Автономная некоммерческая образовательная организация высшего образования
“Сколковский институт науки и технологий”

На правах рукописи

Рахимов Руслан Ильдарович

МЕТОДЫ ПОВЫШЕНИЯ ОБОБЩАЮЩЕЙ СПОСОБНОСТИ МОДЕЛЕЙ В ЗАДАЧАХ 3D КОМПЬЮТЕРНОГО ЗРЕНИЯ

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата компьютерных наук

Научный руководитель:
доктор физико-математических наук
Бурнаев Евгений Владимирович

Москва — 2024

Содержание

1. Введение

Решение задач 3D компьютерного зрения направлено на уменьшение разрыва между цифровым и физическим мирами за счет всестороннего понимания трехмерного пространства на основе двухмерных данных. Несмотря на значительные достижения все еще остается основная проблема - необходимо улучшение обобщающей способности моделей 3D компьютерного зрения для надёжной работы в разнообразных, ранее неисследованных условиях.

В данной диссертационной работе рассматриваются методы повышения обобщающей способности моделей в задачах 3D компьютерного зрения, что позволяет существенно улучшить адаптивность и эффективность этих моделей в различных прикладных задачах. **Цель** данного исследования заключается в повышении возможностей систем компьютерного зрения в трехмерном пространстве для более эффективного решения ряда задач включая генерацию синтетических данных, создание точных трехмерных реконструкций, эффективную визуализацию объекта с новой точки обзора и определение позы человека.

1.1. Контекст и мотивация

При решении основных взаимосвязанных задач 3D компьютерного зрения, каждая из которых критически важна для интерпретации и реконструкции сложной природы окружающего трехмерного мира, требуется, чтобы соответствующие методы имели хорошую обобщающую способность. Среди этих задач есть первоначальный сбор данных и далее регистрация, реконструкция, динамическая интерпретация и визуализация 3D-сред. На каждом шаге моделям необходимо не только понимать и обрабатывать большие объемы даных, но и точно и эффективно работать в сценариях, для которых они не были специально обучены.

В основе 3D компьютерного зрения лежит критически важный процесс реконструкции [24, 83, 58, 59], где сырые данные преобразуются в детализированные 3D-модели, как статические, так и динамические. На первом шаге из различных источников собирается «сырая» визуальная информация, например, RGB изображения с камеры, или синтетические данные. От качества данных зависят все дальнейшие шаги и финальные результаты анализа и реконструкции.

За сбором данных следует критически важный этап регистрации, когда различные наборы данных пространственно выравниваются и интегрируются [5]. Качественное решение задачи регистрации гарантирует, что решение последующих задач, например, 3D реконструкции, будет основано на едином наборе данных, который точно отражает геометрические и пространственные отношения в рассматриваемой 3D сцене.

Этап реконструкции начинается после этапа регистрации. На этом этапе обработанные выровненные данные используются для создания цифровой 3D модели. Алгоритмы интерпретируют и объединяют данные с помощью различных методов, например, триангуляции или реконструкции поверхностей [28, 40], что позволяет получить детализированное трехмерное представление. Способы представления результатов варьируются от облаков точек до сложных форматов, таких как сеточные модели, и даже текстурированные 3D модели, которые позволя-

ют реалистично представлять различные особенности поверхности. 3D CAD-модели, которые имеют важное значение в таких областях, требующих высокой точности, как инженерия и архитектура, могут тоже являться результатом описанного процесса. Традиционные подходы построения 3D моделей по данным реальных измерений зачастую не эффективны при работе с данными высокого разрешения и из-за наличия шума в этих данных.

Задача генерации новых видов часто возникает либо на шаге после процесса реконструкции, либо одновременно с ним. Решение этой задачи включает в себя создание реалистичных изображений с точек обзора, которые изначально не были захвачены во время сбора данных. Значительная проблема заключается в разработке модели, способной эффективно обобщаться на неисследованные сцены и быстро обрабатывать входные данные для создания новых видов.

Точная интерпретация динамических 3D сред, особенно тех, которые включают взаимодействия людей, имеет важное значение. Это особенно актуально в приложениях, таких как оценка позы человека для дополненной реальности и приложений для виртуальной примерочной. В отличие от традиционных методов, которые обычно фокусируются на идентификации ключевых суставов или точек тела [91], плотная оценка позы человека [3] подразумевает полное картирование формы человеческого тела и сопоставление каждого пикселя изображения с соответствующей 3D точкой на карте поверхности тела [51]. Это позволяет более детально «понимать» позу и движения тела человека. Однако, текущие архитектуры требуют значительных вычислительных ресурсов, что мешает их применению в интерактивных сценариях реального мира.

Наконец, в области человекоцентрической 3D реконструкции, важной для создания виртуальных аватаров, существует задача выполнения реконструкции по одному изображению, в отличии от стандартных методов, которые предполагают наличие множества изображений [2, 25, 4]. Решение задачи реконструкции в таком сценарии возможно только если модель будет иметь хорошую обобщающую способность при обработке данных разных людей.

1.2. Актуальность исследования

Область 3D компьютерного зрения достигла значительного прогресса, тем не менее, имеется ряд вызовов, которые ограничивают широкое применение различных методов, особенно если требуется хорошая обобщаемость на разнообразные и сложные среды.

Одна из возможностей решить проблему обобщения моделей 3D компьютерного зрения в различных средах состоит в использовании синтетических данных. Хотя генеративное обучение и позволяет создавать реалистичные синтетические данные, генерация видео остаётся ресурсоёмкой задачей, решение которой часто не имеет желаемого качества [52].

В геометрическом моделировании методы обнаружения особенностей 3D объектов (например, особых кривых, поверхностных линий, вдоль которых поле нормалей испытывает излом) требуют тщательной настройки параметров для каждой модели, соответственно, затруднена масштабируемость [90, 16]. Стандартные стратегии, такие, как сегментация поверхностей и

подгонка патчей (фрагментов), хотя и устойчивы к шуму, тем не менее они не обладают гибкостью и вычислительной эффективностью [50, 9]. Аналогичным образом, модели машинного обучения для классификации особенностей не эффективны при работе с зашумлёнными данными [27, 31].

Традиционные методы решения задачи генерации новых видов, включая интерполяцию видов и рендеринг светового поля, часто неэффективны в случае сложных геометрических особенностей объектов и разнообразных условиях освещения [47, 76]. Продвинутые техники, такие как поля нейронного излучения (NeRF) и воксельные методы, требуют значительных вычислительных затрат и оптимизации параметров моделей [56, 38]. Методы нейронной графики на основе точек (NPBG) улучшают качество рендеринга, но требуют подстройки для каждой сцены, что ограничивает их использование [1].

Современные модели для оценки позы человека позволяют получить точные результаты, но не подходят для использования на мобильных устройствах из-за значительных вычислительных затрат [3, 98]. Хотя и существуют улучшения этих моделей [62, 61], они ещё не достаточно оптимизированы по размеру и скорости для практического использования на мобильных устройствах.

Более того, хотя подходы к моделированию внешности и головы человека в 2D уже достаточно продвинуты, моделирование в 3D часто зависит от ограниченных данных, таких как 3D-сканы [39, 17, 74]. Новые методы, использующие неявные представления, такие как NeuS и VolSDF, имеют потенциал, однако, соответствующие модели затруднительно адаптировать к новым сценам [86, 63, 99, 41].

Вызовы, перечисленные выше, подтверждают необходимость данного исследования с целью повышения устойчивости, эффективности и практичности технологий трёхмерного компьютерного зрения с учетом существующих ограничений и для лучшего соответствия требованиям реальных приложений.

1.3. Цели и область исследования

Цель данной диссертационной работы заключается в разработке и внедрении новых методов и подходов, направленных на улучшение обобщающей способности моделей в задачах 3D компьютерного зрения. Для достижения этой цели были поставлены следующие задачи:

1. Исследовать возможность улучшения обобщения модели видеогенерации при ограничении вычислительных ресурсов при обучении.

2. Разработать метод предсказания геометрических особенностей в 3D моделях с повышенной обобщаемой способностью при работе с новыми, ранее не виданными 3D моделями разных масштабов и наличием шумов сканирования.

3. Разработать подход для генерации новых видов, эффективно обобщаемый на новые сцены и не требующий трудоемкой оптимизации.

4. Улучшить обобщение модели для плотной оценки позы человека, добившись высокой производительности и качества работы при строгих ограничениях на размер и скорость модели.

5. Улучшить обобщающую способность алгоритмов для реконструкции 3D-портретов головы, чтобы они эффективно работали на основе одного входного изображения.

1.4. Результаты

Работа основана на использовании **методологии и методов** машинного обучения, глубокого обучения и компьютерного зрения.

Достоверность результатов обеспечена правильным применением проверенных научных инструментов для исследований и анализа. Разработанные алгоритмы были экспериментально протестированы на множестве задач, с использованием как синтетических, так и реальных наборов данных. Подробные отчеты о проведенных экспериментах, открытый исходный код и доступ к данным позволяют воспроизвести полученные результаты. Исследования были опубликованы в ведущих научных журналах и представлены на конференциях по компьютерному зрению.

Основные положения, выносимые на защиту:

1. Исследование возможности моделирования видео в дискретном латентном пространстве.

2. Регрессионный метод локализации особых кривых 3D объектов, который позволяет надежно обрабатывать зашумленные, высокоразрешенные 3D-данные и превосходит существующие методы.

3. Модель для генерации новых видов сцены по набору изображений этой сцены, которая эффективно обобщается на данные новых сцен без дополнительного обучения.

4. Модель для эффективного решения задачи плотной оценки позы человека, которая может быть развернута на мобильном устройстве.

5. Адаптация алгоритма 3D реконструкции головы человека на основе одного изображения для использования с неизвестными параметрами камеры.

1.5. Значимость работы

В данной диссертационной работе предложены новые подходы, которые позволяют повысить обобщаемую способность решений задач 3D компьютерного зрения на различных этапах построения 3D моделей. Предложенный новый метод к генеративному моделированию видео данных [68] имеет схожую с существующими методами производительность, но при этом требует в разы меньших вычислительных ресурсов для обучения модели. Разработанная модель для предсказания геометрических особенностей по трёхмерным облакам точек [53], обученная на синтетических данных с минимальным дообучением на реальных данных, позволяет

получить точные прогнозы для реальных 3D объектов. Предложенная модель для генерации новых видов [66] не требует дообучения по данным новой сцены и при сравнимом качестве визуализаций позволяет достигать скорости до 22 кадров в секунду, что существенно выше скорости существующих передовых подходов. Для плотной оценки позы человека в реальном времени была разработана модель [67], позволяющая достичь оптимальный баланс между производительностью и качеством работы, за счет которого удалось развернуть модель на мобильном устройстве. Разработана модель для трехмерной реконструкции головы человека, которая способна работать по данным одной фотографии и эффективно обобщается на данные новых людей [8].

Представленные методы имеют широкий потенциал применений в различных областях, в частности, в приложениях дополненной и виртуальной реальности, робототехнике и других отраслях, а также способствуют улучшению понимания способов повышения обобщающей способности моделей в задачах 3D компьютерного зрения.

2. Публикации и апробация работы

Диссертационная работа основана на следующих пяти основных статьях, все из которых индексируются в SCOPUS и Web of Science.

Публикации первого уровня

1. *Рахимов, Р.**, Арделян, А. Т.*, Лемпицкий, В., & Бурнаев, Е. (2022). *NPBG++: Accelerating Neural Point-Based Graphics.* Представлена на конференции IEEE/CVF Conference on Computer Vision and Pattern Recognition (стр. 15969-15979). CVPR 2022. CORE A*.
   https://doi.org/10.1109/cvpr52688.2022.01550
   Резюме: В работе представлена модель для генерации новых фотореалистичных видов сцены с новых ракурсов на основе ограниченного набора изображений. Модель включает новый подход для предсказания нейросетевых дескрипторов из исходных изображений за один проход, устраняя необходимость в трудоемкой оптимизации модели для новой сцены. Предложенная модель значительно сокращает время моделирования сцены и улучшает качество рендеринга. Извлекая признаки на основе архитектуры U-Net с последующей агрегацией дескрипторов, инвариантных к перестановкам, и применяя нейросеть-уточнитель, модель обеспечивает эффективный и высококачественный рендеринг. Эксперименты показывают, что модель по качеству сравнима с ведущими методами для генерации новых видов, но при этом может гораздо быстрее проводить рендеринг и моделировать сцену, что делает модель перспективной рдля реальных приложений в виртуальной реальности, кинематографии и игровой индустрии.

* - Равный вклад

Основной вклад: Рахимов Р.И., как основной соавтор, разработал и реализовал основные модули предложенной модели NPBG++ и внес значительный вклад в эксперименты.

2. Матвеев, А., *Рахимов, Р.*, Артемов, А., Бобровских, Г., Егиазарян, В., Богомолов, Е., Паноццо, Д., Зорин, Д., & Бурнаев, Е. (2022). *DEF: Deep Estimation of Sharp Geometric Features in 3D Shapes.* Представлена на конференции SIGGRAPH. Опубликована в трудах конференции в журнале ACM Transactions on Graphics, 41(4). ACM ToG. CORE A*. https://doi.org/10.1145/3528223.3530140

   Резюме: В работе представлен новый метод DEF локализации особых кривых 3D объектов, основанный на моделировании скалярного поля расстояний от точек на поверхности объекта до ближайшей особой кривой. Метод DEF использует модели на основе глубоких нейросетей, которые по данным локальных патчей (фрагментов) карты глубины рассчитывают прогноз. Агрегация этих прогнозов позволяет повысить его устойчивость и адаптивность, при этом для обучения моделей используются в основном синтетические данные, а также дообучение по небольшому объему реальных данных. Ключевые компоненты метода DEF включают генерацию обучающих данных, предиктивную модель по данным локальных патчей, метод агрегации прогнозов для целой 3D модели объекта и финальное извлечение параметрических особых кривых. Предложенный метод значительно превосходит существующие методы обнаружения особых кривых с точки зрения точности и масштабируемости на различных наборах данных. Вычислительные эксперименты демонстрируют способность метода эффективно обрабатывать большие наборы данных, что делает его мощным инструментом для геометрического анализа и 3D компьютерного зрения.

