Дистилляция диффузионных моделей для создания 3D контента

Ельцов Д.А., СПбГУ, Санкт-Петербург sthfaceless@gmail.com

Аннотация

В данной работе исследуется дистилляция диффузионных моделей для генерации высококачественных 3D-ассетов, что актуально для видеоигр, образования и электронной коммерции. Современные методы сталкиваются с проблемами качества данных и разнообразия объектов. Сравнивая наборы данных и оценивая прямую генерацию и дистилляцию знаний, исследование подчеркивает преимущества мета-дистилляции. Предложенный двух-модельный подход улучшает сходимость моделей и визуальное качество. Основные результаты включают повышение 3D согласованности и визуального качества с помощью моделей MVDream и Stable Diffusion.

Введение

Создание качественного 3D контента важно для таких областей, как видеоигры, образование, архитектура и дизайн. В видеоиграх высококачественные объекты окружения создаются командами профессионалов в течение многих лет. В образовании технологии дополненной реальности делают изучение сложных концептов более интерактивным и наглядным. Архитекторы и дизайнеры используют 3D инструменты для создания макетов будущих зданий и комнат.

Основной проблемой в создании генеративных моделей для 3D контента является недостаток качественных данных. Набор данных LAION [8] для 2D изображений содержит около 5 миллиардов объектов с описаниями, после фильтрации которых всё равно остаётся необъятное число изображений. В то время как крупнейший 3D набор данных ObjaverseXL [1] включает всего 10 миллионов объектов, многие из которых простые и низкого качества. Поэтому современные подходы к генерации 3D контента часто используют предобученные 2D модели генерации изображения по тексту из-за их высокой вариативности.

Первый подход включает дообучение моделей для создания нескольких видов объекта и их последующей реконструкции, что быстро, но часто приводит к низкому качеству мешей. Этот процесс можно визуализировать на Рисунке 1.

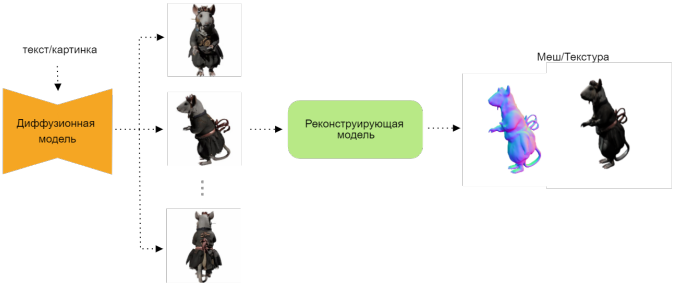

Рис. 1: Прямая генерация 3D объектов с использованием 2D моделей.

Второй подход, являющийся основным объектом исследования, включает инициализацию простым 3D представлением, рендеринг и последующую диффузионную обработку. Этот метод позволяет достичь высокой вариативности и качества рендеринга, зависящих от входного описания или картинки. Однако, он требует значительных вычислительных ресурсов и времени.

Для решения проблемы длительной генерации недавно был предложен оптимизированный подход с использованием дополнительного генератора, который по текстовому описанию создает 3D представление для рендеринга и последующей оптимизации (мета-дистилляция). Этот подход основан на идее, что многие параметры можно обобщить и переиспользовать, что значительно уменьшает среднее время генерации на один объект и улучшает обобщаемость модели. Это делает данный подход наиболее перспективным для исследования. Данный подход проиллюстрирован на Рисунке 2.

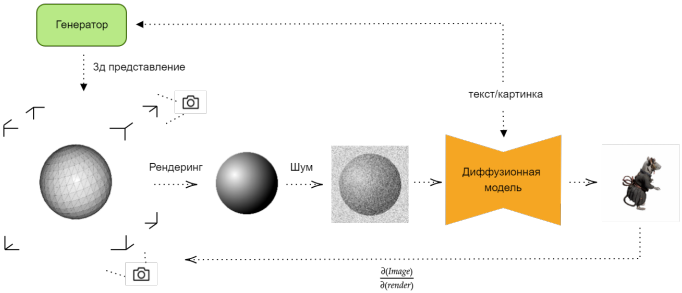

Рис. 2: Мета диффузионная обработка для генерации 3D объектов.

