АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОБРАЗОВАТЕЛЬНАЯ
ОРГАНИЗАЦИЯ ВЫСШЕГО ОБРАЗОВАНИЯ
«СКОЛКОВСКИЙ ИНСТИТУТ НАУКИ И ТЕХНОЛОГИЙ»

На правах рукописи

Хахулин Тарас Андреевич

НОВЫЕ ПРЕДСТАВЛЕНИЯ ДЛЯ ИЗОБРАЖЕНИЙ И 3D СЦЕН.

РЕЗЮМЕ ДИССЕРТАЦИИ

на соискание ученой степени

кандидата компьютерных наук

Научный руководитель:

кандидат физ.-мат. наук

Лемпицкий Виктор Сергеевич

Москва — 2024

Оглавление

Резюме

Прогресс нейронных сетей в обучении на данных, основанных на априорных знаниях, открыл новые возможности в интерпретации данных и обучении представлений. В то время как люди с легкостью преобразуют наблюдения в структурированные формы, ключевой аспект интеллекта, искусственные нейронные сети все еще полагаются на определенные упрощения для выполнения этой сложной задачи. Эта работа рассматривает отсутствие единого подхода к представлению информации в различных сферах 3D и 2D сцен.

В этой диссертации представлены различные архитектурные индуктивные предпосылки для получения новых возможностей для изображений и 3D сцен. Модели демонстрируют новые представления пространственных объектов в различных областях: общие статические 3D сцены, аватары людей и изображения. Мы исследуем новую проблему перевода изображений высокого разрешения из неразмеченных данных. Затем мы показываем, как обойти важность архитектурных предпосылок в задаче синтеза изображений, благодаря кодирование позиции пикселя на двумоерной сетки. Далее используя генеративные состязательные сети, мы демонстрируем эффективный синтез новых обзоров для произвольных трехмерных сцен. Передовое представление сцены позволяет оценить точную реконструкцию из разреженного набора входных изображений, и мы вводим методологии для синтеза нового вида через самоулучшающиеся адаптивные представления сцен и техники коррекции ошибок. Затем мы сокращаем разрыв между 3D и 2D методологиями, облегчая контроль над трехмерными представлениями головы человека без опоры на явные многовидовые наборы данных. Изначально сосредотачиваясь на создании аватаров головы в 3D, исследование исследует структуры-заглушки на основе сетки для генерации реалистичных аватаров. Диссертация далее расширяет свое исследование до синтеза аватаров человека в высоком разрешения с латентным управлением позой и выражением эмоций, открывая новые возможности для данной области.

Список опубликованных работ

1. Ivan Anokhin*, Pqvel Solovev*, Denis Korzhenkov*, Alexy Kharlamov*, Taras Khakhulin, Alexy Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin, ``**High-resolution daytime translation without domain labels"** in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, CORE A**.

2. Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov, ``**Image generators with conditionally-independent pixel synthesis"**, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021, CORE A**.

3. T. Khakhulin, D. Korzhenkov, P. Solovev, G. Sterkin, T. Ardelean, and V. Lempitsky, ``**Stereo magnification with multi-layer images"** *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.*

4. Nikita Drobyshev, Evgeny Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor S. Lempitsky, Egor Zakharov. ``**MegaPortraits: One-shot Megapixel Neural Head Avatars"**. *30th ACM International Conference on Multimedia (ACMMM), 2022, CORE A*.*

5. Taras Khakhulin, Vanessa Sklyarova, Victor S. Lempitsky, Egor Zakharov. ``**Realistic One-shot Mesh-based Head Avatars"**. *17th European Conference on Computer Vision (ECCV), 2022, CORE A*.*

6. Pavel Solovev* , Taras Khakhulin*, and Denis Korzhenkov* , ``**Self-improving multiplane-to-layer images for novel view synthesis"**. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, , CORE A.*

* обозначает совместное первое авторство. Автор диссертации внес следующие вклады в статьи, в которых он не является первым автором:

- **High-resolution daytime translation without domain labels**: совместная разработка метода обучения и конвейера тестирования, обнаружение метрик для оценки предлагаемого подхода, проведение экспериментов с переносом стиля и созданием таймлапсов.