   Основной вклад: Рахимов Р. И. разработал первый из двух основных компонентов метода - регресионную модель для локализации геометрических особенностей. Автор также внес значительный вклад в проведение вычислительных экспериментов на синтетических данных по оценке эффективности предиктивных моделей, обрабатывающих данные отдельных патчей, и был ответственен за дообучение моделей на реальных данных.

3. *Рахимов, Р.**, Богомолов, Е.*, Нотченко, А., Мао, Ф., Артемов, А., Зорин, Д., & Бурнаев, Е. (2021). *Making DensePose Fast and Light.* Представлена на конференции IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1869-1877). WACV 2021. CORE A. https://doi.org/10.1109/wacv48630.2021.00191

   Резюме: В работе представлена модель Mobile Parsing R-CNN для получения плотной оценки позы человека (DensePose) в реальном времени на мобильных устройствах. Существующие модели содержат большое количество параметров и требуют надежной серверной инфраструктуры для своей работы. В данном исследовании предложена новая архитектура, которая позволила значительно сократить размер модели (в 17 раз) и время её

* - Равный вклад

работы (в 2 раза) по сравнению с предыдущими подходами. Эксперименты показывают, что итоговая модель не только обеспечивает высокую точность оценки позы, но и является более эффективной, что делает модель перспективной для приложений, требующих понимания человеческой позы в режиме реального времени на мобильных устройствах.
Основной вклад: Рахимов Р.И., как основной соавтор, сыграл ведущую роль в разработке и реализации архитектурных улучшений модели и внес значительный вклад в эксперименты.

4. Бурков, Е., *Рахимов, Р.*, Сафин, А., Бурнаев, Е., & Лемпицкий, В. (2023). *Multi-NeuS: 3D Head Portraits from Single Image with Neural Implicit Functions.* Опубликована в журнале IEEE Access, уровень Q1. https://doi.org/10.1109/access.2023.3309412
Резюме: В работе предложен подход для реконструкции текстурированной 3D модели человеческой головы из одного или нескольких изображений. Представленная модель Multi-NeuS использует подход на основе неявной нейросетевой поверхности NeuS, в которой различные слои нейронной сети отвечают как за общую для разных сцен информацию, так и за специфичную для данной сцены. Это позволяет одной и той же моделью эффективно моделировать сразу несколько объектов или сцен. Модель эффективно представляет как геометрические, так и текстурные особенности, и, в отличии от предыдущих подходов, не требует значительных наборов данных или результатов 3D сканирования для своего обучения. С помощью мета-обучения Multi-NeuS позволяет получить обобщаемое представление, которое затем дообучается под отдельную сцену. Эксперименты подтверждают способность модели производить высококачественные реконструкции на основе минимального набора входных данных.
Основной вклад: Рахимов Р. И. внес значительный вклад в различные этапы процесса разработки модели, такие как предварительная обработка данных, оптимизация параметров камеры, код для подсчет метрик, извлечение полигонального представления сцены, адаптация алгоритма для работы с изображениями с неизвестными параметрами камеры.

Публикации второго уровня

1. *Рахимов, Р.**, Волхонский, Д.*, Артемов, А., Зорин, Д., & Бурнаев, Е. (2021). *Latent Video Transformer.* Представлена на конференции 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. VISIGRAPP 2021. CORE B. https://doi.org/10.5220/0010241801010112
Резюме: В работе представлена модель для решения задачи генерации видео, заключающейся в прогнозировании будущих кадров видео на основе заданной последовательности начальных кадров. Решение данной задачи имеет потенциальное применение в таких областях как технологии автономного вождения, обнаружение аномалий в видео

* - Равный вклад

данных и создание анимационного контента. Предложенная модель использует комбинацию автоэнкодера (автокодировщика) кадров, в частности архитектуру VQ-VAE для кодирования кадров с помощью представления в дискретном скрытом (латентном) пространстве, и авторегрессионную генеративную модель для предсказания последующих кадров на основе архитектуры трансформера. Модель последовательно генерирует новые кадры в скрытом пространстве с последующим их отображением обратно в пиксельное пространство, что эффективно уменьшает требования к вычислительным ресурсам. Такого рода подход позволил значительно снизить вычислительные затраты на обучение модели при сохранении качества моделирования видео данных в сравнении с существующими методами. Модель была протестирована на наборах данных, таких как BAIR Robot Pushing и Kinetics-600, и продемонстрировала конкурентоспособные результаты.

Основной вклад: Рахимов Р.И, как основной соавтор, предложил идею переноса процесса генерации в дискретное скрытое пространство, разработал общую структура метода, реализовал первую из двух частей модели, автоэнкодер кадров, а также внес значительный вклад в проведение вычислительных экспериментов.

Доклады на конференциях и семинарах:

1. доклад «Making DensePose Fast and Light» на конференции *WACV*, Online, 2021;
2. доклад «Latent Video Transformer» на конференции *VISIGRAPP*, Online, 2021;
3. доклад «NPBG++: Accelerating Neural Point-Based Graphics» на конференции Fall into ML 2022, Москва, Россия;
4. доклад «Multi-Sensor Large-Scale Dataset for Multi-View 3D Reconstruction» на конференции Fall into ML 2023, Москва, Россия.

Автор также внес вклад в следующие публикации

1. Войнов, О., Бобровских, Г., Карпышев, П., Галочкин, С., Арделян, А. Т., Боженко, А., Галочкин, С., Карманова, Е., Копанев, П., Лабутин-Рымшо, Я., *Рахимов, Р.*, Сафин, А., Серпива, В., Артемов, А., Бурнаев, Е., Тетерюков, Д., & Зорин, Д. (2023). *Multi-Sensor Large-Scale Dataset for Multi-View 3D Reconstruction.* Представлена на конференции IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21392-21403). CVPR 2023. CORE A*. Проиндексирована в SCOPUS, Web of Science. https://doi.org/10.1109/cvpr52729.2023.02049

3. Содержание работ

Тема диссертации раскрывается в следующих разделах, в каждом разделе излагается соответствующая статья.

В разделе 3.1 исследуется возможность моделирования видео в дискретном латентном пространстве. Описываются текущие методы, архитектура предложенной модели и проводится сравнительный анализ.

В разделе 3.2 представлен новый регрессионный метод локализации особых кривых 3D объектов. Описывается архитектура предложенного метода и проводятся сравнения с существующими методами.

В разделе 3.3 предлагается новая модель для генерации новых видов по набору входных изображений. Проводится сравнительный анализ с существующими методами как в смысле качества, так и в смысле скорости моделирования и рендеринга сцены.

В разделе 3.4 описывается модель для получения плотной оценки позы человека с улучшенными показателями скорости работы, которые позволяют запускать модель на мобильном устройстве. Проведен обзор влияния различных компонент модели на качество результата и скорость работы.

В разделе 3.5 представлена модель для трехмерного моделирования головы человека по одному изображению. Проводится сравнение эффективности предложенного подхода с существующими методами. Описано применение модели для работы с изображениями с неизвестными параметрами камеры.

3.1. Моделирование видео в латентном пространстве на основе архитектуры трансформера

Рассматривается задача генерации видео, а именно, задача предсказания будущих кадров видео на основе нескольких заданных входных кадров. Эта задача находит практическое применение в различных областях, таких как технологии беспилотных автомобилей, обнаружение аномалий в видео данных, создание тайм-лапсов [60] и создание анимированных пейзажей [22], где точные прогнозы будущих кадров видео критически важны для принятия решений и создания контента.

Несмотря на последние достижения в области генеративного обучения, которые облегчили создание реалистичных объектов высокого качества, включая изображения, текст и речь, генерация видео остается серьезной проблемой. Нейронные сети, даже для генерации кратковременных видео, состоящих из 16 кадров с низким разрешением, требуют значительной вычислительной нагрузки, например 512 тензорных процессоров (ТПУ) [52] для параллельного обучения. Несмотря на объем вычислительных ресурсов, качество получаемого видео остается низким.

Для решения этой проблемы предлагается модель на основе архитектуры трансформер, которая использует авторегрессивную генерацию в дискретном скрытом (латентном) пространстве [84]. Предлагаемый подход позволяет значительно снизить вычислительные затраты на обучение модели, при этом сохраняется качество генерируемых видео. За счет использования комбинации дискретного скрытого представления и рекуррентного подхода к генерации видео удается не только преодолеть ограничения по используюемой памяти ГПУ, но и существенно ускорить генерацию.

Описание модели

Модель предсказывает последующие кадры на основе начального набора заданных кадров. Видеопоследовательность, обозначенная как X, представляется в виде последовательности T кадров $x_{t=1}^{T}$, где каждый кадр $x_t \in \mathbb{R}^{H \times W \times 3}$ имеет размеры H и W с 3 RGB-каналами. Цель - сгенерировать оставшиеся кадры $(T - T_0)$ на основе первых T_0 кадров. Модель состоит из двух основных компонент: автокодировщика (автоэнкодера) кадров и авторегрессионной генеративной модели.

Для автокодировщика кадров используется архитектуру VQ-VAE [84], которая является вариационным автокодировщиком с дискретным скрытым пространством. Модель VQ-VAE, изображенная на рисунке 1, разработана для кодирования входного изображения $x \in \mathbb{R}^{H \times W \times 3}$ с использованием кодовой книги $e \in \mathbb{R}^{K \times D}$. Здесь K представляет собой размер кодовой книги, указывающий на категориальную природу скрытого пространства, а D обозначает размерность представления в кодовой книге.

В широком смысле VQ-VAE состоит из *кодировщика*, который сжимает изображение в более компактное представление, $z_e(x) \in \mathbb{R}^{h \times w \times D}$; *внутренней части*, которая дискретизирует

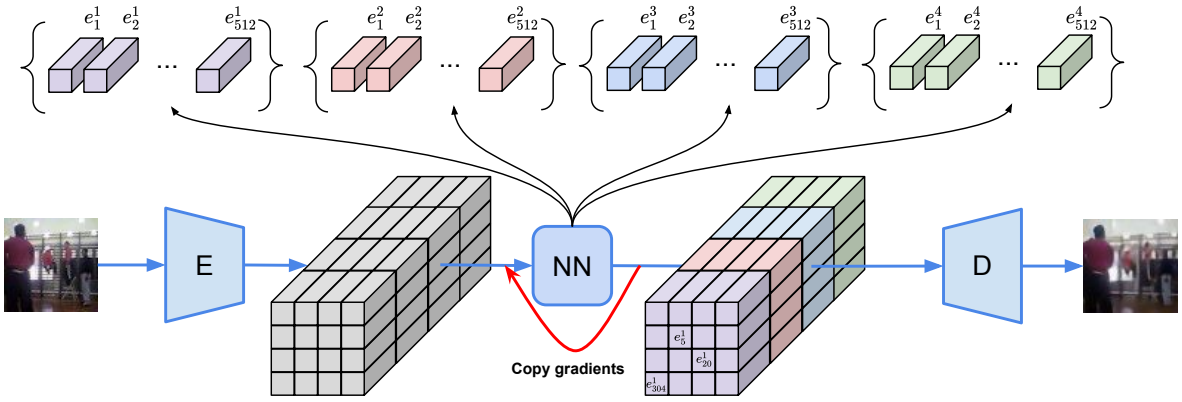

Рис. 1: *Архитектура автокодировщика кадров.* Кодировщик делит входное изображение на $n_c = 4$ части вдоль канального измерения. Пиксели в каждом сегменте затем сопоставляются с ближайшими представлениями из кодовой книги, которые использует декодер в качестве входных данных.

каждый пиксель, сопоставляя его с ближайшим представлением e_i из кодовой книги, чтобы $z(x) \in [K]^{h \times w \times 1}$; и *декодера*, который принимает на вход дискретные скрытые коды, $z(x)$, и отображает их в соответствующие представления, декодируя результат $z_q(x) \in \mathbb{R}^{h \times w \times D}$ обратно в пространство входных пикселей.

Функция потерь для VQ-VAE включает ошибку реконструкции и регуляризацию, выраженные следующим образом:

$$L = \|x - \textit{decoder}(z_q(x))\|^2 + \|z_e(x) - \text{sg}[e]\|^2. \tag{1}$$

Здесь sg[] представляет оператор остановки градиента, который выводит свой аргумент во время прямого прохода нейронной сети и выдает нулевые градиенты во время обратного прохода. Для обновления представлений внутри кодовой книги в ходе обучения используются обновления экспоненциального скользящего среднего (EMA).

Кодировщик кадров преобразует начальные T_0 кадров в дискретное представление, обозначенное как $Z_0 \in [K]^{T_0 \times h \times w \times n_c}$. Затем используется авторегрессионная модель для генерации новых кадров, всего $T - T_0$, на основе начальных Z_0. В качестве такой модели выбрана авторегрессионная генеративная модель [92], адаптированная для работы в скрытом дискретном пространстве вместо оригинального пиксельного пространства. Архитектура видео-трансформера подробно описана в оригинальной статье [92].

Модель принимает тензор $Z \in [K]^{T \times h \times w \times n_c}$ на входе и начинает процесс генерации, инициируя его первыми T_0 скрытыми кадрами, т.е. $Z_{:T_0,:,:,:} = Z_0$. Оставшиеся скрытые кадры могут быть случайно инициализированы, так как процесс генерации зависит исключительно от ранее сгенерированных или инициированных пикселей. Модель использует концепцию подмасштабирования [54], генерируя скрытое видео в виде последовательности неперекрывающихся срезов. С использованием фактора подмасштабирования $\mathbf{s} = (s_t, s_h, s_w)$ скрытое видео

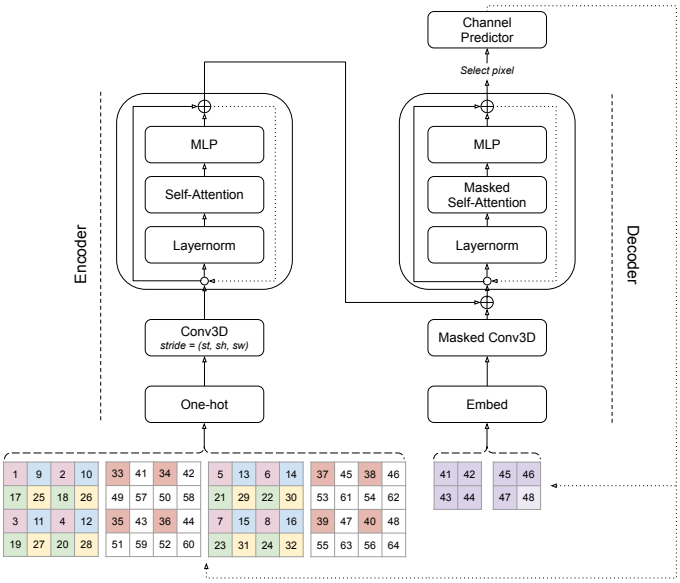

Рис. 2: *Архитектура модели.* Числа обозначают порядок генерации, цветные пиксели представляют собой сгенерированные пиксели, белые пиксели - нулевое заполнение, и пиксели одного цвета принадлежат одному срезу. Пример: генерация последнего пикселя среза $Z_{(1,0,1)}$ для скрытого видео размером $(t, h, w) = (4, 4, 4)$ с факторами подмасштабирования $(s_t, s_h, s_w) = (2, 2, 2)$.

