Deep Learning School Final Project

Stepik ID: 385163654

Student: Viktor Sokolov

Куратор: Нина Коновалова

Image editing via CLIP guidance

Данная работа состояла из 3 этапов:

1. Выбор предобученных моделей: CLIP, ArcFace, StyleGAN

2. Реализация оптимизационного подхода для редактирования изображений в z и w+ пространствах.

3. Реализация тренировочного цикла для Latent Mapper

1. Выбор моделей.

В качестве StyleGAN была выбрана данная PyTorch реализация. Её преимущества в возможности напрямую получать style vectors из w+ пространства размерности (18*512), также имеется встроенный инвёртер StyleGAN, который, как позднее оказалось, не слишком хорошо работает на изображениях, которые не из датасета FFHQ (на котором StyleGAN и обучался).

В качестве ArcFace использовалась следующая модель. ArcFace необходим для построения Identity Loss, которые оценивает степень похожести картинок лиц.

Для возможности редактировать картинки текстовыми запросами применяется модель CLIP от OpenAI.

2. Оптимизационный подход редактирования

Полностью с нуля реализован оптимизационный подход, из оригинальной имплементации была только позаимствована функция подбора скорости обучения для Adam Optimizer. Редактирование произвольного изображения не было реализовано, т.к. не удалось адекватно инвертировать StyleGAN. Поэтому в качестве тестовых изображений используются картинки сгенерированные самим же StyleGAN. Произведено сравнение оптимизации в исходном z-пространстве размерности 512, и в пространстве style vectors размерности 18*512.

Ниже представлены несколько результатов редактирования в обоих пространствах, в большинстве случаев оптимальная величина параметров l2 = 0.001, lambda_ID = 0.008. Количество итераций в диапазоне 1500-3000. Над оригинальной картинкой сверху написан промт и значения параметров.

