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

На правах рукописи

Бурков Егор Андреевич

ОБУЧЕНИЕ ПО ДАННЫМ КАК ОСНОВА МОДЕЛИРОВАНИЯ ПОЗЫ И ВНЕШНОСТИ ЛЮДЕЙ И ВИРТУАЛЬНЫХ АВАТАРОВ

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата компьютерных наук

Научный руководитель:
к.ф.-м.н. Виктор Сергеевич Лемпицкий

Москва — 2024

Содержание

1 Тема диссертации

Введение. Данная диссертация посвящена *human capture* – набору задач, конечной целью которых является автоматическое распознавание и моделирование позы и внешности человека. Значение понятий «распознавание» и «моделирование», а также конкретные задачи широко варьируются в зависимости от задействованных технологий, используемых датчиков и сенсоров, желаемого уровня детализации и, что важно, целевых приложений.

Примеры применения алгоритмов human capture включают телеприсутствие в дополненной или виртуальной реальности, распознавание действий в видеонаблюдении, автоматическое управление спортивными трансляциями или быстрое создание персонажа 3D-видеоигры по заданному образу. Исследовательские же задачи варьируются от классификации жестов и распознавания походки до полной трехмерной оценки формы тела, волос или одежды.

Актуальность. Методы компьютерного зрения для определения позы и моделирования внешности человека лежат в основе десятков практических приложений. Такие приложения могут делать жизнь безопаснее (например, умная охранная система для дома; домашний мониторинг пожилых людей), удобнее (электронный фитнес-тренер; виртуальная примерочная; управление устройством жестами рук) и интереснее (аватары для видеозвонков и трансляций; танцевальные и спортивные симуляторы и игры). В то же время, практически ни один из таких методов компьютерного зрения не совершенен, не решает свою задачу до конца, и, как правило, требует доработки и ограничений для каждого конкретного приложения. Например, существующие алгоритмы для трёхмерной оцифровки человека по одной фотографии часто требуют ручной дообработки художником, и, в целом, обычно не приводят к фотореалистичным результатам. Некоторые приложения, например, управления устройством жестами, до сих применяются ограниченно и только в очень простых случаях, именно, из-за недостатков алгоритмов определения положения пальцев (что связано с низкой точностью, тряской, неустойчивостью к освещению и перекрытиям, с низкой скоростью).

Цель данной диссертации – улучшить некоторые из подобных алгоритмов, исправить часть их недостатков. Таким образом, тема диссертации важна не только с научной точки зрения (недостатки алгоритмов компьютерного зрения, связанных с человеком и виртуальными аватарами, сейчас привлекают огромный интерес исследователей), но и непосредственно с точки зрения практических приложений.

Цели. Человеческому мозгу обычно достаточно полагаться только на визуальные сигналы, чтобы выполнить своего рода «human capture» (например, представить, как кто-то может вы-

глядеть с разных сторон). Мы считаем, что это возможно потому, что люди «учатся на данных», поскольку раньше видели многих других людей и опираются на этот опыт. Однако, по нашему мнению, многие современные алгоритмы human capture недостаточно используют знания о человеческих телах, выученные по данным – используя, например, созданные вручную признаки машинного обучения, или сложные алгоритмы с ручным вмешательством.

Наша основная цель, исходящая из мотивации выше – улучшить некоторые из современных алгоритмов human capture в нескольких сценариях, где обучение на данных может быть использовано гораздо шире, чем сейчас. Для этого мы опираемся на такие инструменты, как богатые наборы данных, нейросетевые модели, обучаемые целиком (end-to-end), и такие парадигмы, как обучение без разметки (self-supervised learning) и метаобучение. Конкретно, в диссертации мы сосредотачиваемся на **решении следующих проблем human capture**:

- определение 3D-позы человека, для применения в первую очередь в телеприсутствии;
- определение позы головы и выражения лица для применения в первую очередь в телеприсутствии и в переносе движения на модель другого человека или аватара;
- 3D-реконструкции модели головы.

Они соответствуют следующим **задачам**:

1. разработать алгоритм для реалистичного и плавного определения 3D-координат ключевых точек тела по видео с одной RGB-камеры;
2. улучшить (по сравнению с современными алгоритмами) точность, плавность и устойчивость к перекрытиям при определении тех же 3D-координат с *нескольких* RGB-камер;
3. создать новый тип представления позы (для лица и головы), который, в отличие от координат ключевых точек, не зависит от человека, т.е. не содержит информации о личности;
4. тщательно оценить применимость этого представления, в частности, при переносе движения на модель другого человека или аватара;
5. разработать алгоритм для восстановления 3D-меша головы человека по одной или нескольким RGB-изображениям.

2 Основные результаты

Основные результаты диссертации:

1. Существующий алгоритм телеприсутствия для людей и аватаров в полный рост был дополнен возможностью отрисовки аватара с произвольных точек зрения благодаря новому алгоритму для оценки 3D-координат ключевых точек тела, лица, стоп и рук по RGB-видео с одной камеры.

2. Для систем из нескольких камер были разработаны два более точных алгоритма определения таких 3D-координат позы тела. Алгоритмы значительно улучшили точность на популярных открытых данных и повысили устойчивость к перекрытиям. Концептуальное новшество алгоритмов относительно существующих в том, что нейросети в них обучаются целиком, напрямую предсказывая 3D-координаты.

3. Получено латентное (т.е. неинтерпретируемое) представление, достаточно точно описывающее позу головы и выражение лица (не хуже существующих представлений вроде 3DMM или координат ключевых точек), но в то же время содержащее гораздо меньше информации для идентификации человека и не требующее вручную размеченных данных благодаря самообучению (self-supervised learning).

4. С использованием этого латентного представления разработана система телеприсутствия портретного формата, которая естественным образом поддерживает произвольных людей в качестве источника позы (driver), сохраняя при этом внешность аватара.

5. Разработан алгоритм, оценивающий 3D-меш головы по нескольким или одному изображению (например, селфи или картина). По сравнению с наиболее релевантным из существующих методов, представленный требует обучения на гораздо более простых данных (100 видео со смартфона против 10.000 3D-сканов).

Новизна этих результатов заключается в следующем:

- Оба представленных алгоритма для триангуляции 3D-позы с нескольких камер основаны на нейросетях, обучаемых непосредственно на предсказание 3D-координат (в отличие от существующих методов, предсказывающих промежуточные 2D-координаты). Эти алгоритмы значительно улучшили точность определения 3D-позы на двух популярных наборах данных.

- Предложенное латентное представление позы формируется без учителя по данным без разметки, потенциально поддерживает неограниченную детализацию позы и, как показывают измерения, не зависит от личности человека.

- Реализованная система телеприсутствия портретного формата поддерживает произвольных людей в качестве источника позы, будучи (на момент публикации) проще и качественнее предыдущих методов с такой же функцией.

- Описаны и опробованы два концептуально новых способа оптимизации нейронных неявных функций (neural implicit functions) под несколько объектов (сцен) одновременно.
- Предложенный алгоритм 3D-реконструкции головы достигает лучших результатов, чем самый точный релевантный метод из того же семейства, при этом будучи обученным на гораздо более простом наборе данных.

Практическая ценность полученных результатов напрямую следует из исходного выбора задач. Было показано, что предложенные алгоритмы определения 3D-позы – как с одной, так и с нескольких камер – позволяют дополнить существующую систему телеприсутствия функцией обзора с произвольных углов (free viewpoint); после этого несложно адаптировать такую систему для дополненной или виртуальной реальности. Независимые от человека латентные дескрипторы позы позволяют обучать генеративные нейросети для анимирования аватаров любыми людьми, не сталкиваясь с «протеканием» их личности (черт лица, формы головы, цвета кожи и т.д.) в отрисованный аватар; такое анимирование найдёт применение в кинопроизводстве или приложениях, связанных с развлечениями. Описанный в диссертации алгоритм 3D-реконструкции головы может быть использован для разработки практических приложений, например, пособий для интерактивного изучения истории или быстрого создания своего аватара для видеоигры.

Кроме того, результаты диссертации имеют и непрямую практическую ценность. В частности, точное и плавное определение 3D-позы без физических маркеров, что может серьёзно упростить и удешевить некоторые продукты (квест-комнаты в виртуальной реальности; захват движения актёра в кинопроизводстве; управление руками в VR-шлемах, без пультов-контроллеров и камер глубины). Дескрипторы позы, не зависящие от человека, могут быть полезны в приложениях, где важна безопасность персональных данных.