разбивается на $s = s_t s_h s_w$ срезов, каждый размером $T/s_t \times h/s_h \times w/s_w$. Процесс генерации разворачивается последовательно, срез за срезом, пиксель за пикселем внутри среза и канал за каналом для каждого пикселя:

$$p(Z) = \prod_{i=0}^{Thw-1} \prod_{k=0}^{n_c-1} p\left(Z^k_{\pi(i)} | Z_{\pi(<i)}, Z^{<k}_{\pi(i)}\right), \tag{2}$$

где $p(Z)$ представляет собой вероятностное распределение по скрытой видео последовательности Z. Пиксели в каждом срезе $Z_{(a,b,c)}$ генерируются в порядке обхода по растровой развертке, в то время как срезы генерируются в порядке подмасштабирования: $Z_{(0,0,0)}$, $Z_{(0,0,1)}$, ..., $Z_{(s_t-1,s_h-1,s_w-1)}$.

Модель трансформера включает в себя кодировщик и декодер. Чтобы сгенерировать новое значение пикселя внутри среза $Z_{(a,b,c)}$, кодировщик сначала создает представление уже сгенерированных срезов $Z_{<(a,b,c)}$. Затем это представление смешивается с представлением уже сгенерированных пикселей внутри текущего среза $Z_{(a,b,c)}$. Авторегрессивный порядок поддерживается через внутренние отступы внутри кодировщика и маскирование в свертках и внимании внутри декодера. После генерации нового значения пикселя соответствующий внутренний отступ заменяется сгенерированным выводом, и процесс генерации повторяется. Процесс генерации, в случае пространственно-временного ($s_t > 0, s_h > 0, s_w > 0$) подмасштабирования, проиллюстрирован на рисунке 2.

Как только процесс генерации завершен, декодер скрытых кадров принимает $Z \in [K]^{T \times h \times w \times n_c}$ на вход (где все значения теперь валидные), отображает его на ранее изученные

представления $Z_q \in \mathbb{R}^{T \times h \times w \times D}$ и декодирует его обратно кадр за кадром в исходное пиксельное пространство $X \in \mathbb{R}^{T \times H \times W \times 3}$.

Эмпирические результаты

Оценивание точности предсказания будущих кадров делается с помощью Fréchet Video Distance (FVD) [30]; кроме того, используется метрика - количество бит на размерность (bits/dim), представляющая собой отрицательный логарифм вероятности, усредненный по всем сгенерированным (скрытым) пикселям и каналам. Также в сравнениях представлено базовое решение: берется последний настоящий кадр и используется в качестве прогноза для всех будущих кадров.

Таблица 1: *Количественные сравнения.* Используется метод сравнения [13, 92], когда обучается видеогенератор с одним кадром на вход и метрики считаются на видео из 16 кадров. FVD и bits/dim вычисляются с пятью и одним инициализирующими кадрами соответственно.

(a) Данные BAIR Robot Pushing

Метод	**bits/dim(↓)**	**FVD(↓)**
Baseline	-	320.90
VideoFlow [44]	1.87	-
SVP-FP [18]	-	315.5
CDNA [23]	-	296.5
LVT (ours, $n_c = 1$)	1.25	275.71 ± 5.41
SV2P [19]	-	262.5
LVT (ours, $n_c = 4$)	1.53	125.8 ± 2.9
SAVP [46]	-	116.4
DVD-GAN-FP [13]	-	109.8
TriVD-GAN-FP [52]	-	103.3
Axial Transformer [32]	1.29	-
Video Transformer [92]	1.35	94 ± 2

(b) Данные Kinetics-600

Метод	**bits/dim(↓)**	**FVD(↓)**
Baseline	-	271.00
LVT (ours)	2.14	224.73
Video Transformer [92]	1.19	170 ± 5
DVD-GAN-FP [13]	-	69.15 ± 1.16
TriVD-GAN-FP [52]	-	25.74 ± 0.66

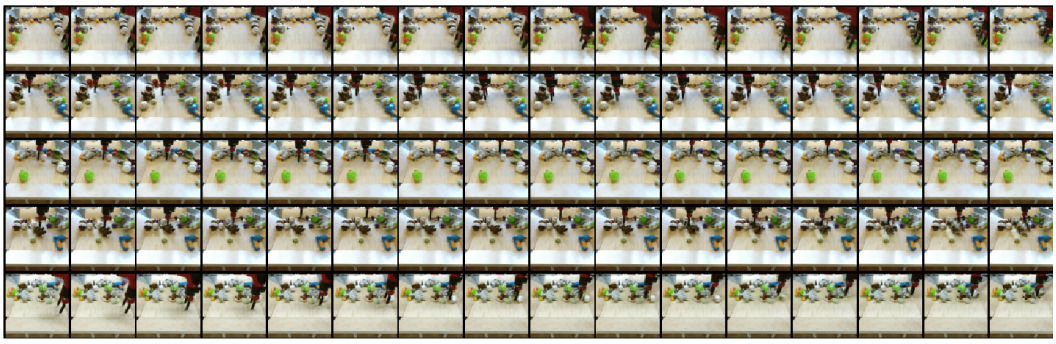

Рис. 3: *Результаты на наборе данных BAIR Robot Pushing.* Каждая строка изображает отдельное видео, показывая первые пять кадров в виде настоящих и последующие кадры в виде сгенерированных.

На двух наборах данных, BAIR Robot Pushing [20] и Kinetics 600 [10] представлены как количественные результаты (Таблицы 1a, 1b), так и качественные результаты (Рисунок 3). Хотя качество полученных результатов похоже на качество работы других методов на наборе данных BAIR Robot Pushing, на наборе данных Kinetics-600 наблюдается низкое качество работы, что потенциально связано с накоплением ошибок внутри модели с архитектурой трансформера по причине повышенной сложности и разнообразия набора данных.

Заключение

Разработана модель генерации видео, основанная на идее моделирования видео в дискретном латентном пространстве. Представленная модель обладает хорошей обобщающей способностью, то есть может генерировать последовательности видео по невиданным ранее входным условным кадрам. Более того, это достигается при использовании ограниченного вычислительного ресурса на этапе обучения, состоящего из 8 графических процессоров V100, в то время как альтернативные методы требуют до 512 тензорных процессоров для обучения.

3.2. Оценка геометрических особенностей 3D объектов на основе глубоких нейросетевых моделей

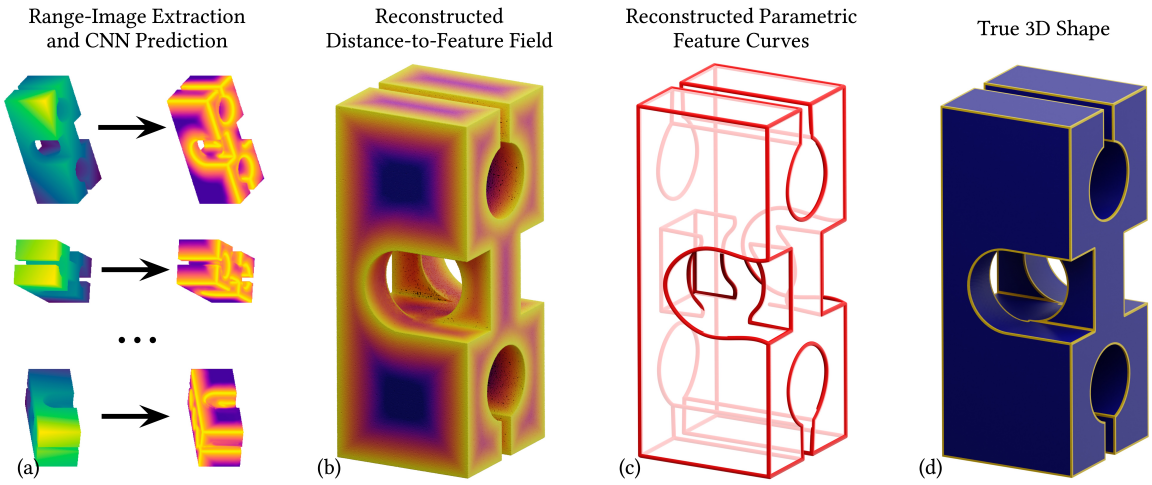

Рис. 4: *Обзор модели DEF.* (a) Разработана эффективная модель для оценки расстояния до особых кривых на картах глубины; (b) Модель объединяет индивидуальные прогнозы с разных ракурсов с шага (a) в единое предсказание для целой 3D модели; (c) Далее извлекаются параметрические особые кривые; (d) Это приводит к точным реконструкциям как прямых, так и кривых особых линий, аккуратно совмещенным с CAD-моделями.

Рассматривается задача локализации геометрических особенностей в 3D объектах (поверхностых кривых и линий, вдоль которых поле нормалей испытывает излом). Решение данной задачи имеет ключевое значение для задачи реконструкции высокоточных 3D CAD моделей без использования ручного аннотирования и настройки модели.

Существующие методы обнаружения геометрических особенностей включают локальные методы оценки [90, 16], которые сосредоточены на вычислении дифференциальных свойств в небольших областях, но требуют тщательной настройки параметров для каждого конкретного случая. Методы сегментации поверхности [50] направлены на определение однородных участков поверхности и классификацию их границ как особенностей, но они неэффективны для неполных моделей. Стратегии подгонки участков [9] включают подгонку предопределенных примитивов к большим областям поверхности объекта, что позволяет обеспечить устойчивость к шуму в данных, но при этом страдает вычислительная эффективность и гибкость из-за зависимости от заранее заданных примитивов. Тем временем наблюдается рост числа различных типов моделей, основанных на данных, особенно моделей машинного обучения, которые позволяют классифицировать точки вдоль особых кривых и линий [27, 31]. Однако, эти подходы сталкиваются с проблемами масштабируемости и устойчивости в условиях зашумленных данных.

Предлагается новый метод, DEF, для регрессии расстояния до особой кривой на локальных участках, отличающийся от стандартной бинарной классификации. Этот метод масштабируем, адаптивен и имеет хорошую производительность при обнаружении особых геометрических кривых, тем самым преодолевая критические ограничения в текущих подходах.

Описание модели

Модель обрабатывает изображения глубины, полученные из реальных сканов или из синтетических наборов, в качестве входных данных для заданного объекта. Модель выдает усеченное значение расстояния до особой кривой для каждой входной точки на поверхности объекта, что продемонстрировано на рисунке 4. Метод включает четыре основных компонента.

Первый компонент, *генерация обучающих данных*, включает создание наборов данных DEF-Sim (синтетических) и DEF-Scan (реальных) для последующего обучения моделей. DEF-Sim, основанный на наборе данных ABC [42], использует граничное представление и аннотации особых кривых для обучения, рассчитывая расстояние до особой кривой для каждой точки p как $d^{\varepsilon}(p) = \min(\|q(p) - p\|_2, \varepsilon)$, где $q(p)$ - точка на ближайшей особой кривой или ребре, а ε - радиус усечения. DEF-Scan включает себя сканы напечатанных объектов, совмещённые с исходными 3D CAD моделями. Сканирование производилось с использованием 3D-сканера структурированного света. Эти наборы данных включают разнообразные условия обучения по разрешению, уровню шума и размерам выборки, которые критичны для разработки точных предиктивных моделей для обнаружения особых кривых.

Второй компонент, *патч-ориентированные глубокие оценщики* (DEF), позволяет оценить расстояния до особой кривой по картам глубины. Сначала обученные на синтетических наборах данных и дополнительно дообученные на реальных данных, эти модели обучаются в процессе минимизации $\min_{\theta} \frac{1}{N} \sum_i^N L(d_i, f(P_i; \theta))$, где d_i — реальное расстояние до особой кривой для патча P_i, $f(\cdot; \theta)$ — модель с параметрами θ, а L — функция потерь. CNN, особенно модель U-Net с

архитектурой ResNet-152, оказались наиболее эффективными в качестве $f(\cdot;\theta)$. Гистограммная функция потерь [36] значительно улучшила качество регрессии, сфокусировав сеть на более узком диапазоне целевых расстояний. Производительность сети стабилизируется при объеме наборов данных более 64000 экземпляров, и DEF может обнаруживать особенности при обучении на выборках разного объема, что указывает на адаптивность модели.

Третий компонент - это *оценка на полных 3D моделях*. Представлен новый подход для решения этой задачи, работающий на основе объединения прогнозов расстояний до особых кривых на уровне патчей с использованием глубоких оценщиков. Этот процесс сначала включает преобразование входной 3D-модели в набор изображений глубины, $I_{i=1}^{n_v}$, по разным направлениям. Каждый патч изображения I_i обрабатывается независимо нейронной сетью, выдающей прогнозы расстояний до особых кривых. Суть предложенного подхода заключается в передаче этих прогнозов между патчами. Для заданной пары $(s,\, t)$ исходного и целевого видов, и с оценкой расстояния до особенности d_s, доступной в исходном виде, используется механизм синтеза видов на основе деформаций, чтобы получить сдеформированный прогноз $\widehat{d}_t^{s\to t}$ для каждого пикселя в целевом виде путем повторного проецирования прогнозов изображения исходного вида. Завершающим шагом является получение согласованной глобальной оценки расстояния, вычисленной как минимум среди деформированных оценок из разных исходных видов $\widehat{d}_t = \min_s \widehat{d}_t^{s\to t}$. Этот метод эффективно интегрирует информацию, чувствительную к особенностям, по всей 3D-модели, что подтверждается результатами различных эмпирических исследований.

