Л

УДК 004.932.2 **DOI** 10.24412/2413-7383-2025-4-39-49-59

О. А. Лямцев[1,2], И. И. Максименко[1,2]
[1]ФГБНУ «Институт прикладной математики и механики»
283048, г. Донецк, ул. Розы Люксембург, 74
[2] ФГБОУ ВО «Донецкий государственный университет»
283001, г. Донецк, ул. Университетская, 24

ОБЗОР И ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ТРЕХМЕРНОЙ ОЦЕНКИ ПОЗЫ ЧЕЛОВЕКА ПО ОДНОМУ ИЗОБРАЖЕНИЮ

O. A. Lyamtsev[1,2], I. I. Maksimenko[1,2]
[1]Federal State Budgetary Scientific Institution "Institute of Applied Mathematics and Mechanics"
283048, Donetsk, Rosa Luxemburg str., 74
[2]Federal State Budgetary Educational Institution of Higher Education "Donetsk State University"
283001, Donetsk, Universitetskaya str., 24

OVERVIEW AND CHALLENGES OF USING DEEP LEARNING FOR THREE-DIMENSIONAL ESTIMATION OF A PERSON'S POSTURE FROM A SINGLE IMAGE

В статье рассмотрена задача трехмерной оценки позы человека по монокулярному изображению, приведено математическое описание, выделены основные подходы, проблемы в задаче и их решения, архитектуры нейронный сетей. Среди архитектур нейронных сетей наиболее популярны графовые сверточные сети, извлекающие локальную информацию из ближайших суставов и трансформеры, моделирующие глобальные зависимости через механизм внутреннего внимания. Среди ключевых проблем при решении задачи можно выделить: 1) неопределенность глубины; 2) окклюзия; 3) недостаточный объем и разнообразие обучающих данных.
Ключевые слова: трехмерная оценка позы человека, компьютерное зрение, глубокое обучение, нейронные сети, монокулярное изображение

The article considers the task of 3D human pose estimation from monocular image, provides a mathematical description, highlights the main approaches, problems in the task and their solutions, neural network architectures. Among neural network architectures, graph convolutional networks are the most popular, extracting local information from nearby joints and transformers, modeling global dependencies through the internal attention mechanism. Among the key problems in solving the task, we can distinguish: 1) depth uncertainty; 2) occlusion; 3) insufficient volume and diversity of training data.
Keywords: 3D human pose estimation, computer vision, deep learning, neural networks, monocular image

Введение

Оценка позы человека в трёхмерном пространстве (*Human Pose Estimation, HPE*) [1] – это активно развивающаяся область, чья актуальность обусловлена прогрессом в глубоком обучении и растущим спросом со стороны прикладных сфер. В отличие от двумерной оценки, эта задача нацелена на точное предсказание пространственных координат ключевых точек тела, что даёт более полное и точное описание его положения. Такая детализация критически важна для решения прикладных задач высокого уровня, таких как дополненная реальность, биомеханический анализ и спортивная аналитика.

Постановка задачи

В рамках работы проводится анализ ключевых проблем и методов глубокого обучения в задаче трёхмерного оценивания позы человека. Для достижения поставленной цели были определены следующие задачи:

1. Сформулировать математическую постановку задачи 3D HPE.
2. Провести анализ существующих аппаратных подходов 3D HPE.
3. Изучить разновидности представлений человеческого тела
4. Классифицировать подходы для оценки позы с точки зрения на модельные и безмодельные
5. Рассмотреть разновидности безмодельного подхода
6. Исследовать наиболее эффективные архитектуры нейронных сетей, применяемые для решения данной задачи
7. Выявить основные проблемы в задаче 3D HPE

Математическая постановка задачи

Пусть у нас есть входное изображение I. Необходимо предсказать 3D-координаты $J = \{j_1, j_2, \dots, j_N\}$ из N ключевых точек тела. Каждая ключевая точка j_k представлена вектором 3D-координат $j_k = (x_k, y_k, z_k)$.

$I \in R^{H \times W \times 3}$ – входное RGB изображение высотой H и шириной W. В случае видео входные данные будут представлять собой последовательность изображений I_t, где t – временной шаг.