- **Image generators with conditionally-independent pixel synthesis**: спектральный анализ генератора, разработка тестового конвейера и проведение экспериментов для выявления свойств модели (например, фовеатный рендеринг, суперразрешение, рисование), написание большой части текста.

- **MegaPortraits: One-shot Megapixel Neural Head Avatars**: совместная разработка второй части модели для получения изображений в высоком разрешении, проведение сравнения базовых условий, включая оценку человека, проведение части экспериментов, написание большой части текста статьи, подготовка рисунков, модельных схем и описаний методов.

Введение

Последний прогресс нейронных сетей в изучении априорной информации из данных заставил нас поверить, что мы можем изучить любое представление из данных, так же как это делают люди. Способность преобразовывать наблюдения мира в структуры является ключевым аспектом интеллекта. Однако, в то время как это может быть простой задачей для людей, в искусственном интеллекте нам все же необходимо вводить некоторые упрощения для упрощения процесса обучения. В настоящее время не существует единого подхода к этому, и даже представление аналогичной информации в 3D и 2D сценах требует различных техник.

Эта диссертация представлена как серия работ, предлагающих различные архитектурные индуктивные предположения вложенные в методы, чтобы раскрыть новые возможности для обработки и получения изображений и представления 3D сцен. В диссертации вводятся различные подходы к обучению представлениям сцен с использованием априорных данных крупномасштабного контента, изученных с применением генеративно-состязательных сетей. Кроме того, исследования расширяют эту линию, охватывая специфичные для человека видеоданные, сокращая разрыв между 3D и 2D подходами и позволяя контролировать представления, позу и эмоции головы человека без явных наборов данных для этого (например без трехмерных сканов).

1.1 Мотивация

Сцена в контексте изображений, видео и 3D-среды — это сложная конструкция, объединяющая физические атрибуты, структуру окружающей среды, временную динамику, эмоциональные, культурные и социальные контексты. В последние годы был достигнут значительный прогресс в анализе сцен, включая распознавание паттернов [25; 36], понимание геометрии [4], текстуры и освещения. Одним из ключевых элементов является представление такой сцены для дальнейшей обработки. В области изображений представление сцены включает в себя интерпретацию и структурирование визуальных

данных на уровне пикселей, распознавание и извлечение информации об объектах, фонах, освещении и пространственных отношениях. Глубокое обучение революционизировало эту область, позволив алгоритмам извлекать сложные паттерны из пикселей в векторы признаков, тензоры или другие форматы.

Изображения, как 2D-сцены, могут быть точно описаны как проекции 3D-сцен, где глубина и пространственные отношения объектов отображаются на плоскую плоскость. Представление сцены в 3D связано со структурированием данных в трехмерном пространстве, где требуется отображение большего количества эффектов. Например, простым способом является определение облаков точек, представляющих объекты. Модели глубокого обучения обрабатывают эти облака точек, чтобы реконструировать и понять 3D-сцены, создавая детализированные 3D-модели городских ландшафтов. Другой подход в 3D — использование вокселей, 3D-эквивалентов пикселей, для представления сцен в виде сетки для детальных объемных представлений. Кроме того, мы можем определить сцену как непрерывное представление, в котором каждой точке пространства сопоставляется функция для отображения ее свойств.

Задача построения представления сцены в 2D и 3D по своей сути сложна из-за высокой размерности и изменчивости данных. Трудности связаны с необходимостью управления разнообразным контентом, различными условиями, фильтрацией шума и даже реалистичным восприятием. Все эти факторы требуют значительной вычислительной мощности и сложных алгоритмов для эффективного понимания и интерпретации контента. Эта сложность побуждает исследователей постоянно разрабатывать и совершенствовать методы, делая представление сцены динамичной и сложной областью на стыке компьютерного зрения и глубокого обучения.

Любая сцена, будь то 2D или 3D, может быть изучена из данных напрямую как пиксельные или воксельные представления, соответственно. Когда мы хотим моделировать определенные распределения и изучать априорные данные из наборов данных, мы пытаемся эффективно извлечь или сжать информацию о сцене из данных. Прямое обучение изображению потребует отдельной пиксельной сетки с N^2 параметрами для каждого объекта в данных. Альтернативный и популярный способ в глубоком обучении — кодировать изображения в векторы для дальнейшего анализа или декодировать обратно для сжатия как можно большего количества информации.