В конце следует этап *извлечения параметрических особых кривых* из облаков точек, объединивший в себе алгоритмы для обнаружения углов, анализ структуры графа и подгонку сплайнов. Это включает классификацию и сегментацию локальных точек, создание графа кривых, подгонку и оптимизацию сплайнов, а также применение постобработки, которая производит фильтрацию примитивов на основе длины кривой с учетом значений метрики качества.

Эмпирические результаты

Эффективность предложенного метода обнаружения особенностей оценивается с помощью нескольких метрик качества, включая среднеквадратичное отклонение (RMSE), полноту, долю ложных срабатываний при различных порогах для оценки качества регрессии расстояния до особенностей и точности оценки особых кривых в 3D-моделях.

Метод DEF сравнивался с пятью ведущими методами извлечения особых кривых из 3D моделей, охватывающими как традиционные подходы, так и методы на основе глубокого обучения. VCM [55], необучаемый подход, использует меры ковариации Вороного. Sharpness Fields (ShF) [65] применяет CNN для предсказания поля. EC-Net [101] использует сеть на основе PointNet++ [64] для обнаружения особых линий, а PIE-NET [88] использует двухэтапный процесс для сегментации особенностей и генерации предложений параметрических кривых.

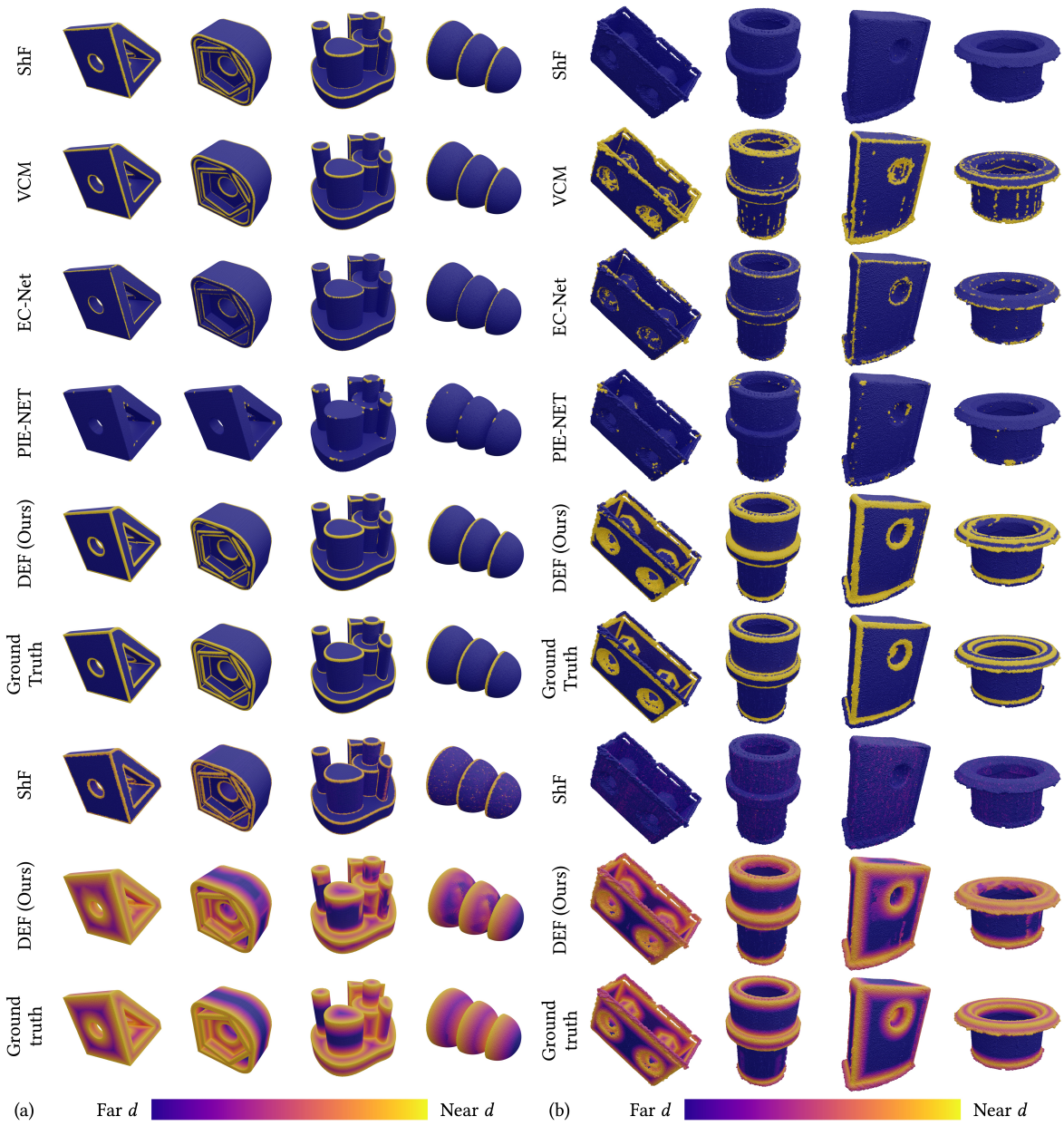

Рис. 5: *Качественная оценка обнаружения особых кривых.* Сравнение с передовыми методами на (a) наборах данных высокого разрешения целых 3D моделей и (b) реальных наборах данных, представляющих полные 3D модели, полученные за счет сканирования реальных объектов. Предложенный метод надёжно реконструирует поле расстояний от точки до особой кривой и масштабируется до 3D моделей, представленных миллионами точек.

В обширных сравнениях на синтетических и реальных наборах данных DEF стабильно превосходит своих конкурентов с точки зрения полноты и доли ложных срабатываний. В сравнении на основе патчей (Таблица 2) DEF превосходит конкурентов, таких как ShF, VCM и PIE-NET. Оценка на полных 3D-моделях и сканированных реальных формах (Рисунок 5) дополнительно подчеркивает превосходство DEF, демонстрируя его способность надежно регрессировать поля расстояний до особенностей и превосходя конкурентов по точности и обобщаемости. В

Таблица 2: *Количественная оценка обнаружения особых линий.* Предложенный метод, предназначенный для *оценки расстояния до особой кривой* и *сегментации особенностей*, превосходит конкурирующие методы по нескольким метрикам оценивания качества сегментации и регрессии (оценивание с использованием синтетических патчей изображений по выборке DEF-Sim). Для получения результатов сегментации DEF применяется порог, равный 0.02, к предсказанному расстоянию.

Метод	RMSE$\times 10^{-3}\downarrow$	RMSE-$q_{95}\times 10^{-3}\downarrow$	Recall %$\uparrow$	FPR %$\downarrow$
VCM [55]	---	---	49.1	3.1
EC-Net [101]	---	---	79.2	2.9
DEF *(Обученный на данных EC-Net)*	124.1	501.1	56.0	0.15
PIE-NET [88]	---	---	32.0	3.8
DEF *(Обученный на данных PIE-NET)*	86.2	451.8	57.1	0.1
ShF [65]	18.0	95.7	**80.9**	0.3
DEF (Ours)	**11.1**	**42.5**	80.02	**0.02**

целом результаты подтверждают статус DEF как мощного и универсального фреймворка для обнаружения геометрических особенностей в 3D моделях.

Заключение

Предложен новый подход, DEF, для предсказания геометрических особенностей в 3D моделях. Традиционные методы опираются на подгонку примитивов или оценку ковариационной меры Вороного, что является времязатратным и не всегда обеспечивает качественные результаты. В отличие от них, DEF, во-первых, работает на основе обучения с использованием больших синтетических наборов данных и минимального количества реальных данных, а во-вторых, обучается регрессии поля расстояний до особенностей на *локальных* участках. Благодаря этим двум факторам, DEF демонстрирует отличную обобщаемость и масштабируемость на новых, ранее невиданных 3D формах различных размеров и вариативности, даже при наличии шумов сканирования.

3.3. Ускорение нейронной графики на основе облаков точек

Рассматривается задача генерации новых видов, которая включает создание фотореалистичных видов сцены с новых точек обзора по ограниченному набору изображений этой сцены. Эффективное решение этой задачи важно в таких приложениях, как виртуальная и дополненная реальность, кинематография и игровая индустрия, где оно позволяет моделировать реалистичные среды по данным разреженных измерений.

Традиционные методы рендеринга, такие как интерполяция видов и визуализация светового поля [47, 76], до недавнего времени являлись основополагающими, но зачастую в сложных сценариях с детализированой геометрией и разнообразным освещением результаты работы этих методов неудовлетворительные. Появление методов на основе полей нейронного излучения (NeRF) [56] позволили достигнуть более эффективных результатов за счет моделирования целых сцен с помощью нейронных сетей, оптимизированных через дифференцируемую объемную визуализацию. Однако, подход NeRF требует значительных вычислительных ресурсов и большого числа входных кадров с разных точек обзора. Также были предложены воксельные методы [38], позволяющие получать структурированное представление 3D-сцен, но при этом результаты работы этих методов имеют не достаточно высокое разрешение, а сами методы - требуют значительных вычислительных затрат. Ещё один значимый подход — нейронная графика на основе облаков точек (NPBG) [1], которая использует облака точек для моделирования геометрии сцены и показывает хорошие результаты в качестве визуализации. Тем не менее, эти методы требуют времязатратной оптимизации для каждой сцены.

Предлагаемая модель, NPBG++, значительно улучшает оригинальный подход NPBG. Путём прогнозирования нейронных дескрипторов напрямую из исходных изображений за один проход, предлагаемый метод упрощает процесс, устраняя необходимость в трудоёмкой оптимизации для каждой сцены. Это развитие не только значительно сокращает время моделирования сцены, но и повышает качество визуализации. NPBG++ значительно улучшает эффективность и обобщаемость решения задачи генерации новых видов, позволяя производить визуализацию высокого качества в реальном времени для разнообразных сцен.

Описание модели

Предложенный метод генерирует изображения с новых ракурсов статичной сцены с использованием набора многовидовых входных изображений, соответствующих параметров камеры и облака точек. В отличие от NPBG [1], который оптимизирует нейронные дескрипторы для каждой новой сцены, предложенный подход напрямую предсказывает дескриптор из входных кадров. Эти нейронные дескрипторы описывают локальные геометрические и фотометрические свойства поверхности сцены.

Рассматриваемый метод, в отличие от методов на основе интерполяции изображений, которые зависят от определения ближайших видов из набора входных изображений для генерации нового вида, строит единую модель сцены. Это достигается путём обработки входных

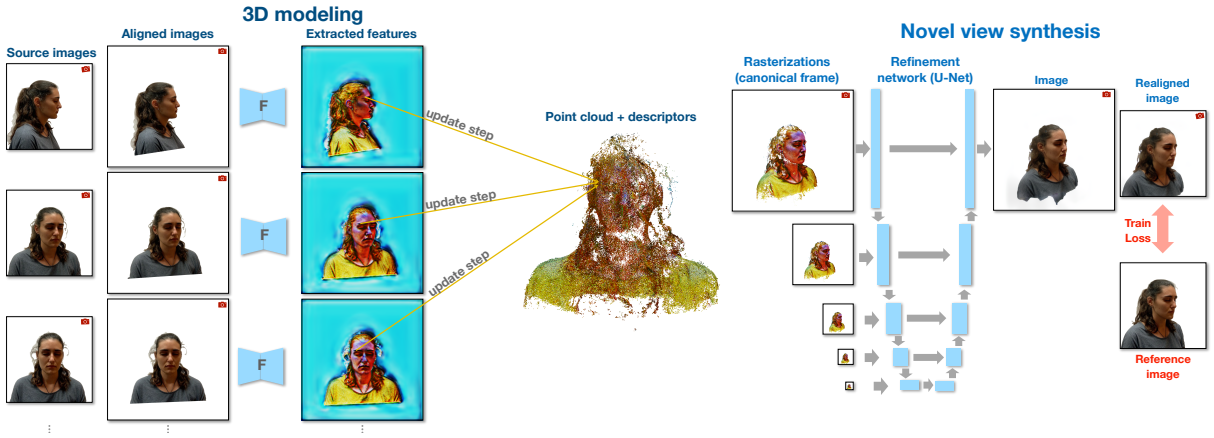

Рис. 6: *Обзор NPBG++.* Сцена представлена в виде облака точек, каждая точка содержит встроенный видозависимый нейронный дескриптор. На этапе 3D-моделирования последовательно обрабатывается каждый входной кадр - выравнивается изображение, извлекаются дескрипторы и выполняется онлайн-агрегация для обновления нейронных дескрипторов, не требующая оптимизации. Для генерации нового вида происходит растеризация дескрипторов облака точек и результат передается на вход нейронной сети-уточнителю, за которой следует постобработка в виде выравнивания.

видов в онлайн-режиме, итеративно обновляя промежуточное состояние дескрипторов облака точек сцены. Потребление памяти при этом не зависит от количества входных кадров. После обработки всех видов вычисляются финальные дескрипторы на основе промежуточных состояний, описанных далее. Система включает в себя два основных этапа: этап моделирования, на котором извлекаются дескрипторы точек путём обработки входных видов, и процесс визуализации, в котором дескрипторы растеризуются и преобразуются в окончательные изображения с использованием сети-уточнителя (см. рисунок 6).

На первом *этапе моделирования* происходит процесс извлечения дескрипторов, где сеть на основе U-Net [72] создаёт плотную карту признаков для каждого пикселя входного изображения той же высоты и ширины, что и входное изображение. Затем облако точек сцены проецируются на карту признаков и билинейно семплируются дескрипторы для каждой точки. Здесь следует отметить два важных момента. Во-первых, входные изображения подвергаются процессу выравнивания для обеспечения согласованности дескрипторов из разных видов. Это выравнивание происходит до извлечения дескрипторов путём поворота входного изображения в каноническую ориентацию, где проекция вертикальной оси мира на плоскость изображения вертикальна (см. рисунок 6 слева). Данное выравнивание критически важно, так как сеть для извлечения дескрипторов не является инвариантной к вращению. Во-вторых, чтобы предотвратить обновление дескрипторов для точек, необозримых в текущем ракурсе, оценивается видимость точек. Для этого создается Z-буфер для каждого пикселя и облако точек растеризуется на уменьшенный размер изображения. Точка в буфере отмечается видимой, если у нее минимальное значение Z (наиболее близкая к камере).