Личный вклад. Все результаты диссертации получены лично соискателем или при его непосредственном участии. Все этапы алгоритмов оценки 3D-позы с одной камеры, определения латентных описаний выражения лица и позы головы (кроме случайных изменений позы при обучении) и восстановления 3D-сетки головы были разработаны и реализованы автором. В соответствующих разделах диссертации соискатель также является автором исходных идей и предшествующим им полных обзоров литературы. При разработке алгоритма по определению 3D-позы с нескольких камер автор полностью отвечал за один из двух наборов данных (Human3.6M), а именно, за его подготовку, валидацию алгоритма на нём и некоторые эксперименты по обучению моделей. Все средства визуализации также были реализованы автором.

Таким образом, **положениями, выносимыми на защиту**, являются:

1. Два метода для триангуляции ключевых точек, полностью обучаемые по данным.
2. Латентное представление позы, получаемое методом самообучения.
3. Система для переноса движения с одного человека на модель другого человека или аватара, основанная на вышеупомянутом представлении.
4. Алгоритм для 3D-реконструкции модели головы по одному изображению с известными параметрами камеры.

3 Публикации и апробация работы

Публикации повышенного уровня

1. E. Burkov, I. Pasechnik, A. Grigorev, V. Lempitsky. **Neural Head Reenactment with Latent Pose Descriptors**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13786-13795, 2020. [Indexed in SCOPUS, CORE A*]
2. E. Burkov, R. Rakhimov, A. Safin, E. Burnaev, V. Lempitsky. **Multi-NeuS: 3D Head Portraits from Single Image with Neural Implicit Functions**. *IEEE Access*, vol. 11, pp. 95681-95691, 2023. [Indexed in SCOPUS, Q1]
3. K. Iskakov, E. Burkov, V. Lempitsky, Y. Malkov. **Learnable Triangulation of Human Pose**. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 7718-7727, 2019. [Indexed in SCOPUS, CORE A*]
4. A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, A. Vakhitov, V. Lempitsky. **Textured Neural Avatars**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2387-2397, 2019. [Indexed in SCOPUS, CORE A*]
5. E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky. **Few-Shot Adversarial Learning of Realistic Neural Talking Head Models**. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9459-9468, 2019. [Indexed in SCOPUS, CORE A*]

Доклады на конференциях и семинарах

1. ”Textured Neural Avatars”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 16–20 июня 2019 г.
2. ”Learnable Triangulation of Human Pose”. IEEE International Conference on Computer Vision (ICCV). 27 октября – 2 ноября 2019 г.
3. ”Few-Shot Adversarial Learning of Realistic Neural Talking Head Models”. IEEE International Conference on Computer Vision (ICCV). 27 октября – 2 ноября 2019 г.
4. ”Neural Head Reenactment with Latent Pose Descriptors”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 14–19 июня 2020 г.

Прочие публикации

1. E. Burkov, V. Lempitsky. **Deep Neural Networks with Box Convolutions**. *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 6214-6224, 2018. [Indexed in SCOPUS, CORE A*]

4 Содержание работы

Основная часть диссертации состоит из трёх глав, каждая из которых пересказана ниже. Первая глава описывает результаты исследований сначала из публикации №4 и затем из №3, вторая основана на публикации №1 и предшествующей ей №5, а третья соответствует №2.

4.1 Трёхмерная поза тела человека

Данная глава вдохновлена системой телеприсутствия для трансляции человека в полный рост, а именно её реализацей, в которой изображение генерирует нейросеть, получая на вход позу человека в виде 2D-координат его ключевых точек (Рисунок 1).

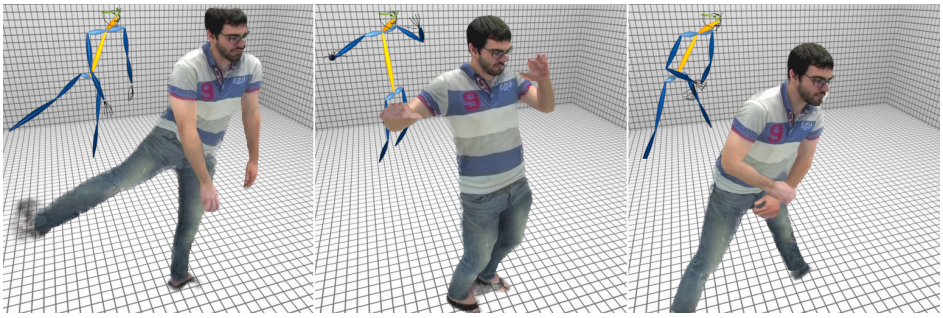

Рис. 1: Примеры входов и выходов системы телеприсутствия для трансляции человека в полный рост. В каждом примере вверху слева целевая поза в формате ключевых 2D-точек, остальное – конечный результат (отрисованный аватар).

Было бы желательно получать целевую позу не в 2D-, а 3D-координатах, ведь тогда её можно спроецировать на произвольные камеры, а значит, естественным образом добавить функцию обзора аватара с произвольных точек зрения – неотъемлемую в приложениях виртуальной или дополненной реальности (Рисунок 2).

4.1.1 Добавление третьей размерности к 2D-позе с помощью информации о движении

В первую очередь мы рассматриваем простой сценарий, где *ведущий* (человек, манипулирующий аватаром, т.е. источник целевой позы) снимается одной RGB-видеокамерой. Мы предлагаем несложный алгоритм для предсказания 3D-позы ведущего по особенностям движения его 2D-позы. Последняя заранее предсказана готовым алгоритмом OpenPose [2]. В основе алгоритма многослойный перцептрон (MLP) для регрессии с некоторыми особенностями (Рисунок 3). В частности, используются специально разработанные процедуры прямой и обратной нормализации данных, а также маски валидности ключевых точек.

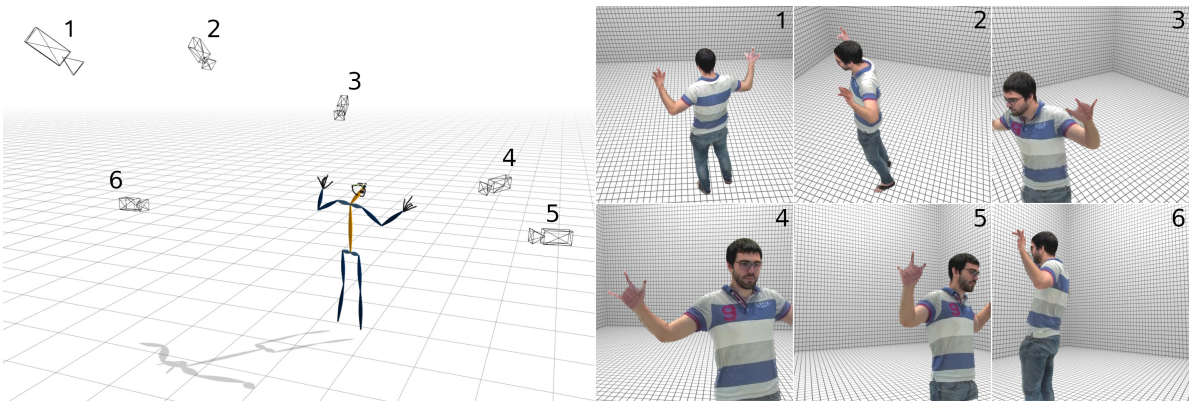

Рис. 2: Улучшение, рассматриваемое в главе 3: захват 3D-позы, благодаря которому возможен обзор аватара с произвольных углов, т.е. выбор произвольной виртуальной камеры. Слева: целевая поза (координаты ключевых точек) и 6 примеров произвольных камер. Справа: отрисованные аватары, соответствующие камерам.

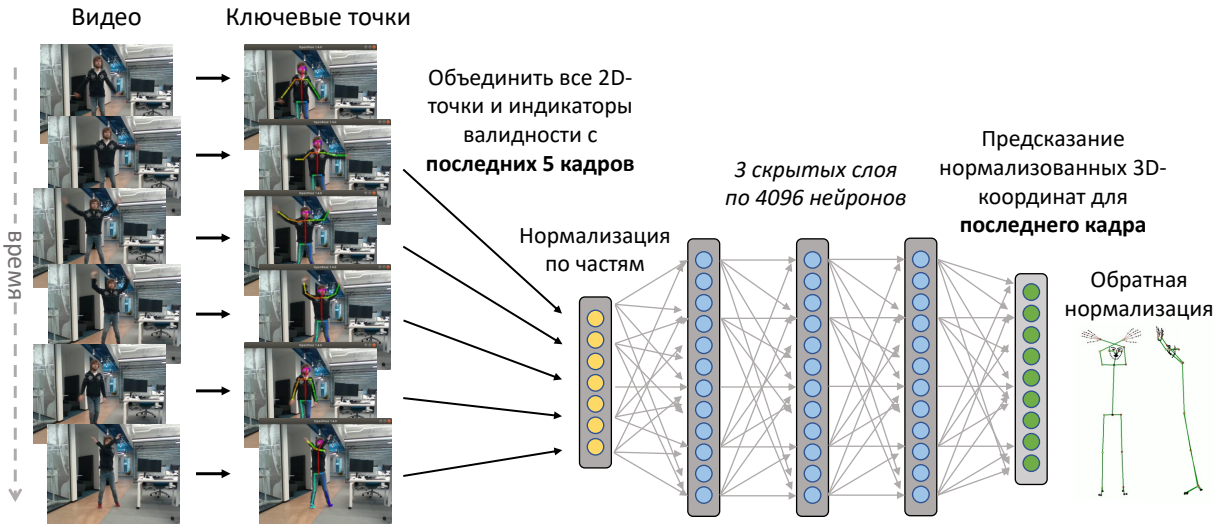

Рис. 3: Регрессионный многослойный перцептрон для предсказания 3D-позы по уже предсказанным 2D-позам для каждого кадра видео.