Архитектура кодировщика-декодировщика в автокодировщиках демонстрирует пример индуктивного предубеждения в глубоком обучении, внедряя специфические предположения об обработке и представлении данных, такие как иерархия признаков и восстанавливаемость. Эта архитектура предполагает, что данные можно понять и представить иерархическим образом, где основная информация дистиллируется и представляется в пространстве с более низкой размерностью. Значимые паттерны — это те, которые способствуют точной реконструкции входных данных при прохождении через декодер.

Индуктивные предположения играют важную роль в представлении 2D и 3D-сцен в машинном обучении [2; 16], так как помогает повысить эффективность и расширить обобщение. Эти предпосылки заложенные в модели, по сути, представляют собой набор предположений, которые алгоритм обучения делает о данных, направляют процесс обучения и позволяют алгоритму эффективно интерпретировать сложные сцены, требуя меньше примеров для понимания основных структур. Этот аспект особенно важен в представлении 3D-сцены, где потенциальные данные и их вариации огромны. Такой подход помогает алгоритму обобщать обучающие данные на неочивидные сценарии, предсказывая и понимая новые сцены с большей точностью. Более того, они играют важную роль в устранении двусмысленностей и шума в реальных данных, обеспечивая структуру для их интерпретации, особенно в 3D, где сложность представляют перспектива, тени и скрытые объекты.

Конкретные предубеждения, такие как понимание пространственных отношений и восприятие глубины, особенно полезны для представления сцен. Они не только уменьшают необходимые вычислительные ресурсы благодаря структурированному подходу к обучению, но и облегчают трансферное обучение, когда модель, обученная на одном типе сцены, может быть адаптирована к другому. Эта фундаментальная база адаптируется к различным типам сцен и поддерживает сложные задачи, такие как автономное вождение, робототехника или виртуальная реальность, где понимание и взаимодействие с 2D/3D-средами имеет решающее значение.

Несмотря на жизненно важное значение индуктивных предубеждений для понимания сцены, они могут стать ключевым элементом для связи 2D и 3D в генеративных задачах. Мы рассматриваем популярную задачу генерации 2D-сцен — генерацию изображения на основе соответствующей семантической карты, где каждый пиксель представляет класс. Первые попытки использования генеративных состязательных сетей уже продемонстрировали высокое качество для этой задачи [32]. Однако результаты не могут быть перенесены на случай согласованной генерации различных видов сцены, представленной набором семантических карт (например, различные виды камеры), главным образом из-за высокой неопределенности входных данных. Чтобы убедиться, что

стол на сцене можно сгенерировать согласованно, можно добавить довольно простое многовидовое ограничение в используемое генератором представление [18], которое и будет индуктивным предубеждением в модели.

При обучении такой модели мы явно будем рассматривать сцену как 3D-объект и внедрять согласованность сцены путем отображения в разные камеры. Подобным образом, для создания реалистичных изображений и движений человека мы изучаем представления цифровых аватаров с существующими достижениями в 3D-графике, чтобы функционально описывать голову или тело и определять процедуру нейронного рендеринга таких сцен. Используя типичные предположения или априорные данные из 3D-графики, многие задачи, кажущиеся неразрешимыми обычными методами, дают разумные результаты в задачах, связанных с человеческими сценами.

Мы исследуем различные способы внедрения индуктивного предубеждения в представления статических 2D и 3D-сцен. Как в предыдущем примере, при обучении представлений 3D-сцен на основе набора изображений мы учитываем то, как сцена выглядит с разных перспектив, и вводим определенные ограничения на функцию представления и результаты рендеринга. Набор ограничений и предположений — основной инструмент, используемый в этой работе для облегчения изучения новых 2D и 3D-представлений, которые впоследствии применяются к различным задач.

