Consistency Models

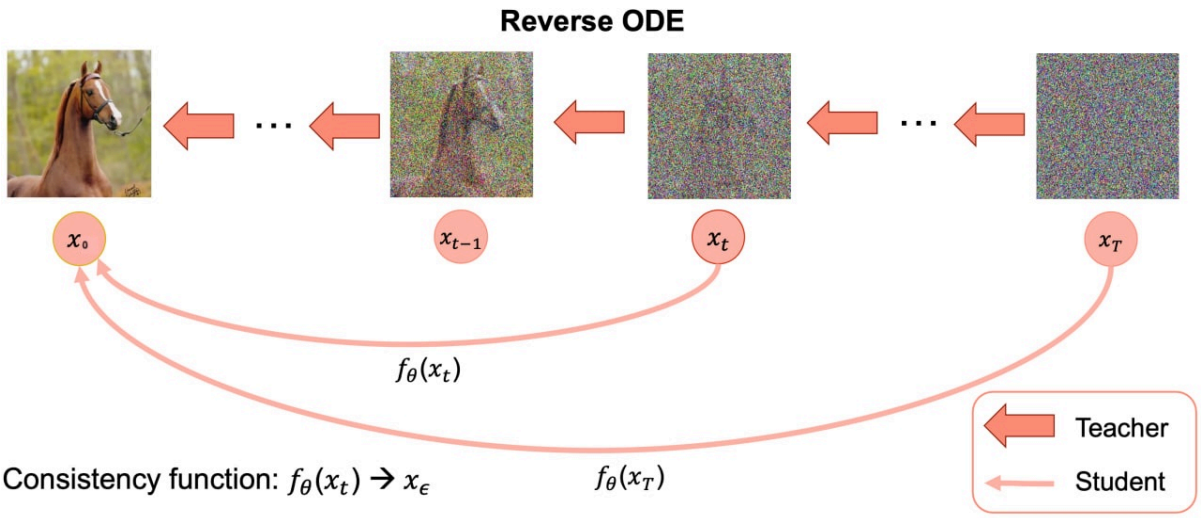
- Выполнил: Грозный Сергей

1. Общая идея

Главная цель дистилляции диффузии - уменьшить количество шагов ДМ, при этом сохранив высокое качество картинок.

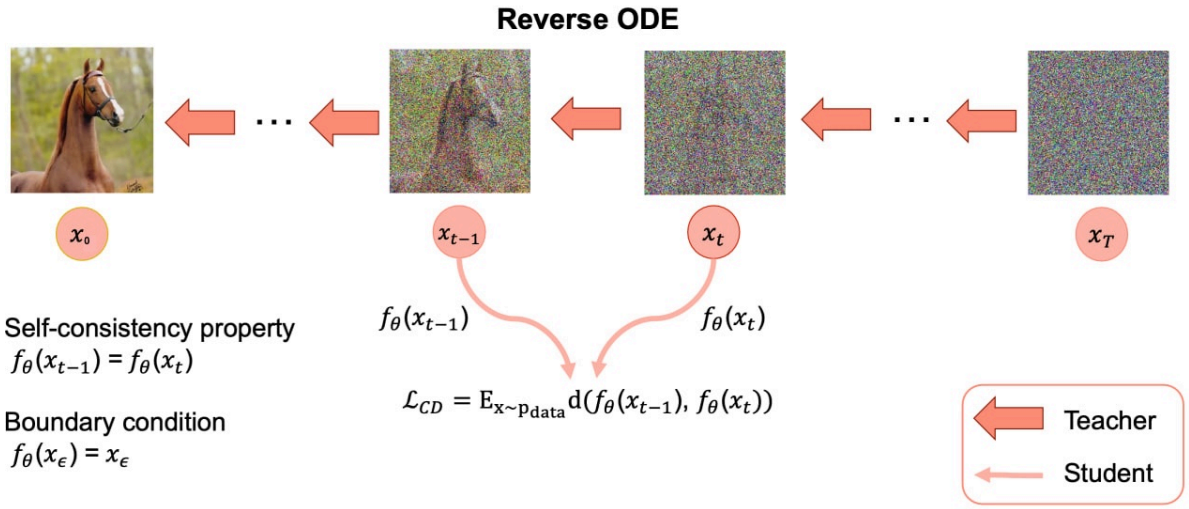
Консистенси модели (Consistency Models | CM) - класс моделей, где мы хотим выучить "консистенси функцию" $f_\theta(\mathbf{x}_t)$ - с любой точки $\mathbf{x}_t$ траектории диффузионного ОДУ **(2)** сразу предсказывать $\mathbf{x}_0$ (чистые данные) за один шаг.
Если мы идеально выучим консистенси функцию, то сможем шагать из чистого шума сразу в картинку, что супер эффективно в отличии от генерации ДМ.

