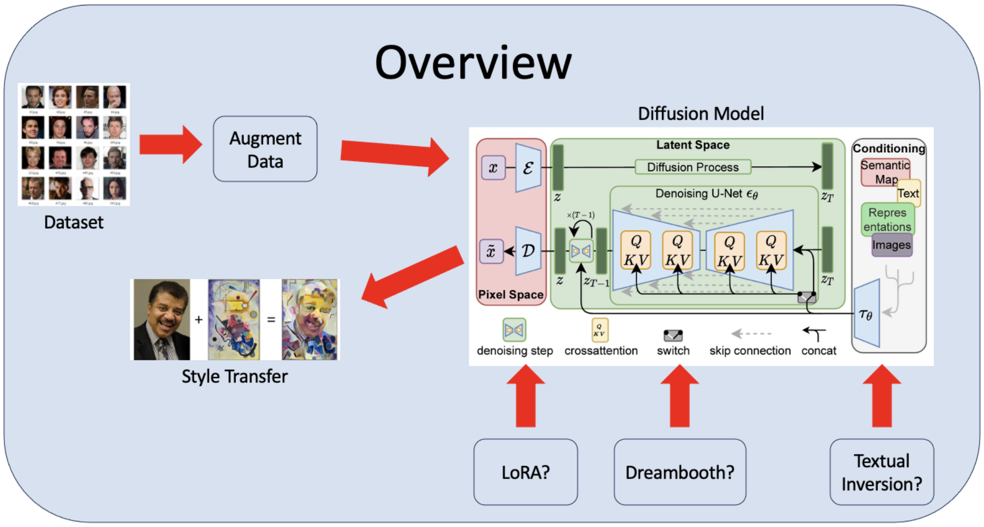
I would like to have an artsy profile picture, but I've never had an artist friend who would draw me. I think this is the perfect time to try out some models to see the best approach for such a task. In class, we've learned about guiding generation using a conditioned latent code. I will experiment with each until I find a model that produces a satisfactory result.

The goal of this project is to produce a pipeline to generate a stylized portrait of a subject. Ideally, the subject can insert themselves into a new scene with a semantically significant textual prompt, then stylize the image as they see fit. In addition to the baseline, I will experiment with various dataset augmentations, training techniques, and different paper's approaches.



Textual inversion [1] optimizes the text embedding to maximize the semantics of an embedding to match an example training image. The method freezes the generative part of the model, which makes the training efficient and less data-hungry. However, textual inversion comes at the cost of producing some artifacts. I can try reducing this problem with some augmented images, but it is unknown if this will be enough. To augment this model, I may try to implement the extended version [5], which does not have an official codebase.

Dreambooth [2] is a holistic fine-tuning strategy that is almost guaranteed to produce a strong result, but will be costly to train. It would also be harder to experiment with multiple models since I'd have to retrain them all. As a result, Dreambooth should be considered a last resort approach.

A middle ground approach would be to experiment with LoRA [3] and SVDiff [4] as a training strategy. We would use significantly less memory, but the training would still be expensive and not particularly reusable since I'm using a limited dataset.

The dataset will range from small to medium sized, as I have limited resources and time. Since I would like to generate a portrait, my dataset will be composed of images of myself, with perhaps some augmentations that I create myself.

The current plan will be to first produce a baseline. I will produce a dataset of myself that I can then use for training. I will start by implementing textual inversion and extended textual inversion to see if I can cohesively add myself to the output. If this fails to return results, I will try LoRA/SVDiff, then Dreambooth, in that order. These two approaches appear orthogonal to textual inversion, so it should be okay to implement both at the same time. After this, I will work on styling. In particular, I hope to leverage the diffusion model's ability to produce a stylistic portrait through textual prompting alone. However, if that is not possible, I can instead produce a pipeline and utilize style transfer to stylize the image itself with a reference image [6].

Although the base goal has been done before, I believe that there are very few implementations that thoroughly investigate different approaches to this problem. In particular, I'm interested in seeing the effects of the extended textual inversion as well as seeing how combining these different approaches may affect the results. Last of all, I believe that style transfer in particular, is not usually a consideration in the pipeline. While a singular end-to-end model is less complex, it may come at the cost of data limitation and reduced ability. This is why a neural style transfer approach may become necessary.

References

[1] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618.

[2] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22500-22510).

[3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[4] Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., & Yang, F. (2023). Svdiff: Compact parameter space for diffusion fine-tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7323-7334).

[5] Voynov, A., Chu, Q., Cohen-Or, D., & Aberman, K. (2023). $ P+ $: Extended Textual Conditioning in Text-to-Image Generation. arXiv preprint arXiv:2303.09522.

[6] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).