1.1.1 Индуктивные предположения в представлениях изображений для переноса стиля и генеративных сетей на уровне пикселей

Проблема переноса стиля является одной из четко определенных задач для представления 2D-изображений. В последние годы появилось множество работ, показывающих, как эффективно извлекать стили из изображений [11] или обрабатывать информацию о содержимом отдельно [46]. Многие модели преобразования изображений используют генеративно-состязательные сети с условными генераторами для внедрения информации о целевом атрибуте или домене [6]. Такие генераторы представлены архитектурой на основе сверточных нейронных сетей (CNN). Глубокие сверточные сети показали высокую эффективность в генеративном моделировании. Задачи стилизации [12] и сверхразрешения [21] также выиграли от использования генераторов CNN. Такие модели могут использоваться в качестве эффективного векторного кодировщика с исходного изображения [20]. Кроме того, разбиение изображения на представления контента и стиля [27] — это эффективная схема для редактирования стиля и получения желаемого переноса домена. Большинство работ нацелены на двухдоменную установку [51] или установку с фиксированными дискретными доменами [6]. Мы исследуем преобразование изображений с целью переноса изображений из одного домена в другой, когда

разница между доменами не представлена в наборе данных (например, может быть сложно аннотировать).

Чтобы преодолеть это ограничение, в данной работе мы разрабатываем общее обучение модели преобразования изображений на обширном наборе несогласованных изображений без меток домена и приводим пример того, что индуктивные предубеждения из набора данных и процедуры обучения архитектуры сети могут обеспечить желаемую трансформацию стиля.

Хотя генерация реалистичных изображений началась с новаторской работы [15], в которой были представлены генеративно-состязательные сети (GAN), они всегда опираются на архитектурные индуктивные предубеждения сверточных нейронных сетей. CNN по своей сути предполагают, что близлежащие пиксели связаны друг с другом больше, чем удаленные, что называется локальной пространственной когерентностью и извлечением текстуры [26; 7; 13]. Эти предположения являются важными для эффективного изучения признаков на изображениях, поскольку они отражают способ структурирования объектов и текстур в реальных 2D-сценах, где связанные признаки часто находятся рядом друг с другом. В более современных генераторах была продемонстрирована способность обучаться модулированному ядру сверточного декодера [19; 22; 23], чтобы генерировать фотореалистичные изображения. Мы представляем генератор изображений на основе простой архитектуры MLP, где каждый пиксель обрабатывается независимо от других в отношении одного и того же шума. Эти модели открывают новые возможности в обработке изображений и имеют интересные спектральные характеристики. Предсказывая цвет каждого пикселя независимо и используя уникальные кодировки координат, CIPS демонстрирует инновационный метод синтеза изображений. Дизайн этой модели не только бросает вызов традиционным методологиям, но и открывает горизонт для более гибких и эффективных архитектур с той же парадигмой GAN.

1.1.2 Обобщаемое эффективное представление сцен для нового синтеза ракурсов сцен

Такие методы демонстрируют важность архитектурного и методологического дизайна для успешного решения некоторых 2D-задач. Когда мы начинаем генерировать 3D-сцены, их представление для обработки глубокого обучения становится чрезвычайно важным, как уже упоминалось. Хотя CNN являются эффективным методом для изучения априорных данных по 2D-сценам, их прямое применение в задачах, таких как синтез новых видов (NVS)[39] или даже дискриминативное распознавание объектов[14], сталкивается с трудностями без использования архитектур, учитывающих 3D-структуры

(например, PointNet [33?]). Интуитивное решение — использовать представление, более удобное для индуктивных предубеждений CNN с двухмерной пространственной локальностью в пространстве пикселей. Со временем были разработаны различные методы для создания новых видов. Эти методы можно классифицировать на объемные [30; 31], сетчатые [52; 41? ; 17] и точечные подходы [1; 24], которые, несмотря на явное использование 3D-априорных данных, обычно требуют значительных вычислительных усилий для рендеринга новых видов. Альтернативой с 2D-пространственной структурой является представление на основе карт глубины, часто получаемых из стереосоответствия или оценки монокулярной глубины [37]. Другой ключевой подход включает многослойные полупрозрачные представления [40], которые развивались вместе с достижениями глубокого обучения и напрямую преобразуют объемы стереопары в аналогичные представления [50], полезные для интерактивных приложений с набором плоскостей изображения.

Наша работа снижает требования к памяти для существующего представления MPI [50], вдохновленного техникой 3D-слоев [37], с оценкой как геометрии, так и цвета + прозрачности, выполненной с помощью глубоких сверточных сетей. Мы представляем эту новую парадигму эффективного представления сцены в обучении с глубокими сверточными сетями для стереоизображений. Мы оцениваем различные схемы параметризации для выхода сети и исследуем возможности использования GAN для более правдоподобных изображений. Более того, чтобы устранить смещение в обучении для каждой сцены или в наборе данных, мы собираем свой собственный набор данных с тысячами статических сцен.