К этому алгоритму нас отчасти привело обилие видеоданных с десятков калиброванных камер в наборе данных CMU Panoptic (Рисунок 4). Из него мы получили (извлекли, оценили и триангулировали) в сумме около 770 000 3D-поз. Это позволило обучить нейросеть не просто закономерностям в разных позах человека, но ещё и в их движении.

Примеры работы алгоритма показаны на Рисунке 5. Наша модель способна предсказывать 3D-позы несмотря на то, что даны только 2D-координаты без какой-либо 3D-информации. Хоть предсказанные позы и не точны, а их ре-проекции на исходную камеру могут отличаться от 2D-позы из OpenPose, они, как правило, реалистичны и в то же время передают позу человека на изображении.

Модель в целом выдаёт достаточно плавные предсказания благодаря «встроенному» временно́му сглаживанию, которое вызвано постоянным обуславливанием нейросети на временной

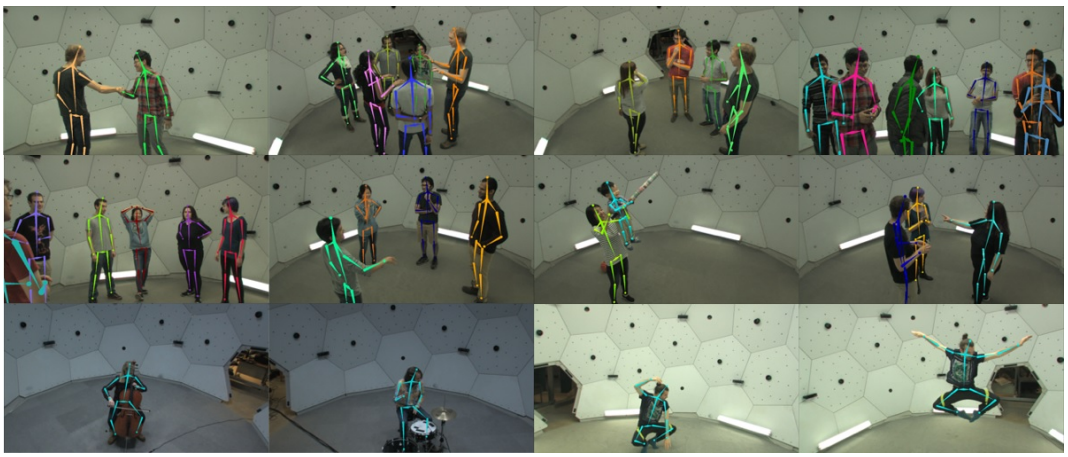

Рис. 4: Примеры кадров из набора данных CMU Panoptic с визуализацией основных ключевых точек тела.

контекст. Важно также, что перевод 2D-позы в 3D лишь незначительно влияет на производительность, тратя около 1.5 мс на кадр.

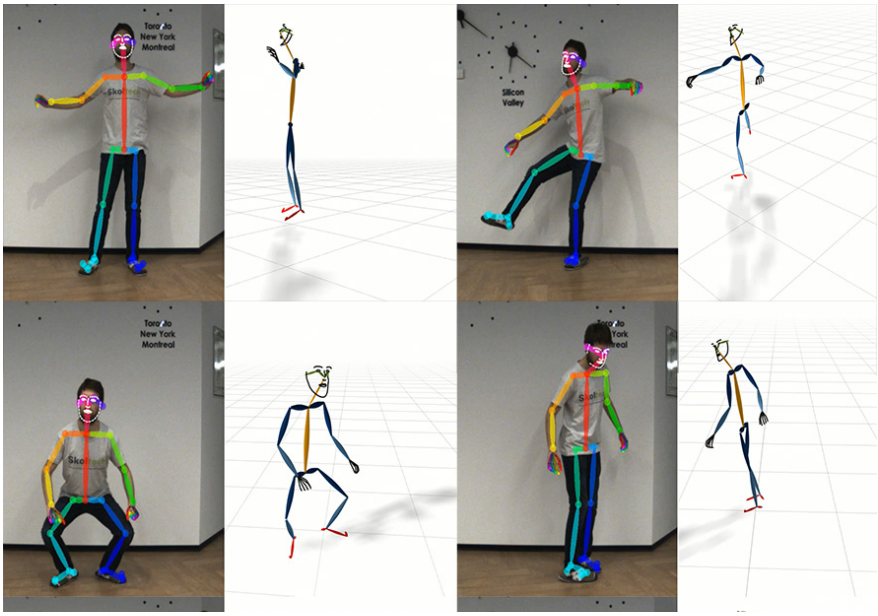

Рис. 5: Примеры работы алгоритма по предсказанию 3D-позы по 2D-позе на видео. Слева: последний кадр видео и визуализация 2D-позы, предсказанной OpenPose. Справа: предсказанная 3D-поза, вид с другого угла обзора.

Существует много способов улучшить этот базовый алгоритм. Например, очевидно, что для нейросети неестественно подавать в неё напрямую бинарную маску валидности. У модели также бывают проблемы с мелкими движениями, где она склонна к «регрессии к среднему»; это можно решить с помощью состязательной функции потерь [3], применённой по маске. Наконец, хотя временной контекст и помогает модели разрешать неоднозначность среди множества 3D-поз, соответствующих данной 2D-проекции, проблемы всё же начинаются, когда человек долго стоит без движения. Если человек замирает в неоднозначной позе более чем на 5 кадров, модель разрывается между позами-кандидатами, и возникает тряска. Эту

проблему могут решить рекуррентные архитектуры.

4.1.2 Обучаемая триангуляция 3D-позы для случая нескольких камер

Во втором разделе главы изучается другой путь определения 3D-позы – на этот раз точный (однако, как в прошлом разделе, с уклоном в реалистичность поз). Дело в том, что по одной камере теоретически невозможно восстановить точную 3D-позу, так как проекция не инъективна. Поэтому мы рассматриваем сценарий с *несколькими* калиброванными камерами.

Если несколько RGB-камер, снимающих один объект, *калиброваны* (т.е. известны их 3D-координаты, матрицы вращения, матрицы проекции), теоретически возможно оценить точные 3D-координаты ключевых точек объекта, определив их 2D-координаты на каждой камере, а затем *триангулировав* эти координаты в 3D. В идеальном случае это почти всегда возможно даже по двум камерам, однако на практике есть много препятствий: ошибки в определении 2D-координат, погрешности калибровки камер, другие источники шума, а главное – заслонённые или выходящие за край кадра точки. В литературе по созданию наборов данных с использованием триангуляции [8, 24] авторы прибегают к чрезмерному количеству камер для достижения достаточной точности 3D-координат – именно из-за перечисленных препятствий.

Мы предлагаем два решения, которые, по сравнению с существующими методами, значительно реже выдают очевидно нереалистичные позы, справляются с точками, не видными ни с одной камеры, и достигают той же точности с значительно меньшим количеством камер.