Отметим, что консистенси модель можно учить как независимую генеративную модель, без предобученной ДМ.



Консистенси дистилляция (Consistency Distillation | CD) - подход, когда для обучения CM, мы используем предобученную ДМ. ДМ нам дает качественную инициализацию модели и уже обученную скор функцию, что сильно упрощает сходимость консистенси моделей.

Обучение CM



Главная принцип обучения консистенси моделей заключается в попытке удовлетворить *self-consistency* св-ву: выход CM на двух соседних точках траектории $\mathbf{x}_t$ и $\mathbf{x}_{t-1}$ должен совпадать по какой-то мере близости, например L2 расстояние: $\|f_\theta(\mathbf{x}_{t-1}) - f_\theta(\mathbf{x}_t)\|_2^2$.

Заметим, что self-consistency св-во удовлетворить очень просто без какого-либо обучения, взяв, например $f_\theta(\mathbf{x}_t) \equiv 0$.

Поэтому, чтобы избежать вырожденных решений, нам необходимо выставить граничное условие (boundary condition), которое будет требовать, чтобы в самой левой точке траектории около 0, модель предсказывала картинку, которую получает на вход: $f_\theta(\mathbf{x}_\epsilon) = \mathbf{x}_\epsilon$.

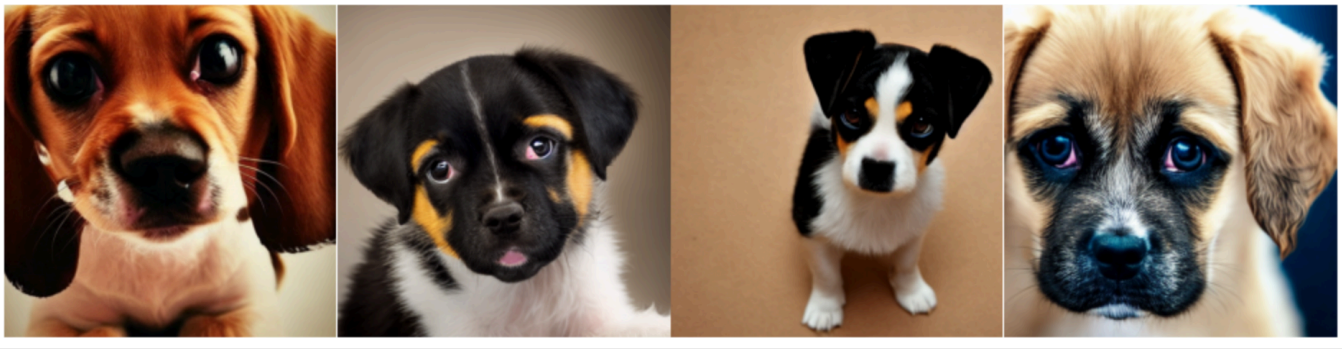
2. Техническое описание

- *Модель*: **SD1.5** - латентная ДМ, т.е. модель работает не в пиксельном пространстве, а в латентном пространстве **VAE**. Таким образом SD1.5 состоит из следующих компонент:
  1. **VAE** - переводит $3\times512\times512$ картинки в латенты $4\times64\times64$ и может декодировать их обратно в картинки.
  2. **Текстовый энкодер** - извлекает текстовые признаки из промпта. Эти признаки будут подаваться в диффузионную модель, чтобы дать модели информацию, что именно хотим сгенерировать
  3. **Диффузионная модель** - UNet, работающий на "латентных картинках" $4\times64\times64$.
- Будем использовать **DDIM** солвер, который является адаптированным методом Эйлера под диффузионный ОДУ. Для консистенси моделей семплирование происходит по следующему алгоритму:
  $x_{t_n} \sim N(0, I)$
  $for\ t_i \in [t_n, \dots, t_1]$:
  - $\epsilon \leftarrow unet(x_{t_i})$