Хотя эффективный синтез новых видов на устройстве может быть достигнут с помощью сквозной глубокой сверточной сети, нас по-прежнему ограничивает количество входных видов. В следующей работе мы делаем шаг дальше, пытаясь создать модель, которая будет генерировать многослойное представление для эффективного рендеринга сцены из произвольного количества видов. Наивный подход к расширению входных данных сети путем объединения всех возможных пар невозможен для большого набора входных изображений. Основная идея заключается в использовании нейронной сети с механизмом внимания, которая объединяет информацию со всех видов в одно представление сцены на основе того же понятия из вышеупомянутого многослойного представления. Чтобы передать как можно больше информации с входных изображений, мы вводим новую технику для прямой коррекции ошибок на основе концепции многоплоскостных изображений в каждой сцене [10]. Этот подход открывает возможность для иерархического улучшения методов представления сцены на основе обучения.

Мы также показываем, что, несмотря на простоту нашего представления из предыдущей работы, мы превосходим методы, которые учитывают зависимость видов и теоретически могут быть более точными в некоторых деталях, но на практике не могут достичь нашего качества на сценах с передней перспективой.

1.1.3 Человеческие априорные данные для синтеза видов и поз

Хотя описанные выше методы общего синтеза новых видов представляют собой невероятно точные механизмы для эффективного рендеринга, основное ограничение связано с тем, что захваченные данные — это несжимаемые сцены. Если что-то движется во время захвата на сцене, когда мы снимаем, метод на основе многослойной концепции создаст размытие в таких областях из-за статической природы данных. Съемка динамических сцен с многокамерными системами чрезвычайно дорога, и невероятно сложно разработать методы, поддерживающие временные изменения для синтеза новых видов [28; 45]. Чтобы сделать шаг вперед от статических сцен без использования дорогостоящих методов сбора данных, мы пытаемся сосредоточиться на данных, специфичных для человека, а именно на головах

1.2 Обзор работ

1.2.1 Перевод изображений высокого разрешения для несопоставленных данных

В последние годы сети преобразования изображений продемонстрировали огромные возможности для моделирования и редактирования контента. Однако для этого необходим тщательно подготовленный набор данных с пометками для каждого пикселя или субъективной метаинформацией.

Наше исследование мотивировано ограничениями существующих методов преобразования изображений, которые обычно требуют хотя бы некоторых меток домена для обучения и вывода. Эти методы были успешными при преобразовании между двумя предопределенными доменами (например, Huang et al. [2018], Isola et al. [2017], Liu et al. [2017], Zhu et al. [2017]), а также между несколькими доменами (Choi et al. [2018], Lee et al. [2018, 2019], Liu et al. [2019]). Однако требование наличия меток домена является серьезным ограничением, особенно когда метки сложно определить или собрать, как в случае с изменением времени суток и условий освещения. FUNIT [29] частично решает эти проблемы, используя изображения целевого домена в качестве ориентира для

преобразования в условиях ограниченного количества примеров. Однако аннотации домена остаются необходимыми во время обучения.

Наше исследование продвинулось дальше этого, обучая модель многодоменного преобразования изображений на несогласованных данных без меток домена, используя лишь слабый внешний контроль на основе грубых карт сегментации для повышения эффективности, который не обязателен при численной оценке.

Наша работа включает два основных вклада. Во-первых, она демонстрирует возможность обучения на несопоставленных наборах данных с использованием внутренних и индуктивных предубеждений архитектуры сети и набора данных. Во-вторых, чтобы обеспечить сохранение деталей, HiDT сочетает пропускные соединения [35] с адаптивными нормализациями экземпляров (AdaIN) [19]. Это архитектурное решение отличается от распространенных архитектур AdaIN, не имеющих пропускных соединений.

Экспериментальная оценка модели HiDT включает сравнение с несколькими передовыми базовыми моделями, используя объективные измерения и опрос пользователей. Результаты показывают эффективность модели в таких задачах, как фотореалистичное изменение времени суток для ландшафтных изображений, а также ее потенциал для многодоменных задач стилизации или перекрашивания изображений. Особенность HiDT заключается в решении задачи применения преобразования изображений на высоком разрешении, что часто является вычислительно сложной задачей. Модель предлагает схему улучшения, позволяющую адаптировать сеть, обученную на среднем разрешении, к изображениям высокого разрешения.