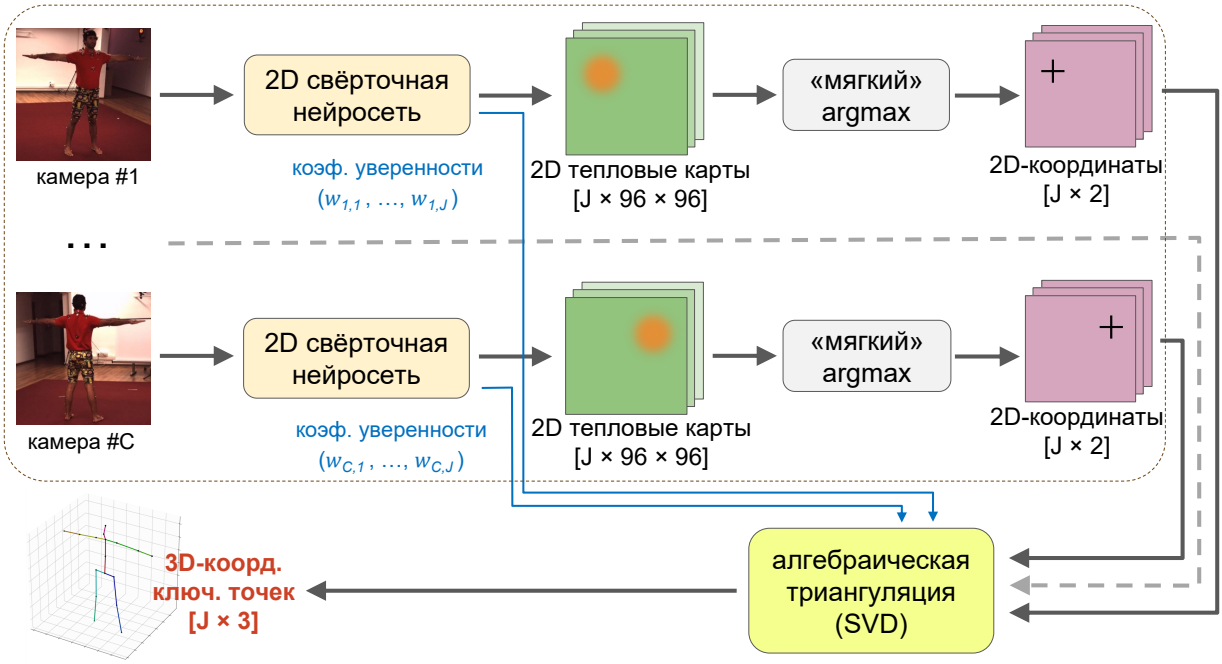

Рис. 6: Схема ”алгебраического”подхода для 3D-триангуляции позы.

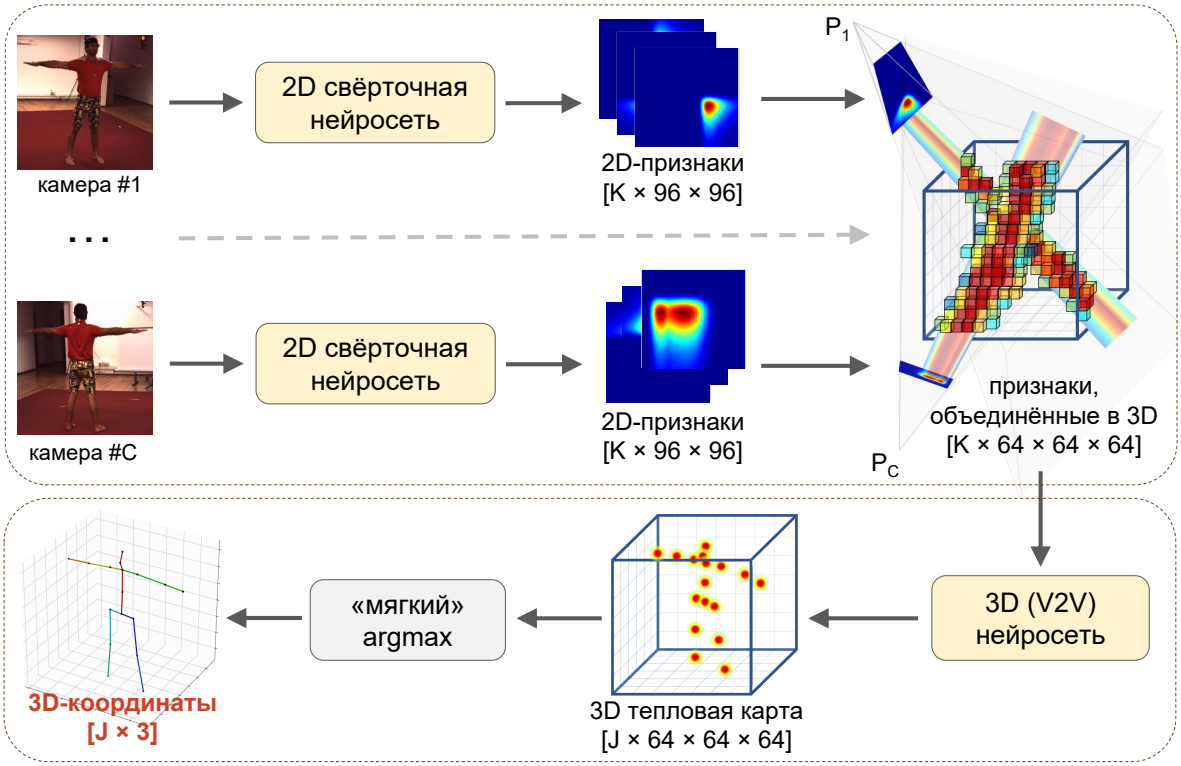

Рис. 7: Схема ”волюметрического”подхода для 3D-триангуляции позы.

В основе обоих решений лежит идея *обучаемой* триангуляции; для обучения мы используем либо точные 3D-координаты, полученные по физическим маркерам на одежде, либо псевдо-истинные координаты, полученные через триангуляцию с огромного числа камер. Важно отметить, что модели в обоих решениях полностью дифференцируемые, что позволяет обучать их целиком без промежуточных этапов.

Первое решение, ***«алгебраическое»***, напоминает классические методы триангуляции, однако полностью состоит из дифференцируемых операций, обеспечивающих обратное распространение ошибки (Рисунок 6). Классическая свёрточная нейросеть определяет 2D-координаты ключевых точек на каждом изображении, предсказывая «тепловые карты» уверенности и далее применяя к ним операцию soft-argmax, а также предсказывает скалярные значения уверенности для каждой ключевой точки на каждом кадре. Затем все предсказания триангулируются в 3D-точки с использованием дифференцируемых операций. Нейросеть обучается на оптимизацию функции потерь между предсказанными и истинными 3D-координатами, а именно «мягкой» модификации среднеквадратичной ошибки.

«Волюметрическое» **решение.** Главный недостаток базового «алгебраического» решения в том, что изображения с разных камер обрабатываются независимо друг от друга, из-за чего в модели не формируется априорное знание о 3D-позах человека, а также нельзя отфильтровать камеры с ошибочной калибровкой.

Для решения этой проблемы мы добавили ещё один этап обучаемой обработки после объеди-

нения информации с разных камер (Рисунок 7). Вместо предсказания 2D-координат, свёрточная сеть теперь предсказывает латентные карты признаков, которые потом заполняют трёхмерную карту признаков через операцию *обратной проекции* в соответствии с матрицами камер. Далее, эта трёхмерная карта обрабатывается ещё одной, трёхмерной, свёрточной нейросетью уже для предсказания интерпретируемых трёхмерных «тепловых карт» уверенности.

Результаты. Мы провели эксперименты на двух больших наборах данных, которые предоставляют изображения с нескольких камер и 3D-координаты ключевых точек: Human3.6M [7] и CMU Panoptic [8, 22, 17] (он же использовался выше для предсказания 3D-позы по одной камере). По результатам экспериментов наши методы превзошли существующие (Таблица 1). Качественное сравнение (Рисунок 8) показывает, что «алгебраический» подход лучше традиционного базового решения, а «волюметрический» ещё и устойчивее к невидимым ключевым точкам. Мы демонстрируем, что модель, обученная на CMU Panoptic, успешно работает на Human3.6M, несмотря на разницу в камерах и форматах ключевых точек. Мы также обнаружили, что, чтобы достичь той же ошибки (17 мм), базовому решению требуется 12 камер, «алгебраическому» – 4, а «волюметрическому» – всего 2.

Методы для одной камеры (MPJPE относительно таза, мм)	
Martinez et al. [11]	62.9
Sun et al.[18]	**49.6**
Pavllo et al. [15] (∗)	**46.8**
Hossain & Little [5] (∗)	58.3
«Волюметрический», 1 камера (†)	49.9
Методы для неск. камер (MPJPE относительно таза, мм)	
Martinez [11] + триангуляция	57.0
Pavlakos et al. [14]	56.9
Tome et al. [19]	52.8
Kadkhodamohammadi & Padoy [9]	49.1
RANSAC (наша реализация)	27.4
«Алгебраический» (без уверенн.)	26.9
«Алгебраический»	22.6
«Волюметрический» (объединение: softmax)	**20.8**
«Волюметрический» (объединение: сумма)	21.3
«Волюметрический» (объединение с уверенн.)	**20.8**

Метод	MPJPE, мм
RANSAC	39.5
«Алгебраический» (без уверенн.)	33.4
«Алгебраический»	21.3
«Волюметрический» (softmax)	**13.7**
«Волюметрический» (сумма)	**13.7**
«Волюметрический» (уверенности)	14.0

Таблица 1: Результаты на данных Human3.6M (**слева**) и CMU Panoptic (**справа**). Методы, использующие временну́ю информацию во время предсказаний, отмечены «(∗)». Наш метод для одной камеры (отмечен «†») использует примерную оценку координат таза, полученную с нескольких камер.

