НИС Машинное обучение и приложения, 2022-2023

Обзор на статью DreamFusion: Text-to-3D using 2D Diffusion

Аржанцев Алексей, группа 192

1 Содержание статьи

В отличие от 2D картинок, для которых давно собраны большие датасеты, достаточно хороших датасетов для 3D моделирования пока не существует, а собрать их значительно сложнее чем для обычных картинок. Чтобы решить эту проблему авторы статьи [4] предлагают метод генерации 3D моделей без использования 3D данных для обучения.

Для генерации 3D сцен используется NeRF. Далее он обучается с помощью диффузионной модели предобученной на обычных 2D картинках. Для обучения используется новый SDS лосс разработынный авторами. На вход он принимает изображение 3D сцены со случайного ракурса и с помощью диффузионной модели оценивает насколько это изображение соответствует промту.

Также авторы используют некоторые приемы чтобы улучшить качество картинок. Среди них – использование только некоторых углов камеры при обучении, classifier free guidance, регуляризация геометрии итоговой модели, продвинутые алгоритмы для теней.

2 Авторы статьи

Авторами статьи являются Ben Poole, Jonathan T. Barron, Ben Mildenhall, Ajay Jain. Первые трое из них из Google Research, а последний из UC Berkeley. Предыдущие работы авторов посвященны диффузионным моделям (в частности, Ajay Jain один из соавторов оригинальной статьи про диффузионные модели), а также различным вариациям NeRF.

До DreamFusion все четверо совместно работали над DreamFields, который во многом схож с разбираемой статьей. Подробнее о DreamFields будет сказано в следующем разделе.

3 Предшественники статьи

Авторы упоминают альтернативные способы представления 3D моделей, такие как воксели и облака точек, однако подчеркивают что все предыдущие статьи по теме использовали для обучение 3D данные для которых у нас нет достаточно хороших датасетов. Также, для генерации реалистичных 3D моделей могут быть использованые GANы, но нам все еще понадобится датасет.

Как пример генерации объемных моделей не использующих 3D при обучении авторы приводят свою собственную статью – DreamFields. Её суть во многом повторят DreamFields, только вместо диффузионной модели используется CLIP.

4 Конкуренты

Я могу выделить 4 статьи на тему генерации объемных моделей вышедшие примерно в одно время или после DreamFusion.

4.1 Get3D [1]

Модель от NVidia для генерации текстурированных моделей. В основе используется GAN с двумя дискриминаторами – первый смотрит на модель с текстурой, второй без текстуры. Для обучения необходим большой датасет с 3D моделям. В данном случае авторы воспользовались ShapeNET. Модель обучается на какой-то конкретный класс объектов и может генерировать только его.

Из плюсов хочется отметить, что есть возможность использовать CLIP Guidance. Это может быть использовано как для улучшения качества картинки, так и для добавление специфических деталей.

4.2 3D Avatar Diffusion [5]

Модель от Microsoft принимающая на вход изображение человека и текстовое описание и строящая соответствующую 3D модель. Использует диффузионные модели.

Несмотря на высокое качество моделей, есть очевидные недостатки по сравнению с DreamFusion. Во-первых, это способность генерировать только узкий класс объектов. Во-вторых, потребность в большом датасете, собрать который может быть очень затратно.

4.3 Point-E [3]

Авторы предлагают быстрый способ генерации моделей по промтам в виде облака точек. Метод состоит из трех этапов. На первом этапе по промту рендерится 2D картинка будущей модели. Для этого используется GLIDE дообученный на рендерах объемных моделей с подписями. Далее с помощью двух диффузионных моделей по плоской картинке строится сначала грубая 3D модель, а потом более детальная.

Несмотря на то, что модели получаются значительно хуже чем при использовании DreamFusion, благодаря намного меньшему времени генерации метод может оказаться полезным.