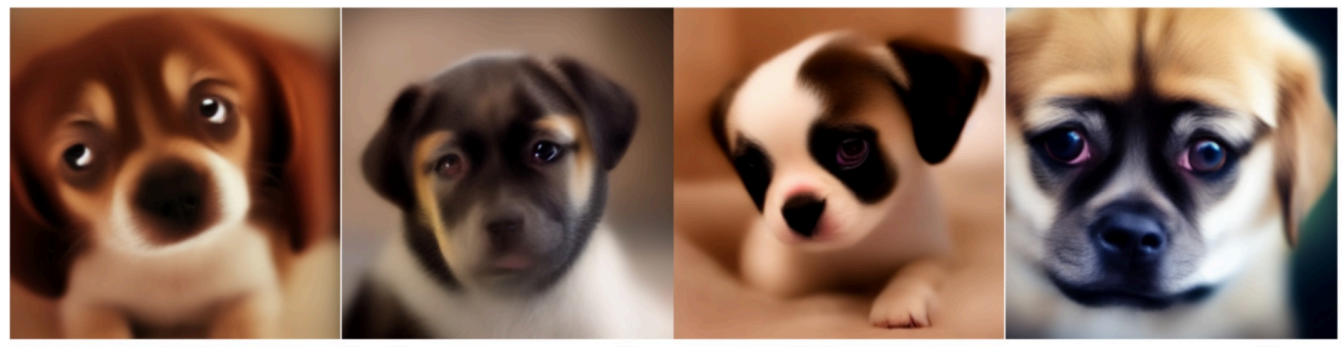
    - $x_0 \leftarrow DDIM(\epsilon, x_{t_i}, t_i, 0)$
    - $x_{t_{i-1}} \leftarrow q(x_{t_{i-1}}|x_0)$
- Чтобы можно было бы запускать обучение при ограничееном бюджете, были сделаны следующие трюки:
    1. gradient checkpointing для обучемой модели
    2. LoRA (Low Rank Adapters) адаптеры, чтобы учить не все веса, а только 10% добавочных весов
    3. Gradient accumulation, чтобы делать итерацию обучения по бОльшему батчу, чем влезает по памяти
    4. Mixed precision FP16/FP32 обучение модели для скорости.
- Обучающая выборка состоит из 5000 пар текст-картинка из COCO датасета. При батчайзе=8 обучение в среднем занимает 30 минут на Tesla T4.

3. Эксперименты

Для ясности, провизуализируем обозначенную проблему. Возьмем предобученный StableDiffusion1.5, зафиксируем сид, промпт и сгенерируем картинки при 50 и 4 итерациях:



Предобученный StableDiffusion1.5. Промпт: "A sad puppy with large eyes". 50 шагов солвера



Предобученный StableDiffusion1.5. Промпт: "A sad puppy with large eyes". 4 шага солвера

Ожидаемо, качество заметно хуже. Хотелось бы иметь модель, которая могла бы генерировать качественные картинки за маленькое число шагов.

Consistency Training

Для начала обучим консистенси модель без учителя (без дестилляции) и провизуализируем результаты семплирования за 4 итерации:

Consistency Training (без учителя). Промпт: "A sad puppy with large eyes". 4 итерации

Картинки теперь менее замыленные, но все равно качество неудовлетворительное.