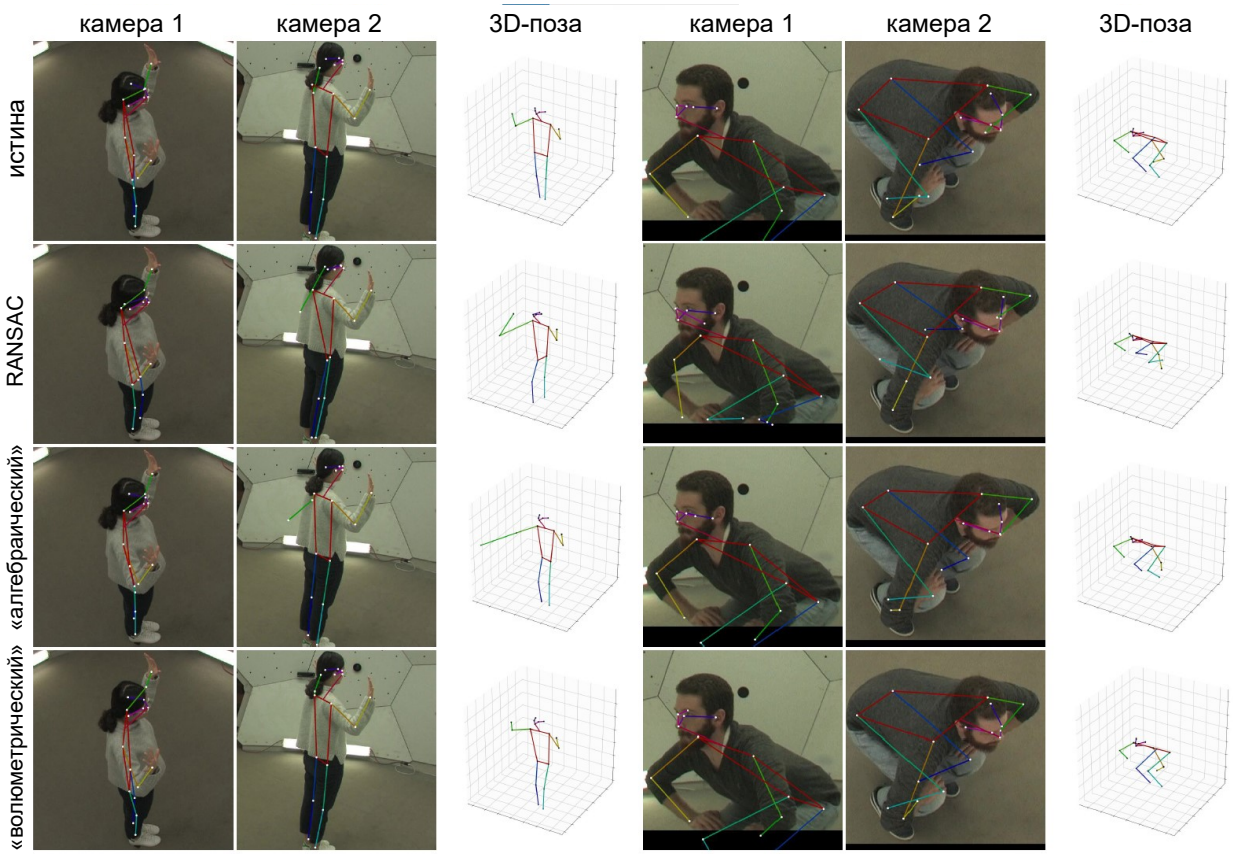

Рис. 8: Результаты триангуляции позы на валидационном подмножестве CMU Panoptic (2 камеры).

4.2 Представление позы головы и лица, не зависящее от личности

Во второй главе мы продолжаем изучать захват позы, и наша главная мотивация – вновь телеприсутствие. В то время как прошлая глава была посвящена улучшению определения ключевых точек, в этой мы обращаем внимание на фундаментальные недостатки ключевых точек как описания позы в целом, и стремимся разработать совершенно новое, лучшее описание позы.

Мы рассматриваем систему для переноса выражения лица и позы головы с ведущего на другого человека или аватара, с референсной реализацией [25], похожей на ту, что в предыдущей главе. Эта система тоже принимает на вход визуализацию ключевых точек, и поэтому подвержена многим их недостаткам. Наиболее проблемные из них – информация о личности содержится в координатах точек (это мешает, если аватар отличается от ведущего), ограниченная описательная способность (Рисунок 9), и временна́я тряска.

Для решения этих проблем мы заменили в данной системе фиксированный детектор 2D-точек на *кодировщик позы* – нейросеть, обучаемую вместе со всей системой. Модель обучается на наборе видео предсказывать новый видеокадр, как показано на Рисунке 10.

Интуитивно кажется, что ничего не мешает кодировщику позы обучиться помещать в дескриптор позы и информацию о личности тоже. В худшем случае, такая система деградирует до автокодировщика, а кодировщик личности становится бесполезен. Сначала мы ожидали, что для решения этой проблемы понадобится специальная функция потерь, например, состязательная [3] или циклического соответствия [26, 6]. По-видимому, чтобы «распутать» (disentangle) позу и личность, оказалось достаточно:

1. Сделать меньше параметров в кодировщике позы, чем в кодировщике личности (в нашем случае, MobileNetV2 и ResNeXt-50 соответственно).

Рис. 9: Снизу: аватары, отрисованные системой, основанной на координатах ключевых точек [25] по позе с соответствующих изображений ведущего (сверху). Форма лица ведущего «протекает» в аватар, вызывая ощутимую непохожесть (a, b, f); также, поза отличается от позы ведущего (c, d, e, f).

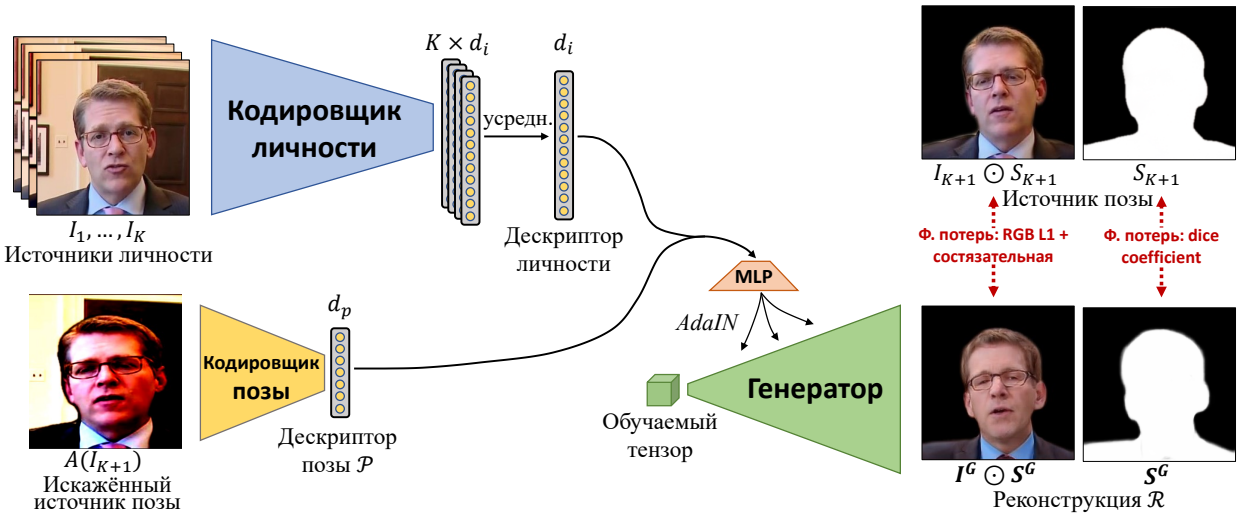

Рис. 10: На каждом шаге метаобучения система выбирает несколько случайных кадров из видео, которые обрабатываются двумя кодировщиками. Меньший из них, *кодировщик позы*, обрабатывает один из кадров (источник позы), остальные обрабатываются *кодировщиком личности*. Полученные векторные описания позы и личности человека соответственно передаются в генератор, чья задача – восстановить источник позы. Поскольку в кодировщике позы меньше параметров, и к его входу сначала применяются случайные искажения личности, то кодировщик позы обучается извлекать и помещать в описание позы только информацию, не зависящую от личности, а остальную извлекает кодировщик личности.

2. Применять в обучении случайные изменения источника позы, сохраняющие позу, но меняющие личность.

3. Предсказывать маску человек-фон и накладывать функцию потерь на изображение с уже удалённым фоном.