Обзор

На данный момент мета-дистилляция является относительно новой идеей, и хороших статей в популярных журналах пока немного. Основные работы можно упорядочить по увеличению числа текстовых описаний объекта (промптов) и качеству диффузионной модели.

1. Att3D [4]: Базовая работа по мета-дистилляции представлена. В качестве генератора используется обычная полносвязная сеть, а 3D репрезентация представлена в виде неявного объема [5] в низком разрешении.

2. AToM [6]: представлен улучшенный генератор на основе трансформера. Первый этап включает неявный объём, а второй - дифференцируемый меш [**?**]. Однако, диффузионная модель не учитывает 3D согласованность объектов и работает в низком разрешении, что приводит к размытости объектов. В данной работе используется 415 промптов.

Данная работа сосредоточена на качественном улучшении второго подхода с перспективой распространить полученные результаты на большее число промптов. Однако, ни одна из работ пока не имеет открытой программной реализации, поэтому все сравнительные результаты получены с помощью собственной реализации [1].

Результаты

Реализация базового решения

Согласно оригинальной статье, в качестве диффузионной модели используется диффузионная модель в пиксельном пространстве RGB. В качестве 3D репрезентации взят объемный рендеринг [5]. Полученные результаты 3 согласованы с результатами статьи. Основная проблема так называемый Janus face (лицо впереди и лицо на спине), так как взятая 2D модель не может учитывать 3D согласованность. Поскольку модель не использует дополнительных автоенкодеров, а работает в пиксельном пространстве, то имеет низкое разрешение из-за вычислительной емкости, что отражается на качестве объектов. Из плюсов можно выделить простоту и хорошее соответствие промпту.

[1] GitHub реализация.

Рис. 3: Результаты базового подхода на основе AToM.

Улучшение 3D согласованности

Для улучшения 3D согласованности была взята модель MVDream [9], дообученная с 2D диффузионной модели на 3D данных из Objaverse. Проблема Janus face исчезла и практически полностью решается проблема 3D согласованности объектов, но появились новые проблемы: ухудшение визуального качества, соответствия промпту и появление точечных артефактов, которые являются отрицательным эффектом дообучения на низкокачественных 3D данных.

Добавление второй модели

Успешным экспериментом стало добавление в процесс оптимизации второй модели — оригинальной Stable Diffusion [2] с небольшим весом, что позволило улучшить визуальное качество объектов, сохранило 3D согласованность и значительно уменьшило количество точечных артефактов. Минусом стала увеличении времени на одну итерацию примерно в 1.5 раза, что решается следующим результатом.

Детерминированное расписание шума

Следующим успешным результатом стала адаптация подхода DreamTime [3] с оптимизации одного объекта на режим мета-дистилляции. Основная идея заключается в переходе на правильное детерминированное снижение уровня шума в обучении взамен привычного случайного уровня. Данный подход улучшил общую реалистичность изображений и ускорил сходимость примерно в 2 раза. Однако возникла проблема со сходимостью отдельных объектов, которые не успели получить нужное число обновлений для геометрии на соотвествующих уровнях шума. В результате чего было принято решение

начинать обучение со случайного шума и поэтапно переходить в детерминированный вариант для конечных результатов.

Визуальные результаты последовательного применения всех подходов показаны на Рисунке 4.

