DreamFusion: Text-to-3D using 2D Diffusion

Ben Poole, Ajay Jain, Jonathan T. Barron, Ben Mildenhall

Бакланов Алексей ПМИ-193

Статья описывает принцип работы и обучения инновационной нейросетевой модели, способной генерировать представления 3d объекта и выдавать его 2d изображение исходя из параметров положения камеры в пространстве относительно объекта. Модель состоит из двух нейросетей: NeRF-модели, генерирующей 3d представление и диффузионной модели, используемой при обучении в замороженном состоянии для улучшения качества 2d изображений, выдаваемых с помощью NeRF. Dream Fusion позволяет генерировать 3d представления объектов, при этом не обучаясь на датасетах с 3d данными. Она способна генерировать как конкретный объект, так и необычную сцену по сложному описанию.

Работа написана в сентябре 2022 года. У статьи четыре автора, первые два из которых занимаются диффузионными моделями, вторые - имеют несколько статей по NeRF-моделям:

Ben Poole - научный сотрудник Google Brain. Участвовал также в написании статей про разные диффузионные модели, среди которых авторегрессионные диффузионные модели и диффузионные модели для генерации видео высокого разрешения.

Ajay Jain - исследователь искусственного интеллекта в Калифорнийском университете. Специалист по диффузионным моделям, и масштабируемым системам ML. Был соавтором самой первой статьи про диффузионные модели: “Denoising diffusion probabilistic models”.

Jonathan T. Barron - старший научный сотрудник Google Research в Сан-Франциско, работает в сфере компьютерного зрения и машинного обучения. Автор нескольких статей по NeRF.

Ben Mildenhall - научный сотрудник Google Research, работает над проблемами компьютерного зрения и графики. Автор нескольких статей по NeRF в соавторстве с Jonathan T. Barron.

Предшественниками Dream Fusion можно считать модели Text-to-Image:

DALL-E: https://arxiv.org/abs/2206.09592
Imagen: https://arxiv.org/abs/2205.11487

Конкурентами модели можно считать следующие проекты:

get3d от nvidia: https://nv-tlabs.github.io/GET3D/. В отличие от dream fusion модель использует 3d данные для обучения, а именно - датасет shapenet. Модель является генеративно-состязательной сетью с двумя дискриминаторами: один для текстуры, второй для формы генерируемой 3d модели. На выходе нейросеть выдает куда более качественные модели в сравнении со stable diffusion, однако не способна генерировать сцены по сложным комплексным описаниям.

3d-avatar-diffusion от microsoft: https://3d-avatar-diffusion.microsoft.com/. Данная модель может по портрету и текстовому описанию генерировать 3d аватар. На вход также подается шум, чтобы аватары получались неодинаковыми. Внутри нейросети происходят трехмерные преобразования с помощью диффузионной модели и на выходе мы получаем аватар.

Плюсы статьи:

- Модель не требует 3d данных, благодаря необычному подходу к обучению модели.
- Модель может генерировать как сложные сцены по необычному описанию, так и конкретные объекты. Примеры можно увидеть на сайте https://dreamfusion3d.github.io/index.html.
- В отличие от таких моделей, как get3d, использующих GAN в своей основе и обучающихся на конкретных объектах, модель Dream Fusion может генерировать какие угодно сцены по случайному описанию.

Минусы статьи:

- У получившихся изображений страдает качество: контуры объектов часто размыты, не естественно падает освещение, у животных может быть несколько лиц.
- Модель довольно тяжелая, из-за чего одна 3d модель может генерироваться больше часа.
- Пока проект выглядит скорее как интересная игрушка, которую не понятно, как использовать в реальных проектах