Consistency Distillation

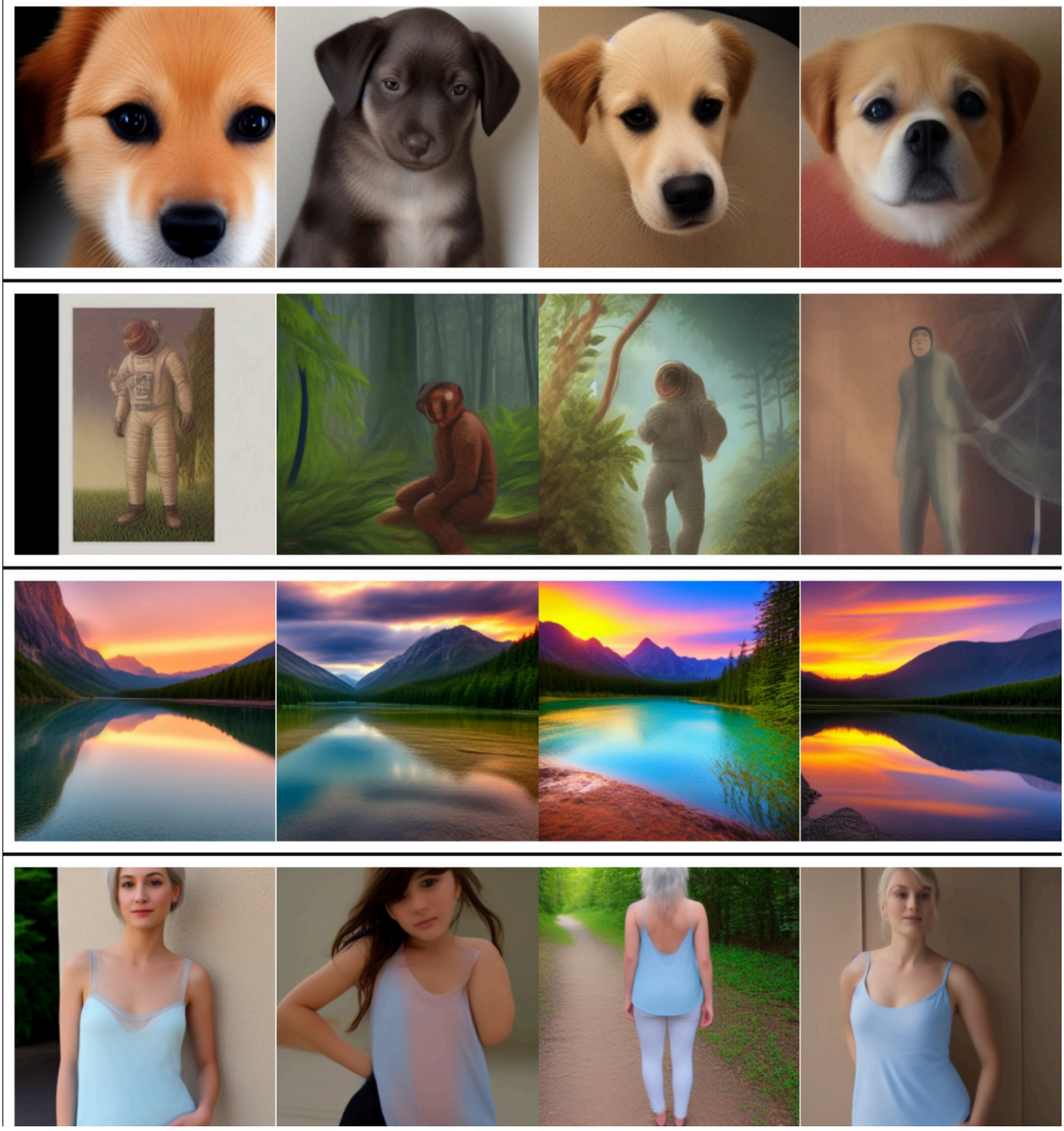
Перейдем к постановке дистилляции, где шаги будут делться с помощью модели учителя (с CFG). Также, заменим стандартный L_2 лосс на pseudo-huber лосс из статьи. Обучим и провизуализиурем результаты:

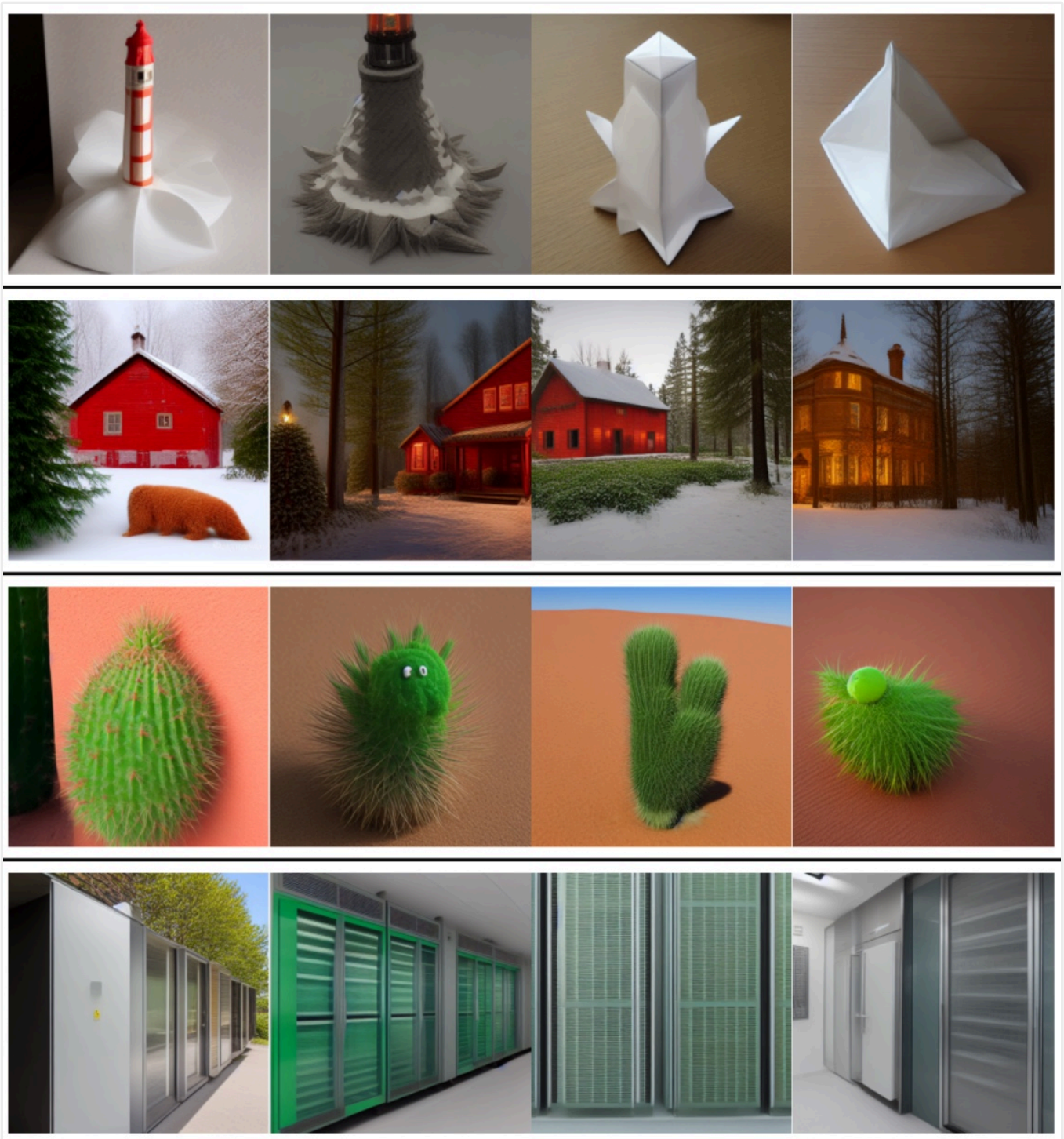
Consistency Distillation (с учителем). Промпт: "A sad puppy with large eyes". 4 итерации

Результат теперь намного лучше: лица собак имеют довольно множество деталей, меньше наблюдаются артефактов.

Провизуализируем теперь несколько промптов:
[
"A sad puppy with large eyes",
"Astronaut in a jungle, cold color palette, muted colors, detailed, 8k",
"A photo of beautiful mountain with realistic sunset and blue lake, highly detailed, masterpiece",
"A girl with pale blue hair and a cami tank top",
"A lighthouse in a giant wave, origami style",

"belle epoque, christmas, red house in the forest, photo realistic, 8k",
"A small cactus with a happy face in the Sahara desert",
"Green commercial building with refrigerator and refrigeration units outside",
]