В целом, модель HiDT представляет метод преобразования изображений высокого разрешения в ситуациях, где отсутствуют метки домена. Это значительный шаг вперед в области преобразования изображений. Мы существенно превзошли все существующие модели и провели комплексную оценку модели, выделив ее преимущества в работе с изображениями высокого разрешения и ее универсальность в различных задачах обработки изображений.

1.2.2 Генераторы изображений без пространственных сверток

Основным строительным блоком всех генераторов была глубокая сверточная сеть. В этом разделе мы представляем исследование о том, является ли это необходимым и можем ли мы достичь качества современных генераторов без использования функций прямого взаимодействия между пикселями. На протяжении многих лет архитектуры

создавались на основе модели DCGAN [34], интуитивной для сети-декодера изображений, с редким присутствием моделей на основе внимания [49]. Однако в этом исследовании мы фокусируемся на модели, в которой соединение между пикселями невозможно, вдохновленной подходами к воспроизведению отдельных сцен [31; 38].

Мы предлагаем дизайн архитектуры, который позволяет достигать аналогичного качества генерации, как у передового сверточного генератора StyleGANv2 [23]. Эксперименты, проведенные для оценки CIPS, включали сравнение с передовыми сверточными генераторами, такими как StyleGANv2. Основным строительным блоком является периодическая функция активации для передачи информации о пространственном положении пикселя (например, месте на 2D-сетке).

Эти эксперименты продемонстрировали качество генерации архитектуры CIPS и ее уникальные свойства, такие как гибкость и эффективность в использовании памяти. Применения CIPS, по результатам этих экспериментов, выходят за рамки традиционных задач генерации изображений, предлагая новые возможности в областях, требующих синтеза и обработки изображений высокого разрешения без ограничений пространственных сверток. Этот новый подход к генерации изображений открывает новые возможности для применения таких сетей, ранее невозможных.

Мы исследуем новые спектральные свойства нашего генератора, генерацию с ограниченной памятью и задачи, в которых некоторые области могут быть перепредставлены (например, сверхразрешение, фовеантный рендеринг).

Более того, мы показываем, что наши генераторы могут быть применены для изображений высокого разрешения с использованием подхода обучения на основе фрагментов, когда мы загружаем в память только части изображений. Координатные сетки позволяют работать со сложными структурами, такими как цилиндрические панорамы, заменяя базовую систему координат.

1.2.3 Эффективные представления сцен с адаптивной геометрией для стереоизображений

Наша основная мотивация заключалась в том, чтобы решить задачу точного захвата и представления сложных сцен посредством стереоизображений, особенно из несвязанных источников. Существующие методы, хоть и эффективны в определенной мере, не могли справиться с комплексными геометриями. Это ограничение особенно заметно в приложениях [37], требующих высокой детализации и реалистичности, таких как виртуальная реальность и продвинутая реконструкция окружения.

В этой работе мы стремимся решить эти ограничения, предлагая более точный, многослойный подход к представлению сцен, обеспечивающий лучшую точность, глубину и качество изображения в стереоизображениях. Мы нацелены на устранение специфической проблемы в области представления стереосцен — способности точно и эффективно обрабатывать сложные сцены с различными уровнями глубины и сложной детализацией. Предоставляя многослойный подход, StereoLayers предлагает решение, которое более адаптируется и способно справляться с вызовами, присущими обработке стереоизображений. Этот метод демонстрирует сильные способности для представления сцен с эффективным использованием памяти для синтеза новых видов.

1.2.4 Самоулучшающиеся адаптивные представления сцен для синтеза новых видов

Хотя система эффективного синтеза новых видов может быть разработана для стереовхода, это явное ограничение по сравнению с существующими методами. В этой работе мы сосредоточились на улучшении качества многослойного представления и на возможности оценки такого представления из произвольного числа изображений. Кроме того, мы разработали систему, которая не только создает высококачественные представления сцен, но и непрерывно улучшает свою производительность посредством обучения и адаптации с прямым распространением ошибок, вдохновленного DeepView [10].