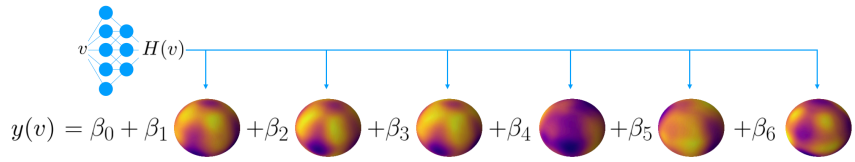

Рис. 7: *Зависящий от направления обзора (видозависимый) нейронный дескриптор.* Дескриптор y: $\mathbb{R}^3 \to \mathbb{R}^c$ моделируется как линейная комбинация обучаемых базисных функций над сферой (H: $\mathbb{R}^3 \to \mathbb{R}^m$), определённая коэффициентами $\beta_i \in \mathbb{R}^c$ (см. уравнение 3). Для каждой новой сцены, используя набор исходных изображений, β_i определяется для каждой точки.

На следующей стадии агрегации решается задача обработки дескрипторов из различных входных видов в онлайн-режиме. На практике требуется, чтобы используемая память не зависела от количества видов, а результат - от порядка обработки входных кадров. По этой причине не используются сети на основе трансформеров [85], LSTM [33] и GRU [12]. Вместо этого выбран метод, инвариантный к перестановкам наблюдений, который добавляет зависимые от точки обзора эффекты в нейронный дескриптор каждой точки. Этот дескриптор моделируется $y : \mathbb{R}^3 \to \mathbb{R}^c$ как линейная комбинация обучаемых базисных функций над сферой (см. рисунок 7):

$$\underset{1\times c}{y(v)} = \underset{1\times m}{H(v)}\ \underset{m\times c}{\beta}\ + \underset{1\times c}{\beta_0}\,, \tag{3}$$

где v представляет направление обзора единичной длины, $H : \mathbb{R}^3 \to \mathbb{R}^m$ является набором из m базисных функций (используется $m = 6$), и β и β_0 — это коэффициенты, которые необходимо определить для каждой точки. Этот подход похож на NEX [94], где моделируют зависимые от точки обзора RGB значения вместо нейронных дескрипторов. В отличие от NEX, решается N задач многомерной линейной регрессии, чтобы найти коэффициенты β и β_0 для всех N точек. Для каждой точки имеется набор пар $\{(v_k, y_k)\}_{k=1}^K$, где K — количество входных кадров (видов), в которых точка оценивается видимой. v_k — это направление обзора единичной длины, а y_k — семплированный дескриптор из входного изображения. Исходя из этого параметры дескриптора находятся следующим образом:

$$\underset{1\times c}{\beta_0} = \frac{1}{K}\sum_{k=1}^{K} \underset{1\times c}{y_k}\,, \tag{4}$$

$$\underset{m\times c}{R} := \frac{1}{K}\sum_{k=1}^{K} \underbrace{H(v_k)^T y_k}_{m\times c} - \frac{1}{K}\sum_{k=1}^{K} \underbrace{H(v_k)^T \beta_0}_{m\times c},$$

$$\underset{m\times c}{\beta} = \left(\frac{1}{K}\sum_{k=1}^{K} \underbrace{H(v_k)^T H(v_k)}_{m\times m} + \frac{\alpha}{K}\ \underset{m\times m}{I_m}\right)^{-1} \underset{m\times c}{R}\,, \tag{5}$$

где I_m — единичная матрица, β_0 описывает средний дескриптор, а коэффициент регуляризации берется равным α=1. Когда приходит новый дескриптор y_k, обновляются пять промежуточных состояний: K, $\sum_{k=1}^K y_k$, $\sum_{k=1}^K H(v_k)^T y_k$, $\sum_{k=1}^K H(v_k)^T$, $\sum_{k=1}^K H(v_k)^T H(v_k)$. Важно

отметить, что финальный размер этих промежуточных состояний не зависит от количества входных видов K. Для каждой отдельной точки продолжают обновляться эти состояния, пока не обработаются все входные кадры. После этого рассчитываются значения для β и β_0. Также удаляются точки из облака точек, которые не были видны ни в одном из входных кадров.

На втором этапе, *этапе генерации нового вида*, происходят три отдельных шага для генерации окончательного изображения на основе заданных параметров камеры. Сначала растеризуются финальные дескрипторы точек, используя подход, аналогичный NPBG [1]. Затем сеть-уточнитель, использующая архитектуру U-Net [100], принимает на вход результат растеризации и выдает на выходе RGB изображение, помогая устранить такие проблемы как просвечивание поверхности. Наконец, на последнем шаге происходит процесс выравнивания полученного изображения. Изначально полученное в канонической ориентации, описанной выше, изображение вращается для выравнивания по отношение к запрашиваемой ориентации камеры, как показано на рисунке 6-право. Этот шаг обеспечивает консистентный процесс визуализации, независимо от ориентации вертикальной оси камеры. Данный фактор ранее упускался из виду в методах, использующих нейронные дескрипторы [1, 45, 93].

Функция потерь в процессе обучения сочетает в себе перцептивную функцию потерь VGG-19 [77], $\mathcal{L}_1$, функцию потерь для оценки разницы между уменьшенными выходными и целевыми изображениями для сохранения цвета и предотвращения сглаживания деталей, а также новую регуляризацию, которая сравнивает целевое изображение с визуализацией, полученной на основе дескрипторов из самого же целевого изображения.

Эмпирические результаты

В экспериментах эффективность предложенного метода оценивается с помощью стандартных метрик для оценки качества изображения: индекс структурного сходства (SSIM), пиковое отношение сигнал/шум (PSNR) и перцептивное сходство изображений (LPIPS) [102]. Предложенный метод сравнивается с несколькими передовыми алгоритмами нейронного рендеринга, включая NPBG [1], NeRF [56], SVS [71] и IBRNet [87]. Эти сравнения проиллюстрированы количественно в таблице 3 и качественно на рисунке 8. Было обнаружено, что предложенный метод может генерировать изображения лучше чем метод SVS и сопоставимые изображения по результатам с IBRNet, ведущим методом генерации новых видов с обобщаемостью на новые сцены. В случае дообучения предложенный метод превзошёл NPBG по всем наборам данных, получив лидирующие оценки на сценах из наборов данных DTU и H3DS, и сходные по качеству результаты на наборах данных ScanNet и NeRF-Synthetic.

Таблица 3: *Количественные оценки.* Для каждого набора данных вычисляются метрики на отложенных кадрах, усреднённые по отложенным сценам. Нижний индекс *ft* указывает на версии методов с дообучением. В случае NPBG++$_{ft}$ коэффициенты (β, β_0) и сеть-уточнитель дообучаются напрямую. В случае NPBG++$_{ft\text{-}system}$ дообучается извлекатель признаков, агрегатор (нейронные базисные функции) и сеть-уточнитель.

Метод	Оптимизация под сцену	Nerf-Synthetic			ScanNet			DTU			H3DS		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SVS[71]	✗	22.81	0.919	0.104	23.32	**0.771**	**0.445**	20.98	0.897	0.162	18.96	0.798	0.210
IBRNet[87]	✗	**29.47**	**0.955**	0.157	**23.34**	0.760	0.494	**25.81**	**0.924**	0.231	20.30	0.791	0.279
NPBG++ (наш метод)	✗	26.06	0.936	**0.071**	23.11	0.766	0.502	23.23	0.915	**0.154**	**21.80**	**0.818**	**0.177**
NPBG[1]	✓	28.62	0.946	0.058	25.09	0.737	0.459	26.00	0.913	0.125	24.68	0.827	0.146
NeRF[56]	✓	32.49	0.970	**0.041**	**25.74**	**0.780**	0.537	**26.92**	0.913	0.198	23.88	0.833	0.178
SVS$_{ft}$[71]	✓	23.37	0.919	0.101	22.31	0.610	0.543	20.72	0.864	0.190	20.12	0.770	0.197
IBRNet$_{ft}$[87]	✓	**32.51**	**0.972**	0.144	24.42	0.774	0.493	23.80	0.917	0.222	24.68	**0.850**	0.195
NPBG++$_{ft\text{-}system}$ (наш метод)	✓	26.24	0.940	0.064	23.48	0.768	0.490	24.05	0.919	0.147	23.79	0.836	0.155
NPBG++$_{ft}$ (наш метод)	✓	28.67	0.952	0.050	25.27	0.772	**0.448**	26.08	**0.928**	**0.123**	**24.91**	0.845	**0.137**

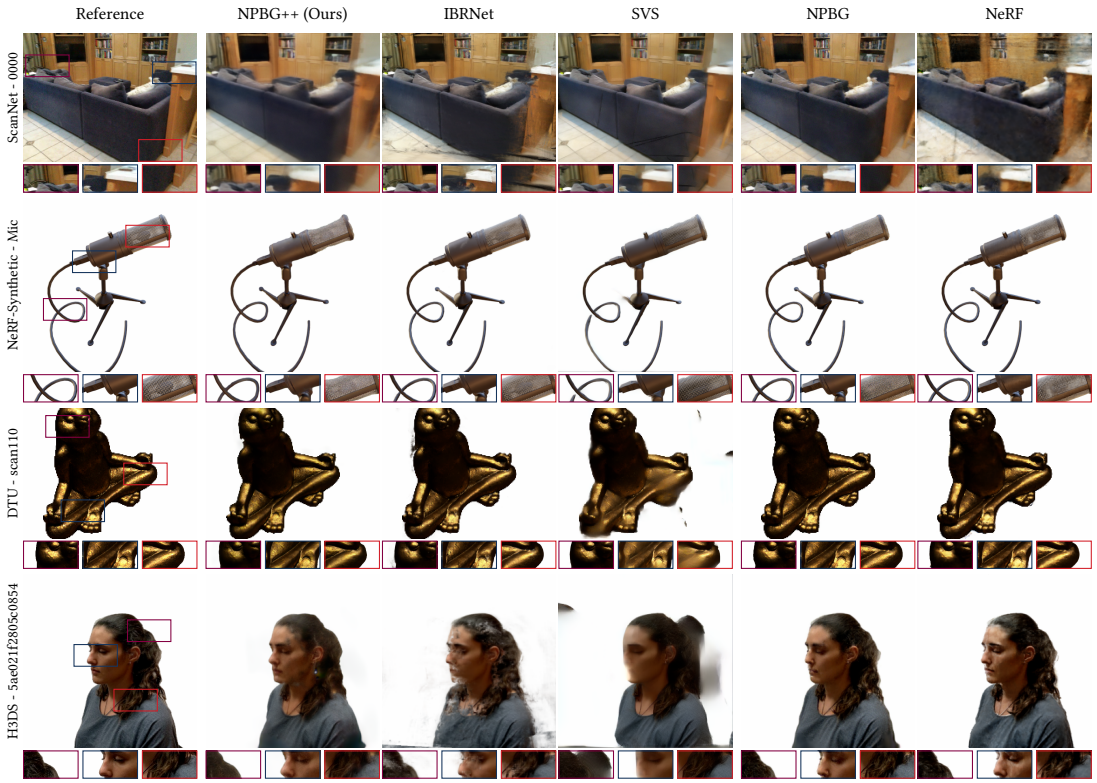

Рис. 8: *Качественные оценки.* Сравнения с оптимизационными подходами (NPBG[1], NeRF[56]) и обобщаемыми подходами (IBRNet[87], SVS[71]) на сценах ScanNet[15], NeRF-Synthetic[56], DTU[37], H3DS[69].

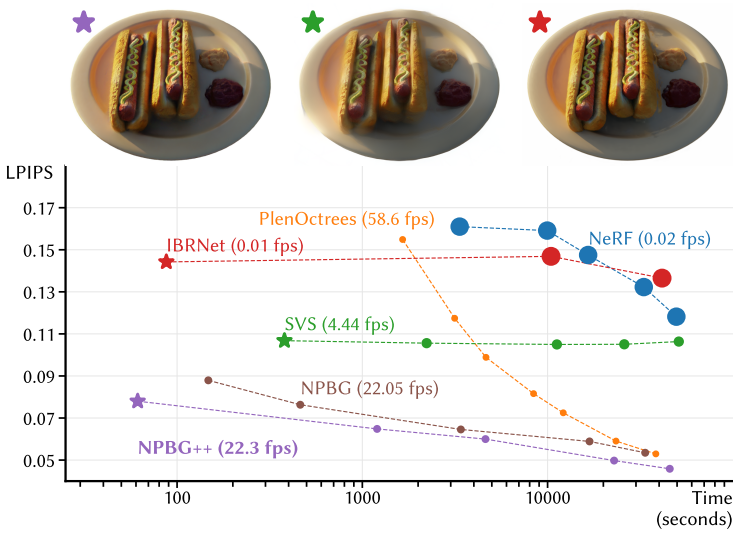

Рис. 9: *Время работы и качество изображения.* Сравнение нескольких методов, выполненное на сцене hotdog из синтетического набора данных NeRF. Ось времени представляет время до получения первой визуализации, т.е. время моделирования сцены + время визуализации одного изображения. Для методов, отмеченных звёздочкой ⋆, первые результаты представлены без дообучения на новой сцене. Время моделирования включает извлечение признаков для IBRNet, оценку геометрии + этап 3D моделирования для NPBG++ и оценку геометрии + создание полигональной сетки для SVS (визуализации сверху предлагают качественные сравнения между этими конфигурациями). Остальные результаты рассчитаны в разные моменты процесса дообучения. Площади кругов пропорциональны логарифмам времени рендеринга (меньше - лучше).

Был проведен анализ времени работы: проведено сравнение скорости нескольких передовых методов на двух этапах рендеринга, как показано на рисунке 9. Первый этап посвящен извлечению алгоритмами данных из исходных изображений. Это включает время на обучение нейронных представлений для некоторых методов, работу модуля для выделения признаков в составе IBRNet и 3D моделирование для предложенного автором работы подхода. Для методов SVS, NPBG и NPBG++ также учтено время получения 3D представления, требуемого на вход моделям. Время этих процессов учитывается в наших сравнениях. Этот этап происходит один раз при обработке заданной сцены. Второй этап включает непосредственно рендеринг новых видов. В частности, NeRF и IBRNet демонстрируют длительное время рендеринга. Время рендеринга IBRNet превышает весь процесс дообучения предложенного метода. В случае NeRF, PlenOctrees и NPBG требуется дополнительно обучать модель заново на сцене, что приводит к увеличенному времени получения качественных результатов. Предложенная в данной работе модель, без дообучения на новую сцену, имеет самое короткое общее время до получения первой визуализации, превосходя SVS и IBRNet, которые ограничены временем оценки поверхности и временем рендеринга соответственно.