$J = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_N, y_N, z_N)\}$ – результирующий набор 3D-координат N ключевых точек, где каждая (x_k, y_k, z_k) – это 3D-координаты k-й ключевой точки.

Задача заключается в нахождении функции f, которая отображает входное изображение I на 3D-позу J: $f: I \to J$ или, в случае с видео: $f: \{I_t\} \to \{J_t\}$. Эта функция f обычно параметризуется нейронной сетью с весами θ. Таким образом, наша задача – оптимизировать θ так, чтобы предсказанные позы $\hat{J}$ были максимально близки к истинным позам J_{gt}.

Многие подходы сначала предсказывают 2D-позу, а затем используют ее для вывода 3D-позы. В этом случае функция f может быть декомпозирована на две подфункции:

$f_{2D}: I \to P_{2D}$, *где* P_{2D} – 2D-координаты ключевых точек.

$f_{3D}: P_{2D} \to J$

Некоторые методы используют параметрические модели тела, такие как SMPL (*Skinned Multi-Person Linear model*), для представления 3D-позы. В этом случае f предсказывает параметры модели (например, параметры формы и позы), а не напрямую 3D-координаты ключевых точек.

$f: I \to (\beta, \alpha)$, β – параметры формы, α – параметры позы.

Затем 3D-поза J получается из (β, α) с помощью дифференцируемого оператора SMPL.

Аппаратные подходы

HPE определяет положение тела и суставов человека по изображениям или видео. Существуют два основных подхода: маркерный и безмаркерный [2].

Маркерная система использует костюм с датчиками, данные с которых фиксируются камерами и преобразуются в 3D-модель. Этот подход обеспечивает высокую точность, но требует дорогого оборудования, ограничивает движения и уязвим к повреждениям датчиков.

Безмаркерная технология основана на компьютерном зрении и использует обычные или специализированные камеры. Для получения 3D-позы можно применять многокамерные системы с последующей триангуляцией или RGB-D камеры, предоставляющие данные о глубине. Однако движения, захваченные с одной точки, часто зашумлены, что требует дополнительной обработки методами машинного обучения.

Альтернативные технологии, такие как RADAR и LIDAR, хорошо справляются с окклюзией и обеспечивают высокое разрешение, но сталкиваются с проблемами разрозненности и разреженности данных, а также являются дорогостоящими. Инфракрасные датчики, например Kinect, подвержены помехам от солнечного света.

Наиболее современным и сложным подходом является использование одной RGB-камеры без данных о глубине, что стало возможным благодаря нейронным сетям. Его ключевое преимущество – масштабируемость и низкая стоимость, однако он сталкивается с ограничениями в виде окклюзии и неопределенности глубины.

Учитывая вышеперечисленное, основное внимание в исследованиях уделяется камерным решениям, поскольку именно они позиционируются как наиболее перспективные и передовые, в то время как работы на основе RADAR, LIDAR или ИК-детекторов сравнительно менее продвинуты.

Представление человеческого тела

Моделирование человеческого тела является ключевым компонентом HPE [3]. Человеческое тело представляет собой сложный нежёсткий объект, обладающий множеством характеристик, таких как кинематическая структура, текстура поверхности, форма и положение суставов. При этом модель не обязательно должна включать все атрибуты тела. В зависимости от сценариев применения в HPE обычно используются скелетные, контурные и объемные модели.

Скелетная модель включает положения суставов и ориентации конечностей, представляя структуру тела. Она интуитивно понятна и успешно применяется в 2D и 3D HPE, обеспечивая гибкое графическое представление, хотя и ограничена в передаче информации о текстуре и форме.

Контурная модель отражает взаимосвязи между частями тела, представляя его форму и внешний вид. В ней части тела аппроксимируются прямоугольниками, соответствующими контурам тела.

Объемная модель используется для 3D-оценки позы, предоставляя точные данные о форме и текстуре с помощью сканирования всего тела, что позволяет получать сетки тела различных форм и поз.