4.4 Magic 3D [2]

Еще одна модель от NVidia. Единственный из методов не использующий 3D при обучении. Используем SDS лосс из DreamFusion. Оптимизация бьется на 2 этапа. На первом этапе получаем модель в низком разрешении. Для представления 3D модели вместо тяжелого NeRF используется более легкий hash grid encoding из Instant NGP. На втором этапе генерируется модель в более высоком разрешении. Для ускорения вычислений используется Latent Diffusion Model, которая предсказывает шум не на оригинальном изображении, а на его представлении более низкого разрешения. На этом этапе оптимизируется

не представлени модели с помощью hash grid encoding, а непосредственно сама модель. Это необходимо, так как разрешение картинки довольно большое. Также такое подход позволяет легче добалять сложные топологические изменения в модель.

Данная статья является непосредственным развитием DreamFusion. Взяв основную идею с использованием диффузионок и SDS лосса, авторы поменяли используемые модели на более легкие, за счет чего смогли в 2 раза сократить время работы и увеличить разрешение с 64x64 до 512x512.

5 Плюсы статьи

К плюсам статьи, в первую очередь, хочется отнести новизну идеи использования 2D моделей для генерации 3D объектов. Авторы сначала применили эту идею в DreamFields, а позже развили её в DreamFusion использовав новые диффузионные модели.

Отличитеьным плюсом полученного метода является то, то он дает возможность генерировать самые разнообразные объекты. Если большинство других способов обучены на какой-то конкретный класс объектов или, как минимум, ограничены тем что было в обучающей выборке, данных метод позволяет генерировать очень сложные сцены. Это отчетливо видно в примере из статьи, где в промту постепенно добавляются новые детали и модель их корректно учитывает.

Также хотелось бы выделить новый лосс предложенный авторами. Авторы показывают из каких соображений они его получили. За счет отбрасывания сложного слагаемого с градиентом U-Net'а лосс становится очень быстрым для вычисления. Также показано, что минимизация данного лосса равносильна минимизации KL дивергенции между некоторыми распределениями на картинках. На мой взгляд, данных лосс может быть применен в других задачах или для оптимизации других моделей помимо NeRF, что мы видели, например, в Magic 3D.

6 Недостатки статьи

Однако использование предложенного метода на практике сопряжено с некоторыми трудностями.

Во-первых качество картинок еще далеко не идеально – на некоторых картинках есть проблемы с освещением, из-за использованного лосса и guidance цвета получаются слишком яркими, у моделей не хватает деталей. Интересным артефактом является то, что у некоторых объектов может появиться 2 лица. Насколько я понимаю, это связанно с тем, что моделька со всех сторон должна хорошо восприниматься диффузинной моделью. Авторы пытаются решить эту проблему передавая в диффузионку также направление камеры, однако, как мы видим, артефакты все еще возможны.

Вторым важным минисом является время работы. Для генерации каждой новой картинки приходится каждый раз с нуля оптимизировать NeRF. Это очень сложный процесс, который занимает больше часа вычислений. Также сложность вычислений не дает возможности сделать разрешение картинки больше чем 64x64, так как тогда время работы

станет неоправдано большим.

К минусам самой статьи хотелось бы отнести малое число экспериментов. По сути авторы сравниваются только со своей предыдущей статьей и с одной единственной другой моделью - CLIP-Mesh. На мой взгляд, стоило добавить больше моделей для сравнений, тем более что многие подходящие модели сами авторы упоминают в начале работы. Но в защиту авторов могу сказать, что использованные модели при обучении используют CLIP и оценка результатов использует CLIP. Так что DreamFusion смог показать более хороший результат даже в сравнении с моделями у которых было явное преимущество.

7 Итог

Идея генерировать 3D модели используя модели для 2D кажется мне очень перспективной. Основными проблемами остаются качество генерируемых картинок и время генерации. Однако эти проблемы видятся мне более техническими и для их решения уже предлгаются методы.

Список литературы

[1] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin1, Daiqing Li1, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. 9 2022.

[2] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. 12 2022.

[3] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point·e: A system for generating 3d point clouds from complex prompts. 12 2022.

[4] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenha. Dreamfusion: Text-to-3d using 2d diffusion. 9 2022.

[5] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. 12 2022.