited Away
52. Bart from the Simpsons
53. Woody from Toy Story
54. a university graduate
55. a firefighter
56. a police officer
57. a butcher
58. a scientist
59. a gardener
60. a hairdresser
61. a man visiting the beach
62. a woman giving a TED talk
63. a child playing with friends
64. a person watching birds in the forest

C Comparison methods

C.1 Direct latent optimisation

We employ code by nshepperd[24] for optimisation of StyleGAN latent vectors based on CLIP similarity score to a text prompt. We modify the existing code such that optimisation is performed in w space (rather than the original $w+$ space) of StyleGAN to more closely match our approach. We otherwise leave optimisation parameters as originally shared in the notebook.

C.2 LAFITE text-free model

We use the text-free MM-CelebA trained model provided by the original authors of LAFITE. As the inference of LAFITE is fast, to most closely compare LAFITE with our method we employ the same CLIP re-ranking approach as clip2latent, and sample 16 generations for each prompt and select that with the best CLIP similiary score.

C.3 DALLE-2

We compare to DALLE-2 using the web interface provided by OpenAI. The web interface was returns 4 images per prompt, to avoid bias the first sample was taken for every prompt. To try and generate images more comparable with the style of other methods trial and error was used to generate an engineered prompt of the form: "A photo headshot, single whole head centred, of $\langle caption \rangle$, flickr dslr, portrait photograph"

D Latent Editability

Our methods generate latent vectors which can effectively be edit using existing editing directions in StyleGAN's latent space. In contrast edits of the same magnitude applied to latents produced by direct optimisation frequently produce artefacts such as changes in composition, poor image quality, and in some cases complete failure to generate a recognisable face, see Figure 2.

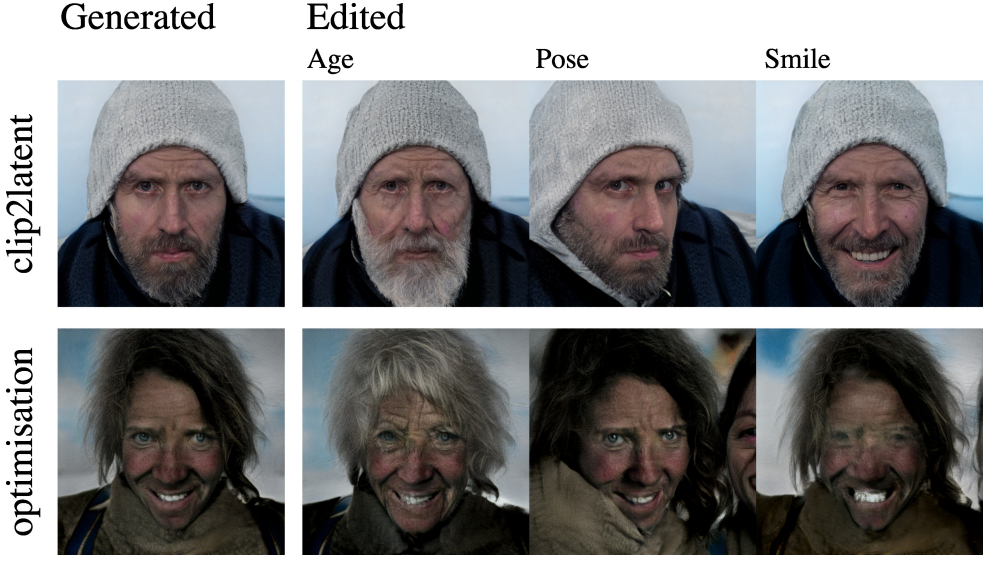

Figure 2: Comparison of editability of latents generated from the text prompt "A photograph of an arctic explorer" generated using clip2latent vs direct optimisation. Edit directions were applied with the same magnitude to both generated latents.

E Other StyleGAN latent spaces

We also explored the training of clip2latent model using the extended latent spaces of StyleGAN. The $w+$ space has been shown to be more expressive then w and is frequently used to accurately embed arbitrary images into the StyleGAN latent space[1]. However generating samples from $w+$ by randomly combining w samples in from all 18 vectors gives rise to low quality generated images[**?**] making it unsuitable for our application.