Заключение

В заключение, предложенная модель NPBG++ значительно улучшает обобщение в задаче синтеза новых видов, что означает, что модель может хорошо работать на разнообразных, невиданных ранее сценах без оптимизации для каждой сцены. Предсказывая нейронные дескрипторы напрямую из исходных изображений за один проход, NPBG++ избегает трудоемкой оптимизации на новой сцене. Это новшество позволяет модели быстро адаптироваться к новым окружениям и позволяет быстро создавать высококачественные рендеринги и сохранять высокую скорость визуализации.

3.4. Ускорение и уменьшение размера модели для плотной оценки позы человека

Рассматривается задача построения плотной оценки позы человека (DensePose) [3], которая включает понимание формы и позы человека на изображениях посредством построения соответствия между пикселями изображения и точек на карте поверхности тела. Эта задача имеет важное значение для различных приложений компьютерного зрения, таких как дополненная реальность и виртуальная примерка одежды, где точное моделирование человеческого тела является необходимым. Задача DensePose включает предсказание UV-координат для каждого пикселя изображения человека, отображая их далее на 3D модель человека, например, модель SMPL (Skinned Multi-Person Linear) [51].

Существующие в этой области модели, такие как DensePose R-CNN [3] и Parsing R-CNN [98], имеют большое количество параметров, что делает их неподходящими для развертывания на мобильных устройствах. Использование такого рода моделей требует надежной серверной инфраструктуры и стабильного интернет-соединения, что ограничивает их практическое применение. Более того, последующие работы [62, 61] внесли вклад в улучшение качества результата, но ни в одной из этих работ не проводилась оптимизация размера и скорости модели для применения на мобильном устройстве. Эти ограничения подчеркивают необходимость сделать модель более доступной и широко используемой в реальных приложениях.

В ответ на эти вызовы предлагается новая архитектура Mobile Parsing R-CNN, которая разработана для того, чтобы быть одновременно легкой и эффективной, позволяя делать оценку DensePose в реальном времени на мобильных устройствах. Оригинальная архитектура модели DensePose R-CNN подверглась серьезным изменениям за счет добавления различных новых компонент и применения методов квантизации. Новая архитектура представляет собой значительный прогресс в рассматриваемой прикладной области, так как она позволила достигнуть сокращение размера модели в 17 раз и улучшение времени её работы в 2 раза по сравнению с базовой моделью. Эти результаты открывают новые возможности для развертывания передовых приложений на основе полученной модели без необходимости использования мощного аппаратного обеспечения и доступа в интернет.

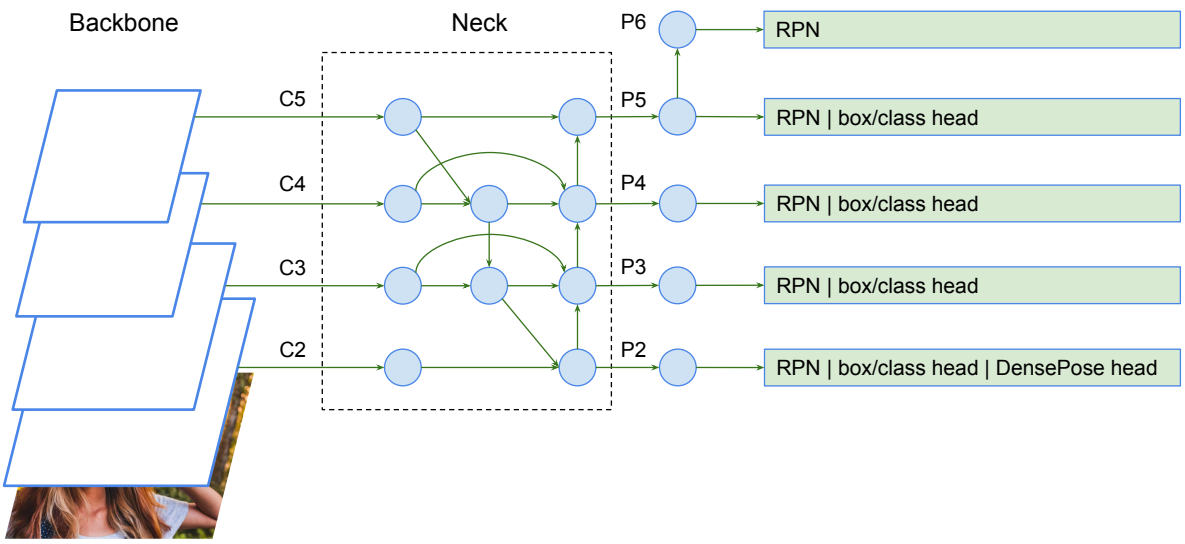

Рис. 10: *Высокоуровневая структура.* C_i, P_i представляют уровни признаков с разрешением $1/2^i$ входного изображения. Карта признаков P_6 получается путем пулинга карты признаков P_5 в два раза.

Описание модели

Архитектура модели Mobile Parsing R-CNN вдохновлена моделью Parsing R-CNN, которая победила в задаче оценки DensePose на конкурсе COCO 2018. Архитектура сети основана на двухэтапном R-CNN пайплайне обнаружения, который включает в себя базовую подсеть для извлечения признаков из входного изображения, «шею» для дальнейшего уточнения этих признаков, сеть для предложения кандидатов-регионов объектов (RPN), голову для классификации объектов и регрессии ограничивающих рамок, а также DensePose голову для финальной плотной оценки позы (см. рисунок 10).

Основной фактор при выборе базовой подсети - эффективная архитектура, подобная MobileNetV1 и V2 [35, 73], которая характеризуется использованием раздельных сверток. Выбор базовой подсети включает исследование использования различных архитектур, таких как MobileNetV3 [34], которая объединяет блок сжатия и нелинейности; MixNet [81], которая предлагает вариант с несколькими ядрами сверток; моделей на основе дифференцируемого поиска нейронной архитектуры MnasNet [79], FBNet [95] и Single-Path [78]; класса моделей EfficientNets [80], позволяющих балансировать точность и размер нейросети; архитектура CondConv [97], которая содержит динамические веса ядер в сверточных слоях.

В качестве части сети, которая представляет собой «шею», была выбрана архитектура BiFPN [82] для слияния карт признаков разных масштабов, полученных на выходе базовой подсети. Выбранная архитектура показала превосходную производительность в задачах обнаружения объектов, оставаясь при этом легковесной и быстрой. Это частично объясняется использованием раздельных сверток.

В компоненте сети для предсказания плотной оценки позы можно получить улучшение за счет увеличения области интереса (RoI) на вход с 14×14 до 32×32, как предложено в [98]. Также можно использовать модуль ASPP [11], за которым следуют сверточные слои, при этом было решено исключить слой медленных нелокальных сверток [89], чтобы улучшить скорость работы.

Эмпирические результаты

В экспериментальной оценке моделей Mobile Parsing R-CNN в основном использовась средняя точность (AP) при различных порогах геодезического точечного сходства (GPS) [3], наряду со средней точностью определения областей изображения с людьми. Модель Parsing R-CNN реализована и модифицирована с использованием пакетов PyTorch и Detectron2 [96]. Эксперименты включали широкое исследование использования различных компонент модели, включая тип базовый подсети, шеи и количество каналов в модели. Детали разных вариантов архитектуры описаны в таблицах 4, 5 и 6.

Было обнаружено, что варианты Mobile Parsing R-CNN (A) и (B) достигают баланса между средней точностью и вычислительной эффективностью, показывая значительные улучшения числа кадров в секунду как на центральном процессеоре, так и на графическом ускорителе. Результат работы показан на рисунке 11.

Таблица 4: *Основные различия между представленными моделями.* Результаты на DensePose-COCO minival. 3x LR обозначает трехкратное увеличение времени тренировки по сравнению с обычной настройкой. P_i обозначает уровень признаков с разрешением $1/2^i$ от размера входных изображений. #Каналов обозначает количество каналов внутри *шеи* и *голов*. LR обозначает шаг обучения.

	DensePose R-CNN (базовый метод) [3]	Parsing R-CNN [98]	Mobile Parsing R-CNN (A)	Mobile Parsing R-CNN (B)
Базовая подсеть	ResNet-50 [29]	ResNet-50 [29]	Single-Path [78]	Single-Path [78]
Шея	FPN[49]	FPN[49]	FPN[49]	BiFPN[82]
RoI разрешение	14×14	32×32	32×32	32×32
Тип пулинга	RoIPool	RoIPool	RoIAlign	RoIAlign
Область/класс головы	2 линейных слоя	2 линеныйх слоя	2 свертки	2 свертки
Уровень признаков для предсказаний	P_2,P_3,P_4,P_5	P_2	P_2	P_2
DensePose голова	8 сверток	ASPP[11]+NL[89]+4 свертки	ASPP[11]+4 свертки	ASPP[11]+4 свертки
#Каналов	512	512	256	64
#Параметров	59.73M	54.36M	11.35M	3.35M
GPU FPS	13.16	10.15	12.03	22.77 (3x LR: **23.55**)
CPU FPS	1.62	1.39	1.42	2.02 (3x LR: **2.10**)
область, AP	57.8	59.609	56.370	55.39 (3x LR: 56.83)
densepose, AP	49.8	54.676	49.512	46.79 (3x LR: 51.08)

Таблица 5: *Исследование абляции базовой подсети, используемой в Mobile Parsing R-CNN (A).* Базовые подсети сортируются по топ-1 точности. Результаты на DensePose-COCO *minival.*

Базовая подсеть	Топ-1 точность (%)	#Параметров	область, AP	densepose, AP	GPU FPS	CPU FPS
ResNet-50 [29]	77.15	33.61M	60.0	**54.7**	11.05	1.34
EfficientNet-B3 [80]	81.636	16.03M	59.027	53.084	8.31	1.37
EfficientNet-EdgeTPU-L [21]	80.534	17.89M	60.069	53.378	8.11	1.34
MixNet-XL [81]	80.120	19.10M	58.444	51.475	8.54	1.32
EfficientNet-B2 [80]	79.688	13.68M	58.041	51.800	9.33	1.38
MixNet-L [81]	78.976	14.62M	57.481	50.649	8.52	1.34
EfficientNet-EdgeTPU-M [21]	78.742	14.57M	58.825	52.302	9.21	1.37
EfficientNet-B1 [80]	78.692	13.03M	57.654	51.053	9.49	1.39
CondConv-EfficientNet-B0 [21, 97]	77.304	18.32M	56.779	49.231	10.63	1.40
EfficientNet-EdgeTPU-S [21]	77.264	13.12M	58.296	51.606	10.03	1.39
MixNet-M [81]	77.256	12.39M	56.834	48.371	9.39	1.35
EfficientNet-B0 [80]	76.912	12.10M	56.271	49.647	10.53	1.39
MixNet-S [81]	75.988	11.52M	55.132	46.685	10.34	1.37
MobileNetV3-Large-1.0 [34]	75.516	12.04M	54.537	47.195	11.54	1.40
MnasNet-A1 [81]	75.448	10.94M	54.648	47.036	11.21	1.38
FBNet-C [95]	75.124	11.49M	55.399	47.983	10.97	1.37
MnasNet-B1 [79]	74.658	11.31M	52.280	47.658	11.24	1.37
Single-Path [78]	74.084	11.35M	56.370	49.512	**12.03**	**1.42**
MobileNetV3-Large-0.75 [34]	73.442	10.92M	52.763	44.736	11.02	1.36
MobileNetV3-Large-1.0 (minimal) [34]	72.244	10.48M	52.464	44.632	11.33	1.36
MobileNetV3-Small-1.0 [34]	67.918	10.07M	49.614	35.808	10.62	1.35
MobileNetV3-Small-0.75 [34]	65.718	9.74M	44.224	32.650	10.16	1.33
MobileNetV3-Small-1.0 (minimal) [34]	62.898	9.58M	45.989	36.522	10.34	1.34

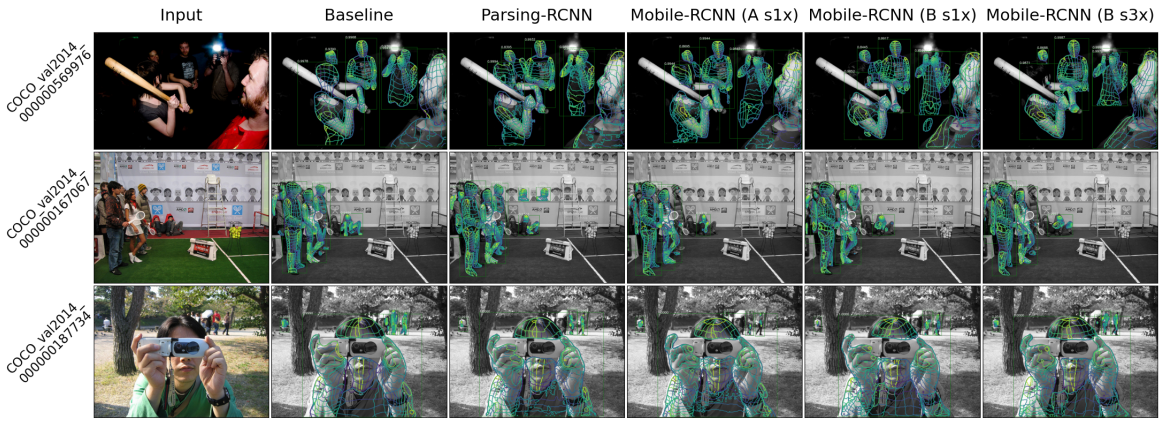

Рис. 11: *Качественное сравнение различных моделей.* В качестве вывода модели изображены контуры с цветовыми кодами координат U и V.

Таблица 6: *Исследование абляции типа шеи и количества каналов.* Количество каналов одинаково в шее и головах. Результаты на DensePose-COCO *minival*.

