Body-Lightning ID Diffusion

Даниил Иванович Казачков
Научный руководитель: А. В. Филатов

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных

2025

Обзор области

VAE $\longrightarrow$ GAN $\longrightarrow$ VLM $\longrightarrow$ Diffusion model

Диаграмма не означает, что вышедшие ранее модели устарели! Выбор инструмента зависит от задачи.

Существующие работы

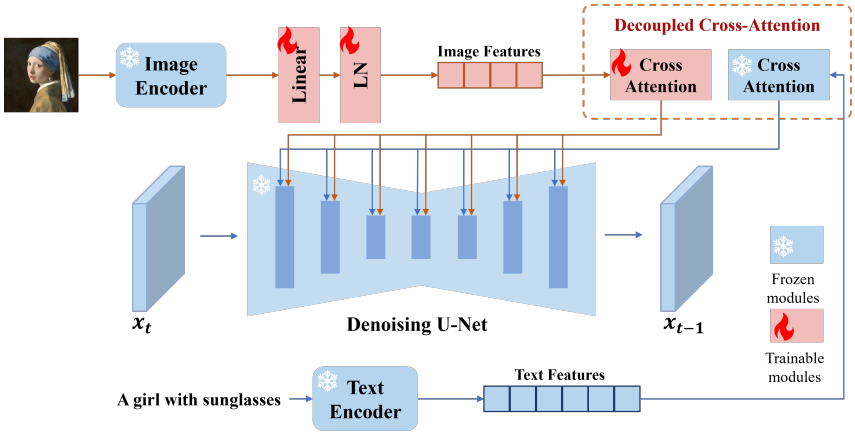


Рис. 1: Архитектура модели IP-Adapter. Автор разделелили cross-attention, позволив диффузионной модели "смотреть"и на текст, и на картинку

Существующие работы

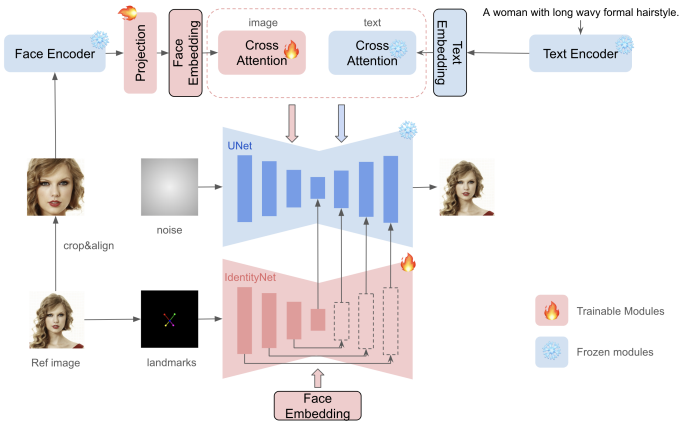


Рис. 2: Архитектура модели InstantID. Использовали идеи ControlNet, отказались от выделения идентичности через CLIP

Существующие работы

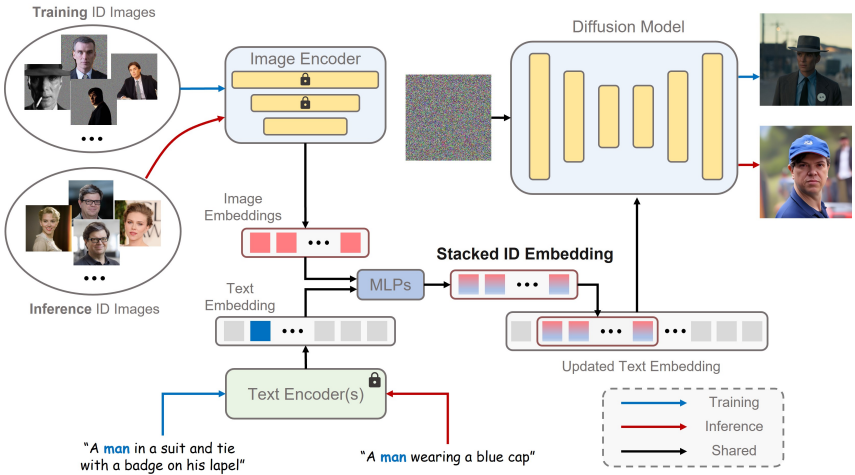


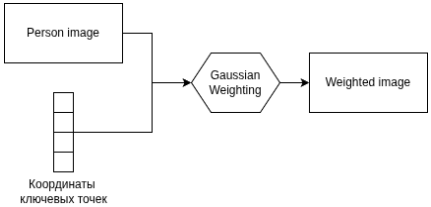
Рис. 3: Архитектура модели Photo-Maker. Авторы используют несколько ракуров и ID-токен

Постановка задачи

Цель: проверить гипотезу о том, что подходы в работах PuLID, InstantID, IP-Adapter распространяются не только на аватары лиц, но и на ростовые аватары.