We briefly explore the use of a more limited latent space comprising three independent w vectors which correspond to low, medium and high resolution layers. We term this latent space $w3$ and find that randomly generated samples are still high-quality but more diverse. We explored training a clip2latent model in this space using the same hyper-parameters as in w space. We found that although the model trains successfully, the overall CLIP similarity scores are not improved and the model converges more slowly. We leave further exploration of how to exploit the greater diversity of extended latent spaces to future work.

F Finding edit directions

ALthough not our main application we also note that our latent generation method can be used to find directions within StyleGAN's latent space. By generating sets of latent vectors corresponding to a "positive" and "negative" prompt we can measure the direction between the two sets in order to obtain a latent direction, see Figure 3.

Figure 3: Examples of editing direction found using text based latent vector generation using clip2latent. Left column: random samples from StyleGAN. Following columns: edits in the directions: smile, curly hair, age, glasses, make up, beard.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? April 2019.

[2] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. CLIP2StyleGAN: Unsupervised extraction of StyleGAN edit directions. December 2021.

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. ReStyle: A Residual-Based StyleGAN encoder via iterative refinement. April 2021.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. September 2018.

[5] Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. November 2020.

[6] Katherine Crowson. Clip guided diffusion hq 256x256., 2021. URL https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj.

[7] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and better Text-to-Image generation via hierarchical transformers. April 2022.

[8] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. June 2021.

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-Guided domain adaptation of image generators. August 2021.

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for Text-to-Image synthesis. November 2021.

[11] Jonathan Ho and Tim Salimans. Classifier-Free diffusion guidance. September 2021.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv [cs.LG]*, June 2020.

[13] Travis Hoppe. Tweet, 2021. URL https://twitter.com/metasemantic/status/1356406256802607112.

[14] Minguk Kang, Joonghyuk Shin, and Jaesik Park. StudioGAN: A taxonomy and benchmark of GANs for image synthesis. June 2022.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. December 2019.

[16] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. June 2021.

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv [cs.CV]*, June 2022.

[18] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. StyleMC: Multi-Channel based fast Text-Guided image generation and manipulation. December 2021.

[19] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. March 2022.

[20] Ryan Murdock. Tweet, 2021. URL https://twitter.com/advadnoun/status/1348375026697834496.

[21] Ryan Murdock. Tweet, 2021. URL https://twitter.com/advadnoun/status/1351038053033406468.

[22] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. February 2021.

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with Text-Guided diffusion models. December 2021.

[24] nshepperd. stylegan3 with clip guidance, 2021. URL https://colab.research.google.com/drive/1eYlenR1GHPZXt-YuvXabzO9wfh9CWY36.

[25] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11), November 2017.

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven manipulation of StyleGAN imagery. March 2021.

[27] Justin N. M. Pinkney. Awesome Pretrained StyleGAN3, .

[28] Justin N. M. Pinkney. Awesome pretrained StyleGAN2, .

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. February 2021.

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image generation. February 2021.

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional image generation with CLIP latents. April 2022.

[32] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for Image-to-Image translation. August 2020.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution image synthesis with latent diffusion models. December 2021.

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv [cs.CV]*, May 2022.

[35] Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. February 2022.

[36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-Filtered 400 million Image-Text pairs. November 2021.

[37] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. May 2020.

[38] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. April 2021.

[39] Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. March 2015.

[40] Yang Song and Stefano Ermon. Improved techniques for training Score-Based generative models. June 2020.

[41] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. February 2021.

[42] Christos Tzelepis, James Oldfield, Georgios Tzimiropoulos, and Ioannis Patras. ContraCLIP: Interpretable GAN generation driven by pairs of contrasting sentences. June 2022.

[43] Phil Wang. Dall-e 2 - pytorch, 2022. URL https://github.com/lucidrains/DALLE2-pytorch.

[44] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for Image-to-Image translation. May 2022.

[45] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. CLIP-GEN: Language-Free training of a Text-to-Image generator with CLIP. March 2022.

[46] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled Text-Driven image manipulation empowered by Pre-Trained Vision-Language model. November 2021.

[47] 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. November 2021.

[48] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. October 2021.