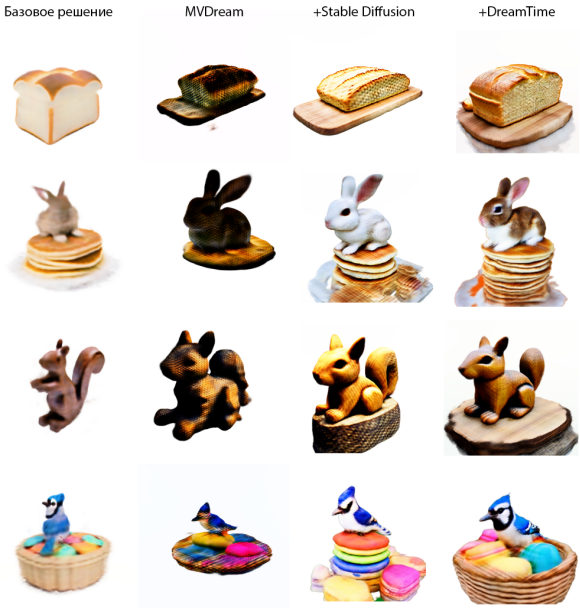

Рис. 4: Визуальные результаты.

Заключение

В результате данной работы был получен ряд успешных экспериментов, значительно увеличивающий 3D согласованность и визуальное качество в сравнении с базовым подходом на основе AToM.

Наиболее перспективные направления для дальнейшего развития включают:

1. Оптимизацию архитектуры генератора для более эффективного использования ресурсов и улучшения соответствия промпту.

2. Дальнейшее улучшение визуального качества и детализации объектов путем перехода на дифференцируемый меш рендеринг.

3. Распространение полученных результатов на большие наборы промптов для получения обобщаемой мета модели.

Список литературы

[1] Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y. and VanderBilt, E., 2024. *Objaverse-xl: A universe of 10m+ 3d objects*. Advances in Neural Information Processing Systems, 36.

[2] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F. and Podell, D., 2024. *Scaling rectified flow transformers for high-resolution image synthesis*. arXiv preprint arXiv:2403.03206.

[3] Huang, Y., Wang, J., Shi, Y., Qi, X., Zha, Z.J. and Zhang, L., 2023. *DreamTime: An Improved Optimization Strategy for Text-to-3D Content Creation*. arXiv preprint arXiv:2306.12422.

[4] Lorraine, J., Xie, K., Zeng, X., Lin, C.H., Takikawa, T., Sharp, N., Lin, T.Y., Liu, M.Y., Fidler, S. and Lucas, J., 2023. *Att3d: Amortized text-to-3d object synthesis*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 17946-17956).

[5] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R., 2021. *Nerf: Representing scenes as neural radiance fields for view synthesis*. Communications of the ACM, 65(1), pp.99-106.

[6] Qian, G., Cao, J., Siarohin, A., Kant, Y., Wang, C., Vasilkovsky, M., Lee, H.Y., Fang, Y., Skorokhodov, I. and Zhuang, P., 2024. *AToM: Amortized Text-to-Mesh using 2D Diffusion*. arXiv preprint arXiv:2402.00867.

[7] Qiu, L., Chen, G., Gu, X., Zuo, Q., Xu, M., Wu, Y., Yuan, W., Dong, Z., Bo, L. and Han, X., 2023. *RichDreamer: A Generalizable Normal-Depth Diffusion Model for Detail Richness in Text-to-3D*. arXiv preprint arXiv:2311.16918.

[8] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. and Schramowski, P., 2022. *Laion-5b: An open large-scale dataset for training next generation image-text models*. Advances in Neural Information Processing Systems, 35, pp.25278-25294.

[9] Shi, Y., Wang, P., Ye, J., Long, M., Li, K. and Yang, X., 2023. *Mvdream: Multi-view diffusion for 3d generation*. arXiv preprint arXiv:2308.16512.

[10] Xie, K., Lorraine, J., Cao, T., Gao, J., Lucas, J., Torralba, A., Fidler, S. and Zeng, X., 2024. *LATTE3D: Large-scale Amortized Text-To-Enhanced3D Synthesis*. arXiv preprint arXiv:2403.15385.