Задача: улучшить качество генерируемых аватаров, путем обуславливания картинки всем телом человека. Создать подход, позволяющий отказаться от необходимости входныз изображений, сделав возможным генерацию по одному текстовому описанию.

Создание латентного вектора идентичности



Гауссово взвешивание изображения ключевыми точками

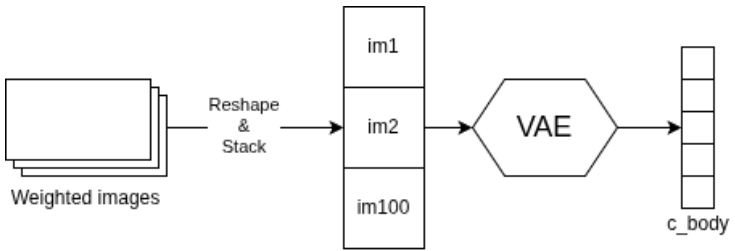
Для ключевой точки с координатами (x_i, y_i), $i = \overline{1, 17}$ вес пикселя с координатами (x, y) определяется гауссовой функцией:

$$w_i(x, y) = \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right), \tag{1}$$

Итоговый вес для пикселя (x, y) - это максимум весов от всех точек:

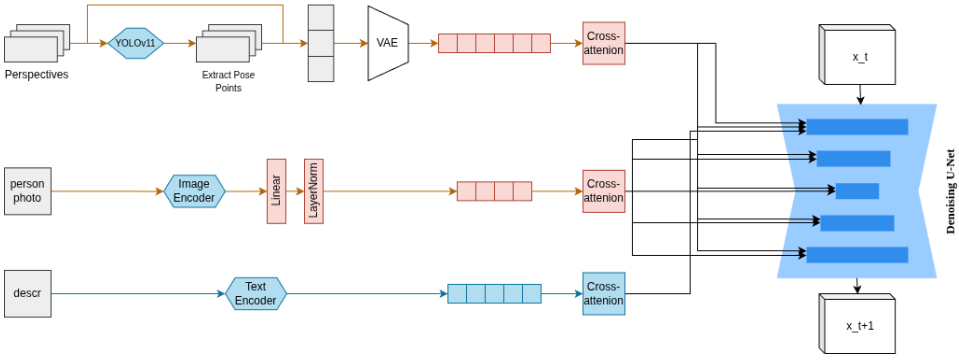
$$W(x, y) = \max_{i=1,\dots,K} w_i(x, y) \tag{2}$$

Создание латентного вектора идентичности



VAE-энкодер выделяет вектор c_{body}, содержащий информацию о теле человека

BoLID: Body-Lightning ID Diffusion



Архитектура модели BoLID, позволяющая на стадии вывода использовать только текстовый запрос

BoLID: Body-Lightning ID Diffusion

вход: Ref Image + TextPrompt + Image Prompt (optional)
выход: Output Image

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \alpha_{\text{id}}\,\mathcal{L}_{\text{id}} + \beta_{\text{CLIP}}\,\mathcal{L}_{\text{CLIP}}, \tag{3}$$

где

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon}\left\|\epsilon - \epsilon_\theta(x_t, \mathbf{c}_\tau, \mathbf{c}_{\text{ref}}, \mathbf{c}_{\text{body}}, t)\right\|^2,$$

$$\mathcal{L}_{\text{id}} = \sum_{k=1}^{17} \left\|p_k^{\text{gen}} - p_k^{\text{ref}}\right\|^2,$$

$$\mathcal{L}_{\text{CLIP}} = 1 - \cos\left(\text{CLIP}(x_{\text{gen}}),\, \text{CLIP}(\tau)\right).$$

x_0 исходная картинка

ϵ шум из распределения $\mathcal{N}(\mathbf{0}, \mathbf{I})$

c_τ, c_{ref} текстовый и визуальный промпты

t шаг диффузионного процесса

p_k^{gen}, p_k^{ref} — координаты k-й точки

Датасет

В качестве датасета были выбраны высококачественные (с разрешением не менее 1000×1000 px) фотографии знаменитостей. Для каждого человека собирался пакет из не менее 100 фотографий с сайта theplace с помощью написанного парсера.

Обработка датасета состояла в удалении изображений с водяными знаками и более чем одним человеком на изображении. Для обнаружения водяных знаков использовалась модель watermark-detection, а модель YOLOv11n-face-detection для проверки количества людей.

Выводы

В ходе данной работы я:

- познакомился с диффузионными моделями
- создал прототип модели, позволяющей агрегировать ракурсы через механизм self-attention
- создал архитектуру модели, позволяющую отказаться от дополнительных элементов контроля
- научился собирать и обрабатывать датасет (парсинг сайта, валидация изображений) для персонализированной генерации.