Consistency Distillation (с учителем). 4 итерации

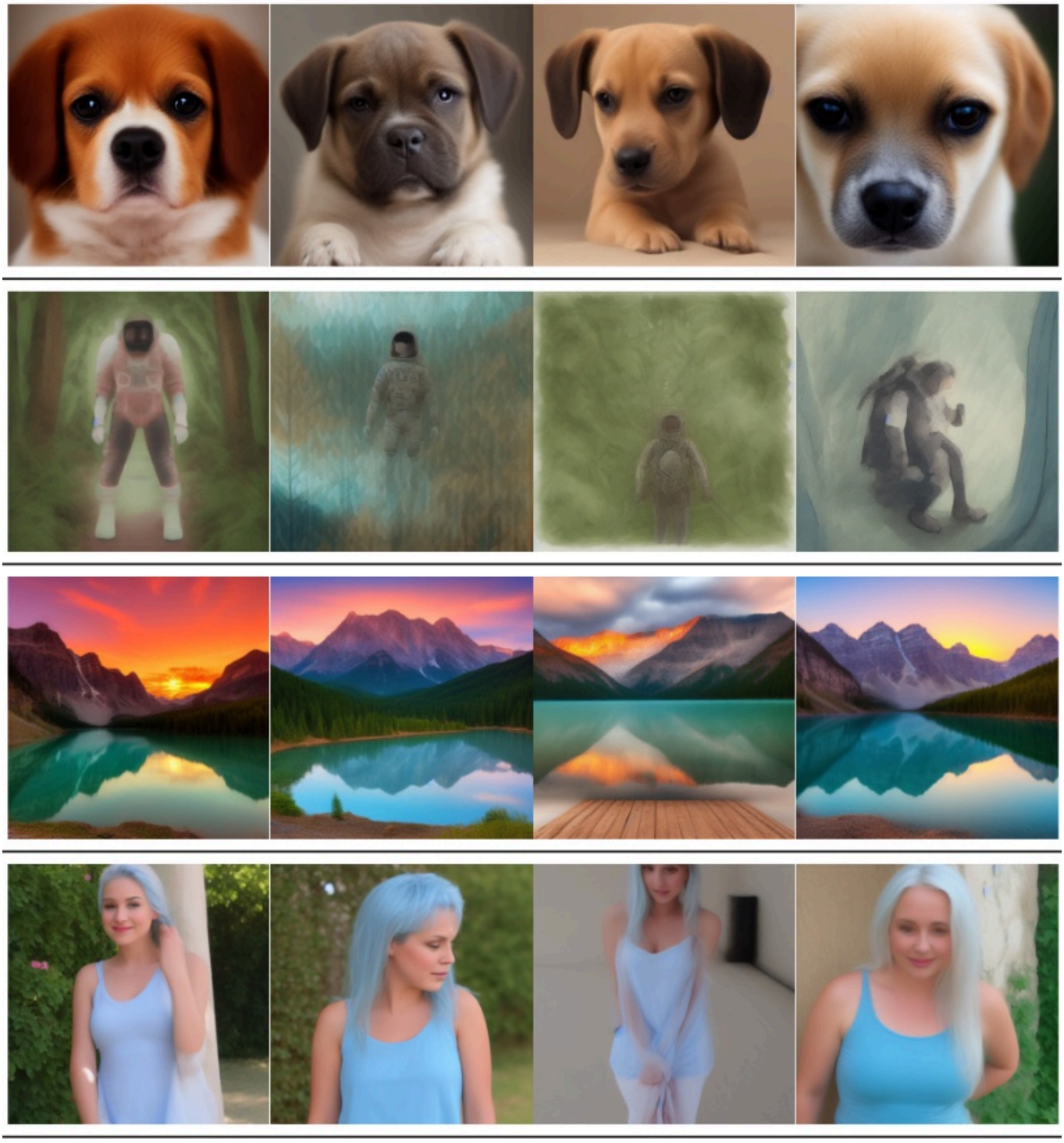
В целом, генерация получилась качественной, однако на всех изображениях не хватает деталей.

Multi-boundary Consistency Distillation

Напоследок, рассмотрим недавнюю модификацию CD, *Multi-boundary CD*, где интегрируем не всю траекторию сразу и потом сэмплируем с возвращением назад, а разбиваем траектории на K отрезков и применяет CD внутри каждого отрезка

независимо. Например, на картинке выше у нас два отрезка: зеленым и красным выделены две граничные точки. Для классического CD, рассмотренного ранее, у нас только одна граничная точка в $t = 0$.

Разобьем на $K = 4$ отрезков, применим к каждому CD и провизуализируем результат:



Multi-boundary Consistency Distillation. 4 итерации

4. Выводы

В рамках исследования были рассмотрены методы Consistency Training, Consistency Distillation и Multi-boundary Consistency Distillation. С каждым этапом качество генерации улучшалось: от изображений с артефактами до детализированных и фотореалистичных.

Обучение модели прошло не без проблем: небольшое отклонение от "правильной" реализации приводило к некачественной генерации (вероятно текущий результат

является ненаилучшим среди всех возможных). Также, несмотря на "лайфхаки", возникали сложности с обучением на видеокартах в силу ограниченности видеопамяти, что показывает трудоемкость обучения таких моделей.

Итого, после проведенных экспериментов, Multi-boundary Consistency Distillation подтвердила свой статус SOTA дистилляции на момент написания отчета. Сгенерированные изображения являются высококачественными и наполнены деталями, а также они сгенерированы всего за 4 шага, что не сравнится с 50 шагами, необходимых для качественной генерации учителя
.