Генеративный и дискриминативный подход

До того, как методы, основанные на обучении, заняли основное место, генеративные модели были популярны и доминировали в 3D HPE [4]. Генеративные методы строят свои решения на основе моделей человеческого тела, тогда как дискрими-

нативные подходы ориентированы на нахождение соответствий между наблюдаемыми данными и позой человека без использования явной модели тела. Таким образом, их рабочие направления являются полностью противоположными.

Генеративные подходы к оценке позы основаны на использовании априорных знаний, заложенных в параметрических моделях человеческого тела. Восстановление позы осуществляется либо инвертированием кинематики из двумерных ключевых точек, либо поиском в пространстве конфигураций, который обычно формулируется детерминированно как задача нелинейной оптимизации или максимизации правдоподобия. Данный процесс включает два этапа: на этапе моделирования формируется функция правдоподобия, интегрирующая признаки изображения, кинематику модели, параметры камеры и анатомические ограничения; на этапе оценки определяется набор параметров позы, наилучшим образом согласующийся с наблюдениями. Хотя эти методы не требуют обучения на парных 3D-данных и обеспечивают высокую точность и физическую состоятельность результатов, их главным ограничением является зависимость от точной начальной инициализации и склонность к сходимости к локальным минимумам. Тем не менее, они демонстрируют высокую обобщающую способность к сложным позам, вариациям одежды и аксессуаров.

В противоположность этому, дискриминативные подходы (безмодельные или основанные на обучении), не используют явные параметрические модели тела и обычно реализуется с помощью машинного обучения. Вместо этого они изучают функцию отображения непосредственно из пространства изображений в пространство поз. Они не требуют инициализации и итеративной оптимизации, что делает их чрезвычайно быстрыми на этапе тестирования. После обучения процесс вывода сводится к вычислительной задаче прямого прогноза или ограниченного поиска, а не к оптимизации в многомерном параметрическом пространстве. Однако эта скорость достигается ценой зависимости от больших объёмов данных с трёхмерными аннотациями, которые зачастую труднодоступны. Кроме того, поскольку многие из этих методов моделируют суставы как независимые точки, они могут принимать физически некорректные позы, например, с нереалистичными углами вращения в суставах или плавающей длиной костей. Как следствие, дискриминативные методы демонстрируют высокую производительность в пределах распределения обучающих данных, но могут быть менее надёжными при столкновении с новыми, нетипичными позами. Но к настоящему моменту генеративные подходы чаще применяются в составе гибридных методов

В попытке преодолеть ограничения каждого из подходов были разработаны гибридные методы, которые комбинируют сильные стороны обеих парадигм. Типичная гибридная схема использует быстрый дискриминативный метод для получения грубой начальной оценки позы. Эта оценка служит качественной инициализацией для последующего итеративного процесса оптимизации на основе генеративной модели. В таком сценарии правдоподобие наблюдения, вычисляемое генеративной моделью, используется для проверки и уточнения гипотез, предложенных дискриминативной моделью. Другой способ интегрирует нейронные сети для прямой регрессии параметров модели, минуя этап явного обнаружения ключевых точек. Например, использование CNN, для прямого предсказания параметров модели SMPL [5], после чего эта модель проецируется на изображение, и функция потерь, оценивающая согласованность в двумерном пространстве, минимизируется для тонкой настройки параметров. Таким образом, гибридные подходы смягчают проблему низкой эффективности генеративных методов и компенсируют слабую обобщающую способность дискриминативных методов, обеспечивая баланс между скоростью, точностью и физической достоверностью итоговой позы.

Таким образом, современные исследования в области 3D HPE в значительной степени сфокусированы на дискриминативных подходах, чей прогресс был катализирован развитием глубокого обучения и благодаря которым также продвинулись гибридные подходы. Далее будут подробно рассмотрены ключевые разновидности дискриминативных методов.

Разновидности дискриминативного подхода

Оценка трёхмерной позы одного человека, в рамках дискриминативного подхода, может быть условно разделена на два основных подхода: прямую регрессию [6] и подъём (lifting) из 2D в 3D [7]. Кроме того, отдельного внимания заслуживает класс методов, основанных на тепловых картах [8].