Для достижения точной реконструкции сцены мы начинаем с прогнозирования фронтально-параллельных полупрозрачных плоскостей низкого разрешения [40; 50], которые содержат основную часть геометрии сцены. Как показано в нашей предыдущей работе, плотная геометрия не является необходимой, и сцена может быть представлена более точно с меньшим количеством параметров. На основе этой идеи мы определяем преобразование прогнозируемых плоскостей в деформируемые слои с учетом метода сквозного обучения. Эта трансформация является ключевым аспектом метода SIMPLI, обеспечивающим более гибкие и точные представления сцен.

Значительной особенностью SIMPLI является процедура прямого улучшения, которая корректирует оцененное представление путем агрегации информации из входных видов. Этот самоулучшающийся механизм обеспечивает непрерывное повышение точности и качества представления.

1.2.5 Аватары головы на основе одной фотографии

Хотя стандартные методы преобразования изображений смогли решить проблему создания аватаров для отдельных объектов [43; 48], такие подходы по-прежнему требуют большого объема обучающих данных и сложны в обучении [44]. Вдохновленные эффективной интеграцией текстур и адаптивной геометрией из предыдущей главы, а также недавними успехами в нейронном рендеринге [42; 31], мы стремимся решить проблему создания персонализированных аватаров.

Одним из ограничений существующих методов, обученных на больших монокулярных видеокорпусах, является узкое поле зрения и ограниченный или сложный контроль над эмоциями и позами. Это нельзя преодолеть без предварительного знания геометрии головы в 3D.

Метод создания реалистичных аватаров головы на основе одной фотографии (ROME) объединяет нейронные сети для рендеринга фотореалистичных 3D-моделей человеческой головы всего по одной фотографии. Наша система использует DECA [9] для точного определения лица и позы в 3D. Следующий шаг включает восстановление сетки головы, здесь мы прогнозируем персонализированную сетку для не лицевых областей, расширяя реконструкцию за пределы лица, чтобы включить всю голову.

1.2.6 Аватары в высоком разрешения на основе нейронных сетей

Иметь прямой контроль над моделью очень удобно для нескольких VR-приложений. Однако такие модели имеют сильную зависимость от базовой геометрической модели (например, FLAME [8]), что ограничивает диапазон эмоций и поз. Несмотря на существующие быстрые латентные модели преобразования изображений для голов людей [48; 47], мы вводим 3D индуктивное смещение в латентное пространство.

Мы определяем каноническую информацию для каждого человека, которую можно извлечь из произвольных изображений, а затем отделяем информацию об эмоциях и позе [3], чтобы создать нейтральные канонические признаки с использованием самообучения латентного движения.

Наконец, мы представляем новый подход к значительному улучшению качества модуля воссоздания по одному изображению путем интеграции набора данных изображений высокого разрешения [23] в учебный процесс, аналогичный первому разделу. Основная идея заключается в сохранении признаков идентичности и обучении модели в самообучающемся режиме [5], когда у нас нет прямого доступа к измененным изображениям.

Заключение

В этой работе были представлены несколько методов для синтеза изображений и 3D сцен. Основная тема, объединяющая работу, - это прямое включение индуктивных предположений о данных, задаче и мире непосредственно в методы синтеза. Этот подход позволяет изучать представления непосредственно из данных, которые затем применимы в дальнейших контекстах.

Ключевые вклады этой работы по нескольким больши направлениям следующие:

1. **Представления изображений**: Высокоразрешающий перевод изображений для несопоставленных данных был установлен как критически важная область, улучшая традиционный перенос стиля новыми техниками.
2. **Генеративные модели**: Был представлен новый подход к генерации изображений, который отходит от традиционных сверточных предубеждений и позволяет использовать продвинутые возможности, такие как распознавание содержания, фокусированное рендеринг и ретуширование изображений.
3. **Синтез сцен**: Были предложены техники эффективного синтеза вида, использующие разреженный набор изображений для реконструкции 3D сцен, что снижает вычислительную нагрузку и увеличивает адаптивность к различным сценам.
4. **Аватары голов**: Был достигнут прогресс в создании аватаров голов за по одному изображению в двух основных направлениях: генеративные аватары, которые описывают 3D голову и параметрическое движение, и аватары на основе латентных представлений предсказанных нейросетями, без использования напрмую 3D моделей.