Результаты. Первая часть экспериментов направлена непосредственно на латентные дескрипторы позы, которые выдаёт обученный кодировщик позы. В задаче распознавания эмоций на данных Multi-PIE [4] наши описания позы лучше, чем другие, работают для соотнесения разных людей в одной позе. Мы также обучаем нейросеть-MLP для регрессии 2D-координат ключевых точек лица по нашим дескрипторам позы и личности; в таком эксперименте наши дексрипторы точнее, чем дескрипторы FAb-Net [21], хотя и менее точные, чем лучшая нейросеть для этой задачи, обученная напрямую по разметке. Примеры на видео (https://shrubb.github.io/research/latent-pose-reenactment/) показывают, что интерполяция наших дескрипторов по сфере приводит к плавным реалистичным изменениям позы, и что на видео они гораздо плавнее, чем ключевые точки.

Далее, мы проводим оценку самой системы переноса позы. В количественном сравнении мы используем две метрики: *ошибка личности* (насколько личность аватара соответствует референсным изображениям) и *ошибка позы* (насколько точно повторены выражение лица и поза ведущего). Рисунок 11 показывает, что наша система достигает лучшего компромисса

между метриками по сравнению с [25] и строго лучше остальных систем-конкурентов. Это подтверждается качественным сравнением (Рисунок 13).

Наконец, мы проводим тщательное абляционное исследование, уменьшая размерность описания позы, увеличивая кодировщик позы, оставляя фон в выходах генератора, и выключая случайные искажения источника позы. Количественные (Рисунок 12) и качественные эксперименты доказывают, что наш выбор итоговой архитектуры даёт наилучший компромисс между ошибками личности и позы.

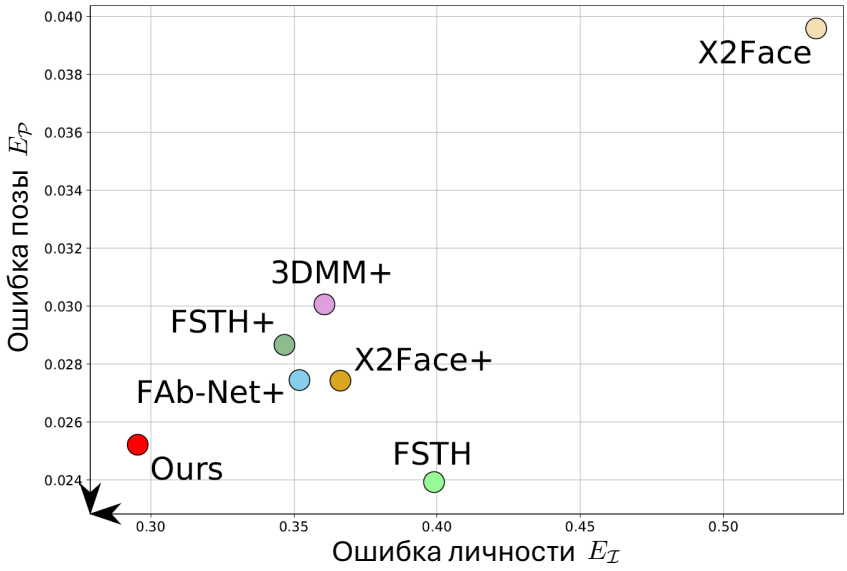

Рис. 11: Сравнение систем для переноса выражения лица и позы головы по качеству сохранения личности аватара и качеству передачи позы (стрелки указывают улучшение).

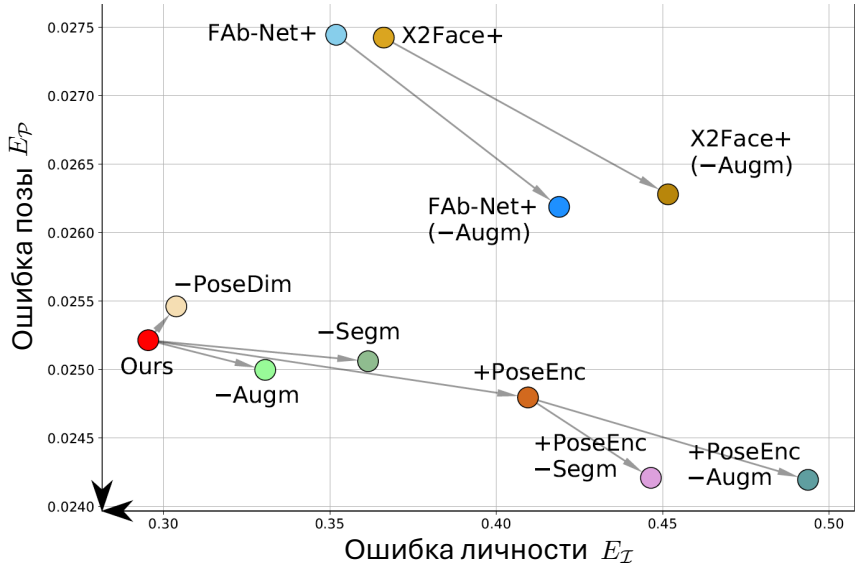

Рис. 12: Влияние некоторых гиперпараметров системы. Легенду см. в диссертации.

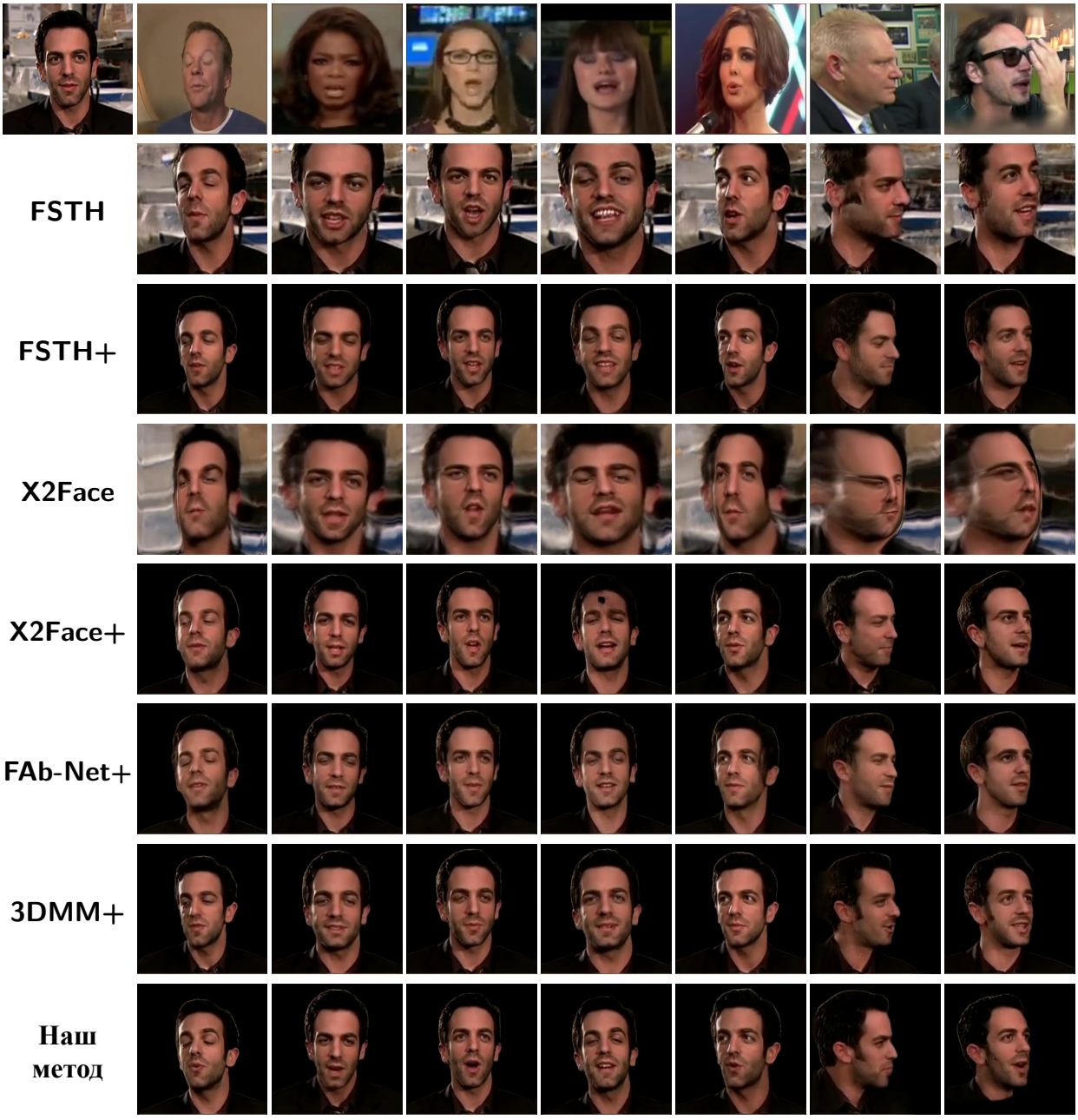

Рис. 13: Сравнение систем для переноса позы на другого человека на тестовой выборке VoxCeleb2. Слева вверху: один из 32 кадров, задающих аватар. Верхний ряд: ведущие (источники позы). Наш метод лучше сохраняет личность аватара и в то же время успешно переносит мимику с ведущего.