dark skin l2 = 0.0 ID = 0.008

w+ Space

z - Space

freckles l2=0.001 ID = 0.05

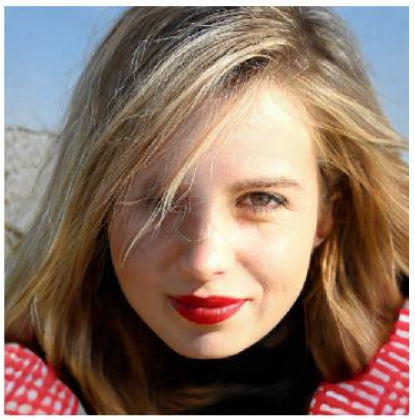
w+ Space

z - Space

red lipstick l2=0.001 ID = 0.008

w+ Space

z - Space



heavy makeup l2=0.001 ID = 0.05

w+ Space

z - Space

thick moustache l2=0.001 ID = 0.05

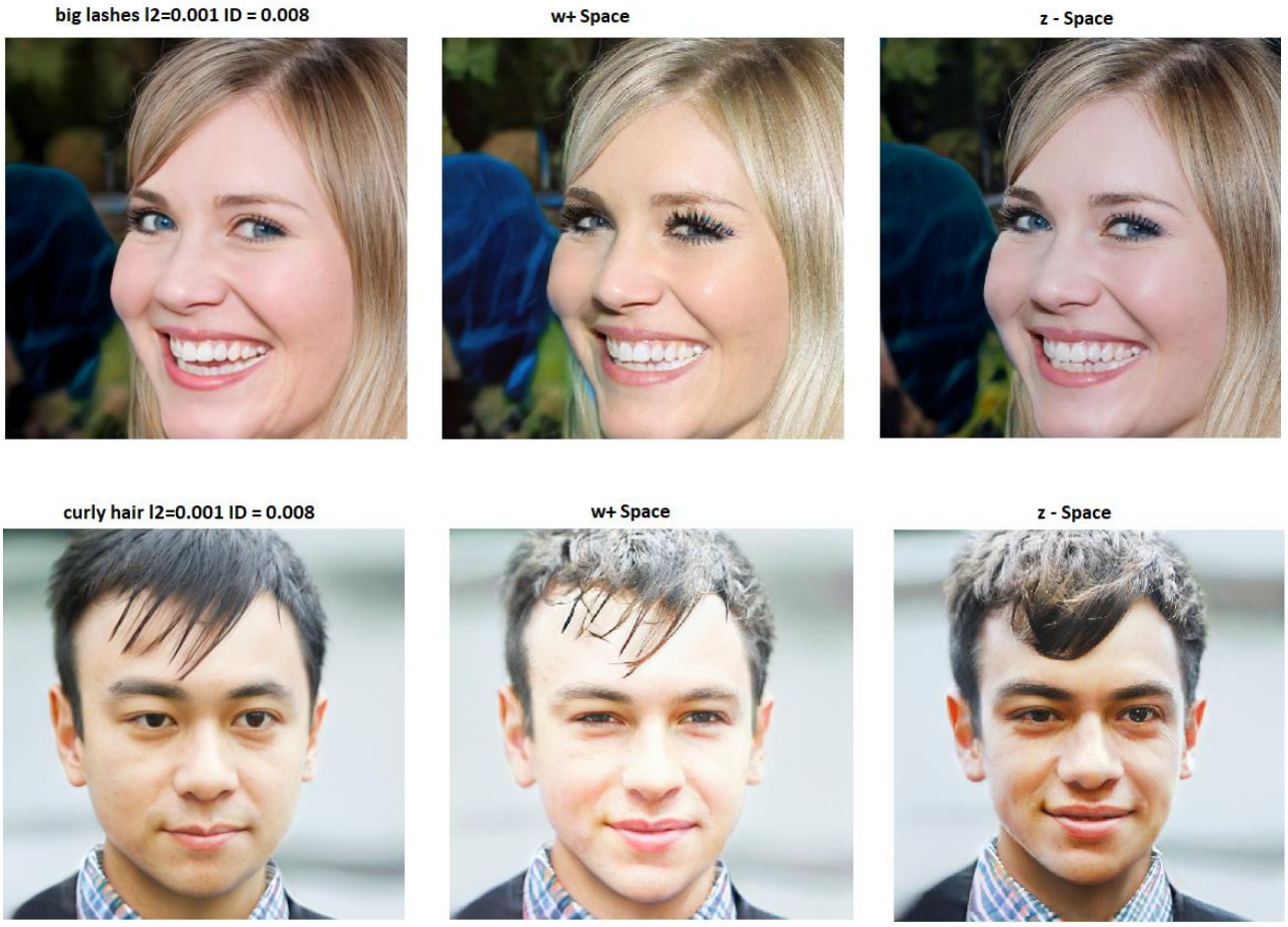
w+ Space

z - Space

red hair l2=0.001 ID=0.05

w+ Space

z - Space



Общие наблюдения: в целом оптимизация в w+ пространстве сходится быстрее и позволяет добиться больше модификаций, например, цвет волос на красный в z-пространстве не сработал. Также чувствуется, что некоторый тип картинок преобладал в FFHQ, на котором обучался StyleGAN. Например, добавление макияжа, губной помады, ресниц и т.п. сходилось очень быстро (порядка 200-400 итераций), тогда как другие модификации занимали более 1000 итераций.

Возможные модификации оптимизационного подхода: известно что пространство style vectors условно можно разделить на coarse, middle и fine подпространства. Первое меняет форму лица, прическу (“большие” черты), второе меняет более мелкие черты, и последнее отвечает за цвет и текстуру. Можно попробовать оптимизироваться только в определённом подпространстве, в зависимости от характера запроса.

3. Тренировочный цикл Latent Mapper

Ввиду отсутствия вычислительных ресурсов был написан только цикл обучения и подготовлены данные, на ноутбук можно посмотреть здесь.

Был подготовлен тренировочный датасет из 20’000 картинок и валидационный из 500, сгенерированный самим StyleGAN, соответствующие style vectors были сохранены.

Изначально была идея сделать “conditional” Latent Mapper, который бы менял цвет волос людей, например, на 8 разных цветов (т.е. принимал 8 разных промтов) и который бы

менял только fine часть style vectors, которая как раз отвечает за цвет. Первая попытка обучить была предпринята только для 1 промта: green hair, но ввиду недоступности вычислительных ресурсов пройдя одну эпоху на 20000 дальше обучаться не было возможности, и также не было видно даже намёка на успешное редактирование. В оригинальной статье указано 10-12 часов обучения на Latent Mapper, и учитывая необходимость многих экспериментов для подбора удачной архитектуры LatentMapper не было возможности довести эту часть проекта до конца.