	Шея	#каналов	#параметров	область, AP	densepose, AP	GPU FPS	CPU FPS
Mobile Parsing R-CNN (A)	FPN	256	11.35M	56.371	49.512	12.03	1.42
	BiFPN	256	10.53M	58.106	52.80	12.05	1.41
	BiFPN	112	4.41M	56.41	49.64	19.04	1.78
	BiFPN	88	3.82M	56.08	48.19	20.43	1.87
Mobile Parsing R-CNN (B)	BiFPN	64	3.35M	55.39	46.79	22.77	2.02

Заключение

Исследование посвящено улучшению обобщающей способности модели DensePose для плотной оценки позы человека при строгих ограничениях на размер и скорость работы модели. Улучшение достигнуто благодаря тщательному подбору компонентов в различных частях модели, таких как выбор оптимальной базовой подсети для извлечения признаков, архитектуры "шеи", а также архитектуры "голов" для детекции людей и предсказания DensePose. Благодаря этим нововведениям, модель стала работать быстрее и качественнее, что позволило впоследствии запустить ее локально на мобильном устройстве.

3.5. Моделирование трехмерной модели головы человека по одному изображению

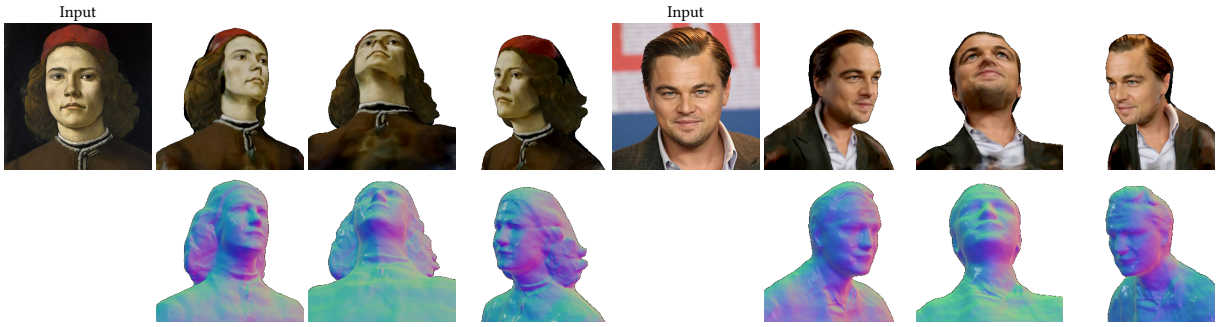

Рис. 12: *Моделирование трехмерной модели головы человека по одному изображению.* Модель Multi-NeuS может моделировать реалистичные 3D портреты головы по одной фотографии.

Рассматривается задача автоматического получения текстурированной трехмерной модели головы человека по одному изображению. Эта задача крайне важна в различных областях, таких как кинопроизводство, дополненная реальность (AR), виртуальная реальность (VR), расширенная реальность (XR) и игровая индустрия. Решение задачи направлено на автоматическое моделирование как геометрических, так и текстурных деталей, что позволяет обойти необходимость в трудоемком и длительном ручном создании моделей. Важность этой задачи заключается в её потенциале изменить создание контента в упомянутых областях.

В области моделирования внешности головы существует несколько методов, в основном основанных на 2D-представлениях [39, 17, 74]. Для 3D-моделирования существующие подхо-

ды, такие как H3D-Net и NeuralHeadAvatars [70, 26], часто опираются на 3D-сканирование или синтетические данные, что ограничивает их практическое применение. Недавнее появление метода NeuS и связанных методов, таких как UNISURF и VolSDF [86, 63, 99, 41], открыло новые возможности в моделировании 3D объектов с помощью нейросетевых неявных представлений, но эти методы не обобщаются на новые объекты и сцены.

Для преодоления этих ограничений предлагается новая архитектура, *Multi-NeuS*, на основе нейронных неявных функций, которая эффективно адаптируется к множеству объектов одного класса (например, модели головы человека) и реконструирует поверхность по набору фотографий с разных ракурсов. Multi-NeuS является улучшением модели NeuS [86] за счет добавления набора параметров, общего для нескольких моделируемых объектов одновременно. Этот подход позволяет достичь хорошей скорости и эффективности с точки зрения объема требуемых данных. В отличие от своих предшественников, Multi-NeuS может генерировать высококачественные 3D-портреты головы с одной или нескольких фотографий, что является значительным прогрессом в области 3D-портретирования (см. рисунок 12).

Описание модели

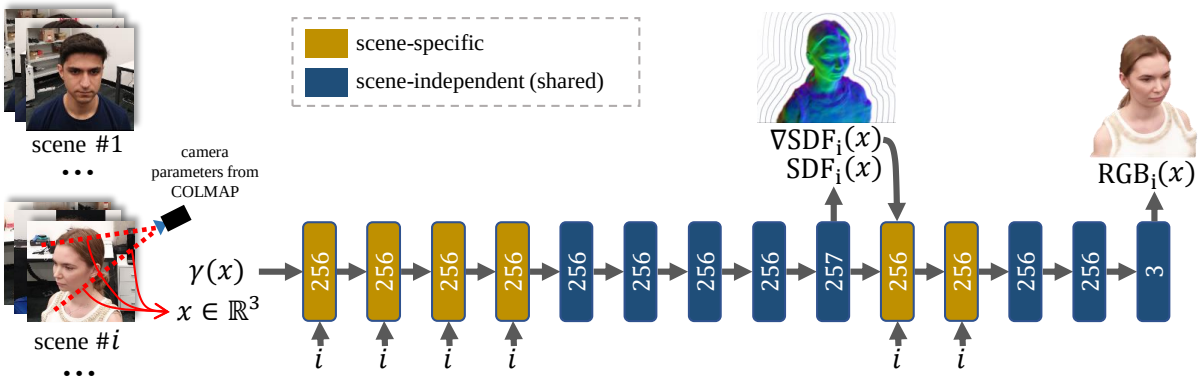

Рис. 13: *Обзор Multi-NeuS.* Модель на основе нейронных неявных функций может представлять несколько объектов одного класса. Общие слои (синего цвета) содержат априорные знания о классе объектов. Модель обучается на задаче генерации новых видов с использованием объемного рендеринга и попиксельной функции потерь на наборе данных из нескольких сцен. При адаптации к новой сцене сначала корректируются слои, специфичные для сцены (желтого цвета), затем следует дообучение всех слоев.

В исследовании представлена новая модель Multi-NeuS, на основе нейронных неявных функций, предназначенная для одновременной реконструкции нескольких объектов определенного класса. Основываясь на принципах NeuS [86] и NeRF [57], Multi-NeuS стремится преодолеть ограничения в сценариях, где доступно только одно или несколько изображений для построения реконструкции сцены. Рассматривается сценарий моделирования трехмерной головы человека как пример реализации подхода.

Рассмотрение новой модели начинается с метода реконструкции NeuS, модификации NeRF, предназначенной для моделирования непрозрачных объектов, в частности, для моделирования поверхности объекта. NeuS моделирует объект в виде неявной функции - представляет поверхность объекта как область пространства где функция расстояния до поверхности (SDF) равн нулю: $\{x \in \mathbb{R}^3 \mid \text{SDF}(x) = 0\}$. Две нейронные сети используются для моделирования SDF и RGB цвета в любой 3D точке, с плотностью, определенной как колоколообразная функция SDF, достигающая максимума в нуле, который соответствует поверхности объекта. Эти сети оптимизируются через дифференцируемый объемный рендеринг на задаче генерации новых видов. Далее обученная модель позволяет уже получить 3D поверхность объекта.

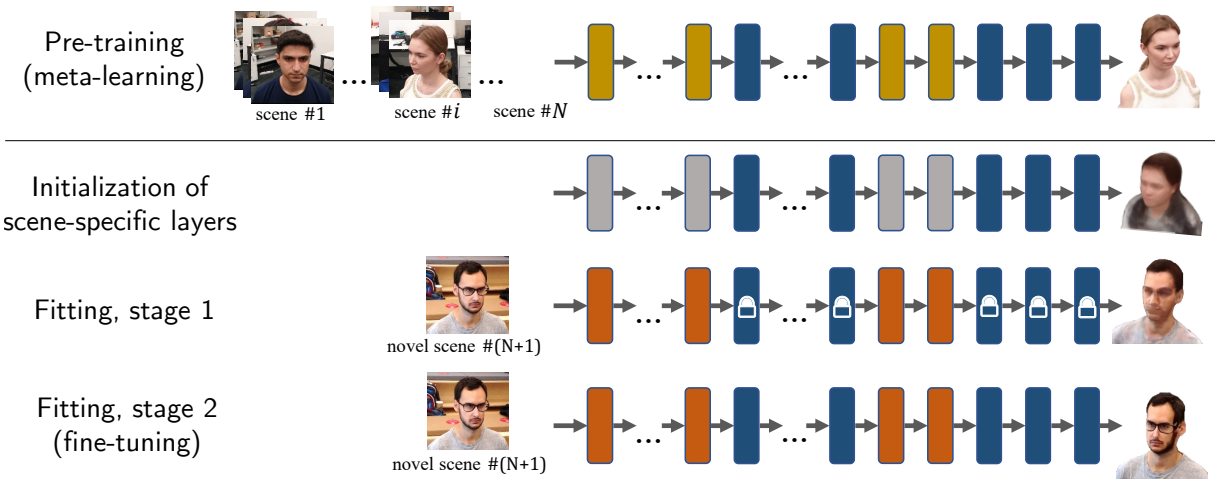

Рис. 14: *Этапы обучения Multi-NeuS.* Ряд 1: модель обучается для представления N сцен, включая слои, специфичные для сцен. Ряд 2: слои, специфичные для сцен, инициализируются с использованием взвешенной агрегации для адаптации к новой сцене. Ряд 3: только слои, специфичные для сцен, дообучаются для новых индивидов с ограниченным количеством изображений. Ряд 4: все слои дообучаются с уменьшенной скоростью обучения.

Предлагаемый метод Multi-NeuS расширяет метод NeuS для моделирования нескольких сцен за счет включения общих и специфичных для сцены слоев в его архитектуру (см. рисунок 13). Эта конструкция помогает переносить выделенные из данных априорные знания о классе сцен (объектов) в новые сцены, улучшая реконструкцию по нескольким доступным о сцене изображениям.

Процесс обучения Multi-NeuS состоит из двух основных этапов: мета-обучение и адаптация (см. рисунок 14). Во время мета-обучения модель предварительно обучается на наборе данных многовидовых изображений из нескольких сцен, что позволяет Multi-NeuS выделить общее представление этого класса сцен. Затем, на этапе адаптации, в модель добавляются и оптимизируются слои, специфичные для новых сцен, начиная с инициализации, представляющей «средний» объект из набора данных.

Далее ставится задача применения модели к случайным изображениям в естественных условиях, например, взятым из Интернета. Для модели требуются как внешние (экстринсики),

так и внутренние (интринсики) параметры камеры, которые обычно неизвестны для случайных фотографий.

Сначала вручную инициализируется матрица внутренних параметров, усредняя параметры камер, использованных во время стадии мета-обучения. Чтобы уменьшить возможные ошибки инициализации, все изображения из наборов для обучения, валидации и из естественных условий обрезаются вокруг области лица с отступом. Для нахождения внешних параметров камеры используются 3D координаты ключевых точек лица, по отношению к которым выравнивались все обучающие примеры на этапе мета-обучения. Для фотографий в естественных условиях предсказываются 2D координаты ключевых точек лица с использованием обученного детектора [7]. Имея 2D позиции ключевых точек, соответствующие 3D координаты и матрицу внутренних параметров, внешние параметры камеры можно оценить с помощью PnP [14].

Полученная оценка достаточно грубая, поэтому в ходе обучения матрица внешних параметров умножается на корректирующую матрицу, параметризуемую с использованием $\mathfrak{se}(3)$ алгебры Ли следуя [48]. Для корректировки внутренних параметров камеры задаются обучаемые коэффициенты, на которые умножаются фокальные расстояния матрицы. Все новые параметры оптимизируются одновременно с параметрами сети на каждой итерации обучения с использованием стохастической оптимизации.

Эмпирические результаты

Модель обучается на подмножестве набора данных SmartPortraits [43], состоящего из видео, снятого на смартфоны. С помощью программного обеспечения COLMAP [75] были извлечены кадры и соответствующие параметры камеры, а сами сцены были выровнены для обеспечения согласованных 3D координат.

Оценивается реконструкция по одному виду (изображению) и предоставляются как количественные (таблица 7), так и качественные (рисунок 15) сравнения. Предложенный метод Multi-NeuS показал сопоставимую производительность с H3D-Net [70] несмотря на то, что он был обучен на другом наборе данных с гораздо меньшим объемом и без доступа к 3D-сканам. Результаты по фотографиям, сделанным в естественных условиях, представлены на рисунке 16.

Рис. 15: *Качественное сравнение качества реконструкции по одному кадру.* Сравнение на первых четырех сценах набора данных H3DS. H3D-Net [70], изначально разработанный для реконструкции по трем кадрам, также может быть оценен в случае одного кадра на вход. H3D-Net был обучен на 10 000 3D-сканах из того же распределения, что и эти тестовые примеры, в то время как предложенный метод достигает сопоставимого качества, будучи обученным всего на 100 примерах. Кроме того, предложенный подход лучше восстанавливает ключевые черты лица человека и избегает эффекта «среднего лица».

Таблица 7: *Количественное сравнение качества реконструкции по одному кадру.* Сравнение с H3D-Net на наборе данных H3DS [70]. Рассчитывается однонаправленное расстояние Чамфера в миллиметрах, измеренное после выравнивания с помощью метода ICP [6]. Этот показатель применялся как к областям лица, так и к полным моделям головы. Меньшие значения указывают на лучшую производительность. 'F/L/R' обозначает ракурс входного изображения: 'фронтальный/левый/правый'.