4.3 Восстановление 3D-поверхности головы по одному изображению

Две прошлых главы рассматривали захват позы человека, вторая также изучала точность оценки личности, а в последней мы полностью сосредотачиваемся на оцифровке личности. Наша задача здесь – автоматическая генерация 3D-модели головы по нескольким и особенно по одному изображению.

Как и выше, мы, насколько это возможно, опираемся на обучение по данным, а также хотим, чтобы наша модель хорошо обобщалась на несколько (1-2) новых примеров. Поэтому наша цель – создать мета-архитектуру в духе двух систем, описанных в прошлой главе. Кроме того, нам желательно избежать сложных данных вроде 3D-сканов или искусственных 3D-моделей.

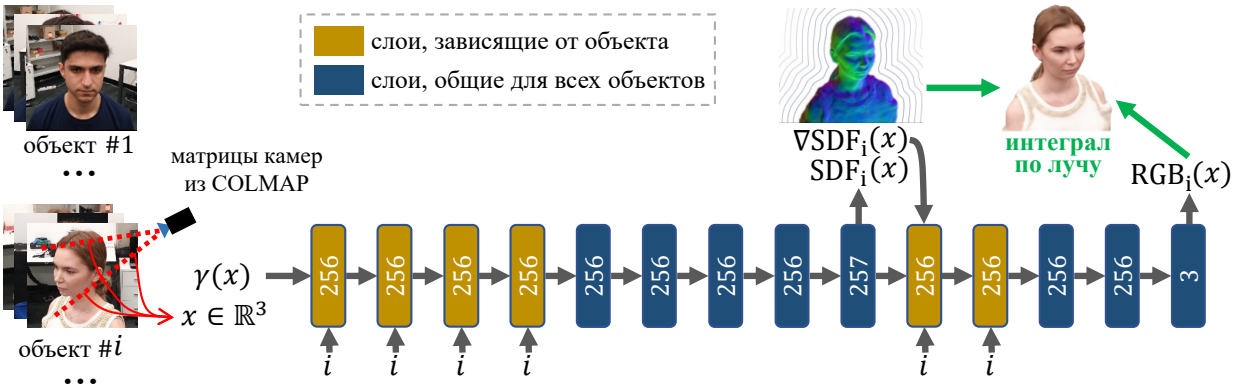

Рис. 14: Архитектура Multi-NeuS, трёхмерной нейронной неявной функции, способной одновременно задавать несколько объектов одного класса (прямоугольники – полносвязные слои, указана размерность выхода; γ – кодирование 3D-координат). Так как некоторые слои (синие) используются для всех объектов, они способны выучивать общее знание о всех объектах выборки и потом применять его на новых, позволяя задавать объекты лишь по нескольким примерам. Модель обучается через объёмный рендеринг и попиксельную функцию потерь, как и NeuS [20], но под несколько объектов сразу. Затем, когда дан новый объект, сначала под него обучаются слои, зависящие от объекта (жёлтые), а после – все слои вместе.

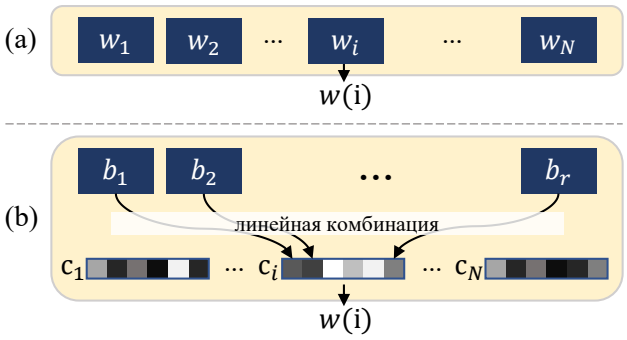

Рис. 15: Два варианта общих слоёв в наших экспериментах, *независимый* и *низкоранговый*. Оба – полносвязные слои, чьи веса $w(i)$ зависят от номера объекта i. В независимом слое обучается свой набор весов для каждого из N объектов, в низкоранговом – r общих наборов и дополнительно коэффициенты их линейной комбинации для каждого объекта.

По этим причинам мы прибегаем к *нейронным неявным функциям*, которые хорошо показали себя в восстановлении 3D-объектов только по калиброванным RGB-изображениям [20, 13, 23] и основаны на обучаемых нейронных сетях. Конкретно, наша базовая архитектура – NeuS [20], вариант [12] для непрозрачных объектов. Это регрессионная нейросеть-MLP, получающая на вход 3D-координату и выдающая расстояние со знаком до поверхности объекта, а также направленную RGB-яркость. Обучающие данные – изображения объекта и их параметры камер. В процессе обучения среди них выбирается случайный пиксель и на воображаемом луче камеры, соответствующем пикселю, выбирается ряд 3D-точек. Предсказания нейросети в этих точках интегрируются по лучу и результат сравнивается (в функции потерь) с цветом пикселя на изображении.

Наш метод называется Multi-NeuS (Рисунок 14). Мы модифицируем NeuS для обучения под N объектов одновременно: создаётся N копий NeuS, где *некоторые* из слоёв не копируются, а являются общими для всех копий сразу. Затем мы обучаем эти N копий одновременно, каждый под свою сцену, применяя к скопированным слоям дополнительную регуляризацию (Рисунок 15).

Мы обучаем Multi-NeuS на подмножество набора данных SmartPortraits [10], состоящее из $N = 107$ коротких видео неподвижных людей со смартфона. Затем, для 3D-реконструкции нового человека, мы добавляем новый $(N+1)$-й набор объекто-специфичных слоёв (усреднив существующие N), обучаем только его под изображение(-я) нового человека, а затем «размораживаем» остальные слои и дообучаем их все вместе с матрицами камеры. Если параметры камеры недоступны, например, для фото из интернета, мы грубо оцениваем их через

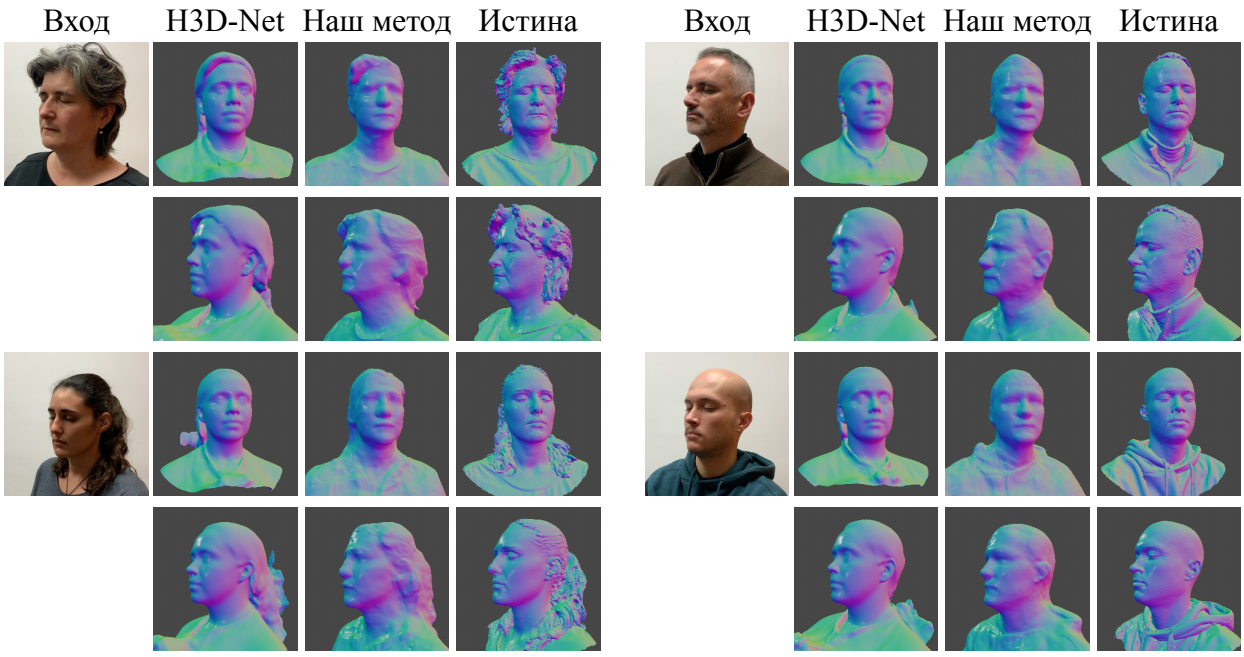

Рис. 16: Восстановление первых четырёх объектов из набора H3DS по одному фото.