Потенциальные направления будущих исследований включают:

- Расширение синтеза новых видов на динамические сцены, решая сложности вызванные движением на видео и обработкой монокулярного сигнала,

- Объединение лучших практик из генеративных и сеточных подходов к созданию аватаров для повышения реализма,
- Расширение методологий для охвата полных человеческих тел, включая дополнительные проблемы, вызванные монокулярным видео,
- Соединение классического рендеринга, физических моделей и нейронного рендеринга для создания реалистичных систем рендеринга людей.

Подход генеративного моделирования остается критически важным для достижения высокореалистичного синтеза, с предложением моделей на основе диффузии для генерации более разнообразных и реалистичных образцов. Данная работа показывает как на основе в основном генеративно-соотязательных сетей достигать качественно новых результатов в различных задачах, но предложенные методы во многом ложатся и на альтернативные варианты генеративных моделей, которые могут изучаться в дальнейшем как часть предложенных в работе методов.

Список литературы

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020.

[2] Peter Battaglia, Jessica Blake Chandler Hamrick, Victor Bapst, Alvaro Sanchez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andy Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Jayne Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. In *NeurIPS*, 2018.

[3] Egor Burkov, I. Pasechnik, Artur Grigorev, and Victor S. Lempitsky. Neural head reenactment with latent pose descriptors. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13783--13792, 2020.

[4] Angel Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. arXiv, 2015.

[5] Shuaijun Chen, Zhen Han, Enyan Dai, Xu Jia, Ziluan Liu, Xing Liu, Xueyi Zou, Chunjing Xu, Jianzhuang Liu, and Qi Tian. Unsupervised image super-resolution with an indirect supervised path. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[6] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789--8797, June 2018.

[7] Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. In *ICLR*, 2017.

[8] Andrew Feng, Dan Casas, and Ari Shapiro. Avatar reshaping and automatic rigging using a deformable model. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pages 57--64. ACM, 2015.

[9] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40:1 -- 13, 2020.

[10] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Styles Overbeck, Noah Snavely, and Richard Tucker. Deepview: High-quality view synthesis by learned gradient descent. In *CVPR*, 2019.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414--2423, June 2016.

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, pages 2414--2423, 2016.

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[14] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *CVPR*, 2022.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, pages 2672--2680, 2014.

[16] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478, 2020.

[17] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. In *ACM TOG*, 2018.

[18] Hsin-Ping Huang, Hung-Yu Tseng, Hsin-Ying Lee, and Jia-Bin Huang. Semantic view synthesis. In *ECCV*, 2020.

[19] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510--1519, Oct 2017.

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-Image Translation. In Vittorio Ferrari, Martial Hebert, Cristian

Sminchisescu, and Yair Weiss, editors, *Computer Vision -- ECCV 2018*, pages 179--196, Cham, 2018. Springer International Publishing.

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, pages 694--711, 2016.

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4396--4405, 2019.

[23] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proc. CVPR*, pages 8107--8116, 2020.

[24] Leif Kobbelt and Mario Botsch. A survey of point-based techniques in computer graphics. *Computers & Graphics*, 2004.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 -- 90, 2012.

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278--2324, 1998.

[27] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: diverse image-to-image translation via disentangled representations. *CoRR*, abs/1905.01270, 2019.

[28] Tianye Li, Miroslava Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, S. Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022.

[29] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[30] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. In *ACM TOG*, 2019.

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proc. ECCV*, pages 405--421, Cham, 2020. Springer International Publishing.

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2019.

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211--252, 2015.

[37] M. L. Shih, S. Y. Su, J. Kopf, and J. B. Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020.

[38] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Proc. NeurIPS*. Curran Associates, Inc., 2020.

[39] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019.

[40] Richard Szeliski and Polina Golland. Stereo matching with transparency and matting. 1999.

[41] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. In *ACM TOG*, 2019.

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[43] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: real-time face capture and reenactment of rgb videos. *ArXiv*, abs/2007.14808, 2019.

[44] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[45] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022.

[46] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic Style Transfer via Wavelet Transforms. *arXiv:1903.09760 [cs]*, March 2019. arXiv: 1903.09760.

[47] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor S. Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020.

[48] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. 2019.

[49] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019.

[50] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM TOG*, 2018.

[51] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242--2251, October 2017.

[52] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. In *ACM TOG*, 2004.