	лицо				голова			
Входные кадры	*F*	*L*	*R*	*среднее*	*F*	*L*	*R*	*среднее*
H3D-Net, 3 кадра	-	-	-	1.34	-	-	-	10.53
H3D-Net, 1 кадр	**1.82**	1.83	1.91	1.85	13.83	**13.01**	12.51	13.12
Наш метод, 1 кадр	1.89	**1.77**	**1.86**	**1.84**	**13.00**	13.27	**11.95**	**12.74**

Заключение

В данной работе представлен *Multi-NeuS*, новый подход для реконструкции 3D портретов головы по одному или нескольким изображениям, улучшая обобщение в задачах 3D компью-

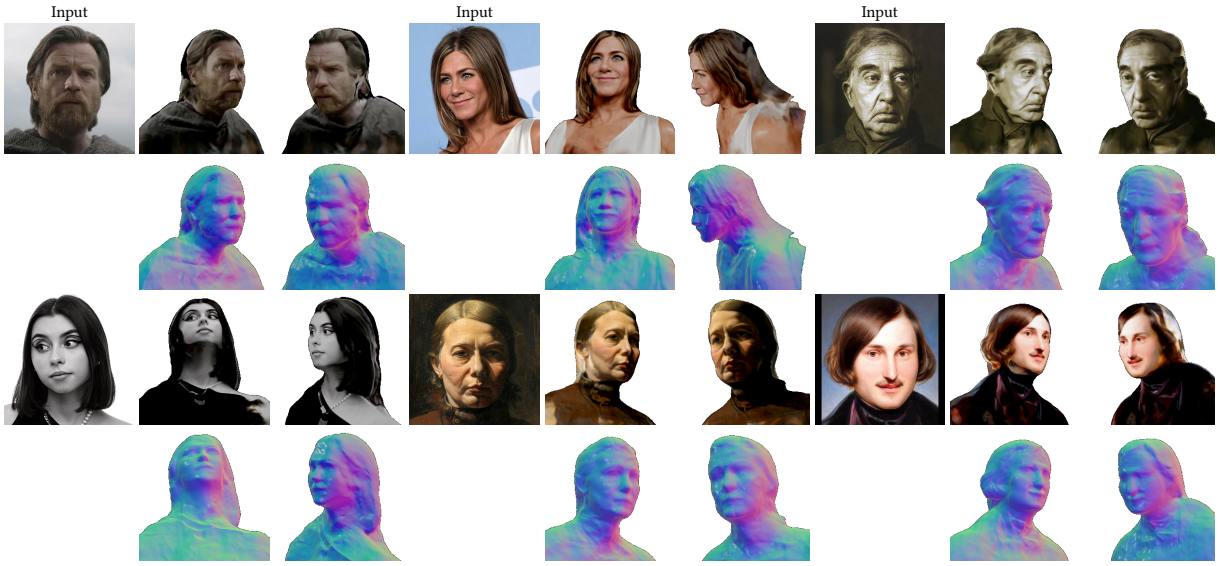

Рис. 16: *3D реконструкция по фотографии в естественных условиях.* Метод справляется с разнообразными прическами и хорошо работает на изображениях, выходящих за распределение набора данных для обучения SmartPortraits. Возможные артефакты на задней стороне обусловлены отсутствием разнообразия углов съемки в данных для обучения.

терного зрения. Обобщение, то есть способность модели точно реконструировать *новые* лица по одной или паре фотографий, достигается путем добавления априорных данных через предобучение на большом наборе изображений различных людей. Этот шаг предобучения позволяет модели захватывать специфические для класса особенности, снижая необходимость в длительной оптимизации для каждой сцены. Комбинируя обучение общим параметрам с адаптацией к конкретным сценам, Multi-NeuS эффективно реконструирует текстурированные поверхности. У метода есть ограничения, в первую очередь возникающие из-за ограниченного разнообразия набора данных для обучения. В будущем расширение набора данных и архитектурные улучшения позволят далее усовершенствовать возможности подхода.

4. Заключение

В данной диссертационной работе рассмотрены и предложены методы для улучшения обобщающей способности моделей в задачах 3D компьютерного зрения. Все представленные методы направлены на повышение эффективности и точности работы моделей в разнообразных, ранее невиданных условиях, что является ключевым фактором для успешного применения этих технологий в реальных сценариях.

Первое исследование представило модель генерации видео, основанную на моделировании видео в дискретном латентном пространстве. Уникальность данного подхода заключается в его способности генерировать видеопоследовательности по невиданным ранее входным условным кадрам, что достигается при значительно меньших вычислительных ресурсах по сравнению с существующими методами. Использование всего 8 графических процессоров V100 для

обучения модели, в то время как альтернативные подходы требуют до 512 тензорных процессоров, демонстрирует значительное улучшение эффективности без ущерба для качества обобщения.

Во втором исследовании предложен новый метод DEF для предсказания геометрических особенностей в 3D моделях. В отличие от традиционных методов, которые опираются на подгонку примитивов или оценку ковариационной меры Вороного, DEF использует обучение на больших синтетических наборах данных с минимальным количеством реальных данных. Метод обучается регрессии поля расстояний до особенностей на локальных участках, что повышает обобщающую способность и масштабируемость на новых, ранее невиданных 3D формах, даже при наличии шумов сканирования.

Третье исследование фокусируется на модели NPBG++, которая значительно улучшает обобщение в задаче генерации новых видов. Эта модель предсказывает нейронные дескрипторы напрямую из исходных изображений за один проход, избегая трудоемкой оптимизации на новой сцене. Такое нововведение позволяет модели быстро адаптироваться к новым окружениям, создавая высококачественные рендеринги с высокой скоростью визуализации, что делает ее эффективной по сравнению с существующими подходами.

В четвертом исследовании достигнуто значительное улучшение обобщающей способности модели DensePose для плотной оценки позы человека при строгих ограничениях на размер и скорость работы модели. Оптимизация различных компонентов модели, таких как базовая подсеть для извлечения признаков, архитектура "шеи" и "голов" для детекции людей и предсказания DensePose, позволила повысить производительность и качество работы модели, что в конечном итоге позволило запустить её локально на мобильном устройстве.

Наконец, в пятом исследовании представлен подход Multi-NeuS для реконструкции 3D портретов головы по одному или нескольким изображениям. Улучшение обобщающей способности достигается за счет предобучения модели на большом наборе изображений различных людей, что позволяет захватывать специфические для класса особенности и снижать необходимость в длительной оптимизации для каждой сцены. Комбинируя оптимизацию общих параметров с адаптацией к конкретным сценам, Multi-NeuS эффективно реконструирует текстурированные поверхности.

Таким образом, все представленные в работе методы демонстрируют значительное улучшение обобщающей способности моделей в задачах 3D компьютерного зрения. Каждое из предложенных решений не только превосходит существующие подходы по эффективности и точности, но и обеспечивает более широкое применение в различных прикладных задачах, таких как генерация синтетических данных, точная 3D реконструкция, эффективная генерация новых видов и определение позы человека. Эти достижения подчеркивают важность и значимость разработанных методов, открывая новые возможности для дальнейшего развития технологий 3D компьютерного зрения.

Список литературы

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XXII 16*, pages 696--712. Springer, 2020.

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98--109. IEEE, 2018.

[3] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297--7306, 2018.

[4] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1):1--21, 2021.

[5] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586--606. Spie, 1992.

[6] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239--256, 1992.

[7] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[8] Egor Burkov, Ruslan Rakhimov, Aleksandr Safin, Evgeny Burnaev, and Victor Lempitsky. Multi-neus: 3d head portraits from single image with neural implicit functions. *IEEE Access*, 2023.

[9] Yuanhao Cao, Liangliang Nan, and Peter Wonka. Curve networks for surface reconstruction. *arXiv preprint arXiv:1603.08753*, 2016.

[10] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834--848, 2017.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[13] A Clark, J Donahue, and K Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.

[14] Toby Collins and Adrien Bartoli. Infinitesimal plane-based pose estimation. *International Journal of Computer Vision*, 109(3):252–286, sep 2014.

[15] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[16] Kris Demarsin, Denis Vanderstraeten, Tim Volodine, and Dirk Roose. Detection of closed sharp edges in point clouds using normal estimation and graph theory. *Computer-Aided Design*, 39(4):276--283, 2007.

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690--4699, 2019.

[18] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018.

[19] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018.

[20] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *arXiv preprint arXiv:1710.05268*, 2017.

[21] Efficientnet-edgetpu: Creating accelerator-optimized neural networks with automl. https://ai.googleblog.com/2019/08/efficientnet-edgetpu-creating.html.

[22] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: self-supervised learning of decoupled motion and appearance for single-image video synthesis. *arXiv preprint arXiv:1910.07192*, 2019.

[23] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64--72, 2016.

[24] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1--8. IEEE, 2007.

[25] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653--18664, 2022.

[26] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Niessner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proc. CVPR*, 2022.

[27] T. Hackel, J. D. Wegner, and K. Schindler. Contour detection in unstructured 3d point clouds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1610--1618, 2016.

[28] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770--778, 2016.

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626--6637, 2017.

[31] Chems-Eddine Himeur, Thibault Lejemble, Thomas Pellegrini, Mathias Paulin, Loic Barthe, and Nicolas Mellado. Pcednet: A lightweight neural network for fast and interactive edge detection in 3d point clouds. *ACM Transactions on Graphics (TOG)*, 41(1):1--21, 2021.

[32] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735--1780, 1997.

[34] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314--1324, 2019.

[35] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[36] Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *International conference on machine learning*, pages 2157--2166. PMLR, 2018.

[37] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406--413, 2014.

[38] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[39] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[40] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, page 0, 2006.

[41] Petr Kellnhofer, Lars C. Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proc. CVPR*, June 2021.

[42] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9601--9611, 2019.

[43] Anastasiia Kornilova, Marsel Faizullin, Konstantin Pakulev, Andrey Sadkov, Denis Kukushkin, Azat Akhmetyanov, Timur Akhtyamov, Hekmat Taherinejad, and Gonzalo Ferrer. Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis. In *Proc. CVPR*, June 2022.

[44] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A flow-based generative model for video. *arXiv preprint arXiv:1903.01434*, 2(5), 2019.

[45] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440--1449, 2021.

[46] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[47] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31--42, 1996.

[48] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proc. ICCV*, 2021.

[49] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117--2125, 2017.

[50] Y. Lin, C. Wang, B. Chen, D. Zai, and J. Li. Facet segmentation-based line segment extraction for large-scale point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):4839--4854, 2017.

[51] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1--248:16, October 2015.

[52] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020.

[53] Albert Matveev, Ruslan Rakhimov, Alexey Artemov, Gleb Bobrovskikh, Vage Egiazarian, Emil Bogomolov, Daniele Panozzo, Denis Zorin, and Evgeny Burnaev. Def: Deep estimation of sharp geometric features in 3d shapes. *ACM Transactions on Graphics*, 41(4), 2022.

[54] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. *arXiv preprint arXiv:1812.01608*, 2018.

[55] Quentin Mérigot, Maks Ovsjanikov, and Leonidas J Guibas. Voronoi-based curvature and feature estimation from point clouds. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):743--756, 2010.

[56] 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 *European conference on computer vision*, pages 405--421. Springer, 2020.

[57] 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 *Proc. ECCV*, 2020.

[58] Theo Moons, Luc Van Gool, Maarten Vergauwen, et al. 3d reconstruction from multiple images part 1: Principles. *Foundations and Trends® in Computer Graphics and Vision*, 4(4):287--404, 2010.

[59] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255--1262, 2017.

[60] Seonghyeon Nam, Chongyang Ma, Menglei Chai, William Brendel, Ning Xu, and Seon Joo Kim. End-to-end time-lapse video synthesis from a single outdoor image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1409--1418, 2019.

[61] Natalia Neverova, David Novotny, and Andrea Vedaldi. Correlated uncertainty for learning dense correspondences from noisy labels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 920--928. Curran Associates, Inc., 2019.

[62] Natalia Neverova, James Thewlis, Riza Alp Guler, Iasonas Kokkinos, and Andrea Vedaldi. Slim densepose: Thrifty learning from sparse annotations and motion cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10915--10923, 2019.

[63] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. ICCV*, 2021.

[64] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099--5108, 2017.

[65] Prashant Raina, Sudhir Mudur, and Tiberiu Popa. Sharpness fields in point clouds using deep learning. *Computers & Graphics*, 78:37--53, 2019.

[66] Ruslan Rakhimov, Andrei-Timotei Ardelean, Victor Lempitsky, and Evgeny Burnaev. Npbg++: Accelerating neural point-based graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15969--15979, 2022.

[67] Ruslan Rakhimov, Emil Bogomolov, Alexandr Notchenko, Fung Mao, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Making densepose fast and light. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1869--1877, 2021.

[68] Ruslan Rakhimov*, Denis Volkhonskiy*, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *VISAPP 2021: 16th International Conference on Computer Vision Theory and Applications*, 2021.

[69] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. *arXiv preprint arXiv:2107.12512*, 2021.

[70] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proc. ICCV*, 2021.

[71] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216--12225, 2021.

[72] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234--241. Springer, 2015.

[73] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510--4520, 2018.

[74] Sergio Saponara, Abdussalam Elhanashi, and Alessio Gagliardi. Reconstruct fingerprint images using deep learning and sparse autoencoder algorithms. In *Real-Time Image Processing and Deep Learning 2021*, volume 11736, pages 9--18. SPIE, 2021.

[75] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016.

[76] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, volume 4067, pages 2--13. International Society for Optics and Photonics, 2000.

[77] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[78] Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. Single-path nas: Designing hardware-efficient convnets in less than 4 hours. *arXiv preprint arXiv:1904.02877*, 2019.

[79] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820--2828, 2019.

[80] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[81] Mingxing Tan and Quoc V Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[82] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. *arXiv preprint arXiv:1911.09070*, 2019.

[83] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21--22, 1999 Proceedings*, pages 298--372. Springer, 2000.

[84] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306--6315, 2017.

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998--6008, 2017.

[86] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proc. NeurIPS*, 2021.

[87] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690--4699, 2021.

[88] Xiaogang Wang, Yuelang Xu, Kai Xu, Andrea Tagliasacchi, Bin Zhou, Ali Mahdavi-Amiri, and Hao Zhang. Pie-net: Parametric inference of point cloud edges. *Advances in Neural Information Processing Systems*, 33, 2020.

[89] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794--7803, 2018.

[90] Christopher Weber, Stefanie Hahmann, and Hans Hagen. Sharp feature detection in point clouds. In *2010 Shape Modeling International Conference*, pages 175--186. IEEE, 2010.

[91] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724--4732, 2016.

[92] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.

[93] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467--7477, 2020.

[94] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534--8543, 2021.

[95] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734--10742, 2019.

[96] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[97] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, pages 1305--1316, 2019.

[98] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 364--373, 2019.

[99] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Proc. NeurIPS*, 2021.

[100] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471--4480, 2019.

[101] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Ec-net: an edge-aware point set consolidation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 386--402, 2018.

[102] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.