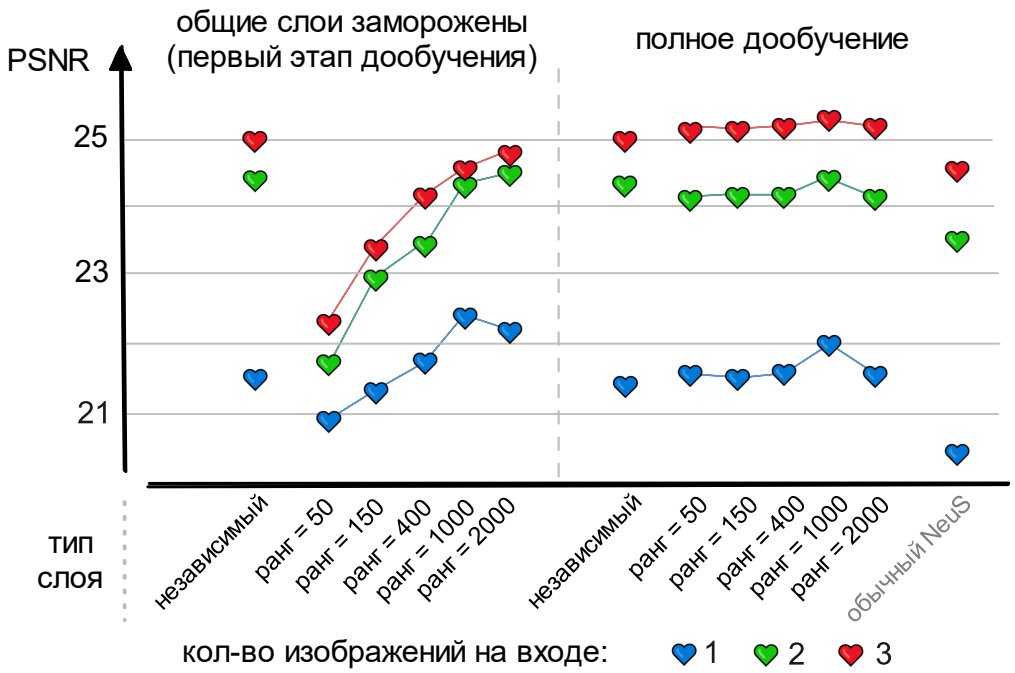

Рис. 17: Качество восстановления нового ракурса в зависимости от типа слоёв, зависящих от объекта. Метамодели с низким рангом недостаточно гибки первом этапе дообучения; модели с бóльшим рангом подгоняются к объекту лучше и поэтому дают более удобную инициализацию для второго этапа.

примерные 3D-координаты ключевых точек лица [1].

Результаты. Мы оцениваем наш метод на наборе данных H3DS [16] – он содержит 3D-сканы голов и поэтому позволяет провести численное сравнение. В нём наша лучшая модель показала себя не хуже главного метода-конкурента – H3D-Net [16] (Табл. 2), при этом обучаясь на гораздо более простых данных (1̃00 видео против 1̃0 000 3D-сканов). Помимо этого, у нашего метода менее выражена «регрессия к среднему» (Рисунок 16).

Мы также демонстрируем дополнительные результаты восстановления по одному изображению на некоторых фото из интернета и картинах на соответствующем рисунке (см. полный текст) и в видео (`https://shrubb.github.io/research/multi-neus`).

Наконец, последний этап экспериментов посвящён валидации архитектуры. Количественное

	лицо				голова			
Положение камеры	*F*	*L*	*R*	*средн.*	*F*	*L*	*R*	*средн.*
H3D-Net, 3 фото	-	-	-	1.34	-	-	-	10.53
H3D-Net, 1 фото	**1.82**	1.83	1.91	1.85	13.83	**13.01**	12.51	13.12
Наш метод, 1 фото	1.89	**1.77**	**1.86**	**1.84**	**13.00**	13.27	**11.95**	**12.74**

Таблица 2: Ошибка восстановления 3D-сетки (Chamfer distance) на данных H3DS. ”F/L/R”означает ”анфас/слева/справа”.

(Рисунок 17) и качественное сравнение способности к обобщению в зависимости от разных настроек и типов слоёв, зависящих от объекта, показывают, что низкоранговый слой с $r = 1000$ оптимален для восстановления по одному изображению. Мы также перебрали и сравнили конфигурации расположения таких слоёв (оптимальное расположение отражено на Рисунке 14).

5 Выводы

Данная диссертация вдохновлена множеством вызовов в области автоматического распознавания и моделирования позы и внешности человека, а также преимуществами глубокого обучения по данным, когда непосредственно модель решает бóльшую часть целевой задачи. Таким образом, мы выбрали 4 задачи, в которых, на наш взгляд, такая парадигма (например: большие наборы данных; нейронные модели, обучаемые целиком без промежуточных этапов – end-to-end; самообучение; метаобучение) имеет огромный потенциал, но пока используется недостаточно.

Сначала, стремясь в первую очередь улучшить системы телеприсутствия, мы рассмотрели задачу определения 3D-координат ключевых точек тела. Для сценария с одной камерой мы предложили простой регрессионный подход, опираясь на большой датасет 3D-поз. Для сценария с несколькими камерами мы описали два новых решения, каждое из которых использует свёрточную нейросеть, обучаемую напрямую под функцию потерь в 3D. Благодаря этому оба решения значительно точнее и устойчивее к перекрытиям и сложным позам, чем существующие. В процессе этих исследований мы обнаружили фундаментальные недостатки ключевых точек как представления позы в целом, и, изучая их, пришли к методу обучения нейросети для оценки *неинтерпретируемых* дескрипторов позы по набору видео в режиме самообучения. Эта нейросеть обучается как часть новой системы для переноса выражения лица и позы головы на другого человека или аватара с сохранением внешности. Далее мы сосредоточились на захвате внешности и разработали метод для восстановления 3D-поверхности головы по одному или нескольким RGB-изображениям. Для этого мы использовали недавно ставший популярным метод нейронных неявных функций, который позволяет обучать наш метод всего по сотне людей.

Мы рассчитываем, что какие-то из результатов выше расширят рамки практического применения алгоритмов human capture – например, позволят запускать их на менее мощном или более простом оборудовании, либо улучшат точность или качество, чтобы освободить ресурсы для более требовательных приложений. Что не менее важно, наша работа привносит в область более общую и концептуальную новизну, такую как схема для «распутывания» (disentanglement) двух признаков методом самообучения (адаптированная под головы людей) или метод для метаобучения нейронных неявных функций. Результаты наших экспериментов также подсказывают новые направления будущих исследований, к примеру: добиться полной чистоты латентных дескрипторов позы от информации о личности с помощью более систематических методов для «распутывания» – например, из литературы по генеративным моделям; более компактные механизмы переиспользования весов в слоях для более

быстрого и точного обучения нейронных неявных функций для нескольких объектов.

Список литературы

[1] 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 *Proc. ICCV*, pages 1021–1030, 2017. 22

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 9

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, 2014. 11, 16

[4] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society, September 2008. 17

[5] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, pages 69–86. Springer, 2018. 14

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proc. ECCV*, 2018. 16

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 14

[8] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 12, 14

[9] Abdolrahim Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3D human pose regression. *arXiv*, 1804.10462, apr 2018. 14

[10] 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. 21

[11] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 14

[12] 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. 21

[13] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. ICCV*, 2021. 21

[14] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1253–1262, 2017. 14

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. *arXiv*, abs/1811.11742, 2018. 14

[16] 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. 22

[17] Tomas Simon, Hanbyul Joo, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *CVPR*, 2017. 14

[18] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 14

[19] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking Pose in 3D: Multi-stage Refinement and Recovery for Markerless Motion Capture. In *2018 International Conference on 3D Vision (3DV)*, pages 474–483. IEEE, sep 2018. 14

[20] 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. 20, 21

[21] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *Proc. BMVC*, 2018. 17

[22] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *arXiv preprint arXiv:1812.01598*, 2018. 14

[23] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Proc. NeurIPS*, 2021. 21

[24] Zhixuan Yu, Jae Shin Yoon, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi 1.0: Human multiview behavioral imaging dataset. *arXiv preprint arXiv:1812.00281*, 2018. 12

[25] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proc. ICCV*, 2019. 16, 18

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2017. 16