Прямая регрессия предполагает непосредственное предсказание трёхмерных координат суставов по входным изображениям или видеопоследовательностям с помощью единой нейросетевой модели. Несмотря на концептуальную простоту, такой подход страдает от фундаментальной проблемы неоднозначности глубины: из одного 2D-изображения невозможно однозначно восстановить 3D-структуру сцены без дополнительных предположений или контекста. Это ограничивает точность и обобщающую способность прямых методов, особенно в условиях сложных поз или частичных окклюзий.

В отличие от прямой регрессии, подходы на основе подъёма (2D-to-3D lifting) разделяют задачу на два этапа. На первом этапе с использованием хорошо зарекомендовавших себя 2D-детекторов поз – таких как OpenPose [9], DeepCut [10] или BlazePose [11] – из входного изображения извлекается двумерный скелет человека. На втором этапе специализированная модель преобразует полученную 2D-позу в её трёхмерный аналог. Такой декомпозиционный подход позволяет эффективно использовать достижения в области 2D-оценки поз и значительно смягчает проблему неоднозначности глубины за счёт явного моделирования геометрических и кинематических ограничений человеческого тела. Эмпирически показано, что методы подъёма обеспечивают существенно более высокую точность, чем прямые регрессионные модели. Более того, использование временных последовательностей 2D-скелетов в качестве входа для лифтинга позволяет учитывать динамику движения и дополнительно улучшает качество 3D-реконструкции по сравнению с обработкой отдельных кадров.

Отдельную категорию составляют методы на основе тепловых карт. В таких подходах нейронная сеть не предсказывает координаты напрямую, а генерирует для каждой ключевой точки трёхмерную тепловую карту, обычно моделируемую как трёхмерное гауссово распределение с центром в истинном положении сустава. Окончательная 3D-поза восстанавливается на этапе постобработки путём поиска локальных максимумов на этих картах. Такой подход обеспечивает более устойчивую оценку позы за счёт мягкого представления пространственной неопределённости и часто используется в гибридных архитектурах, сочетающих преимущества регрессии и тепловых карт.

Таким образом, несмотря на теоретическую привлекательность прямых методов, современные решения в области 3D-оценки позы человека преимущественно опираются на двухэтапные стратегии с использованием 2D-детекторов и последующего подъема, а также на представления в виде тепловых карт для повышения точности.

Архитектуры

Графовые сверточные сети (*Graph Convolutional Networks, GCN*) [12,] [13] продемонстрировали высокую эффективность в задачах 3D-HPE, особенно в двухэтапных подходах, где на первом этапе из изображения извлекаются двумерные ключевые точки, а на втором – по ним восстанавливается трёхмерная поза. В этом контексте GCN позволяют явно моделировать пространственные зависимости между суставами, используя графовую структуру скелета для распространения информации между соседними узлами. Благодаря этому модель способна учитывать биомеханические ограничения, такие как фиксированная длина конечностей, и эффективно разрешать неоднозначности, вызванные окклюзиями или отсутствием глубины в 2D-наблюдениях. Однако стандартные GCN обладают принципиальным ограничением: их агрегация информации локальна и, как правило, ограничивается непосредственными соседями в графе. Это затрудняет моделирование долгосрочных взаимодействий между анатомически удалёнными суставами (например, между локтем и коленом), что может снижать точность восстановления сложных или нестандартных поз.

В то же время трансформеры, изначально разработанные для задач обработки естественного языка, показали выдающиеся результаты в компьютерном зрении благодаря своей способности моделировать глобальные зависимости через механизм внутреннего внимания (self-attention) [14], [15]. В контексте 3D-HPE трансформеры позволяют каждому суставу напрямую взаимодействовать со всеми остальными, независимо от их топологической близости в скелете. Это особенно ценно при наличии частичных наблюдений: даже если часть суставов скрыта, модель может корректно предсказать их 3D-положение, опираясь на глобальный контекст и изученные анатомические корреляции. Механизм внимания динамически взвешивает вклад каждого сустава в предсказание других, обеспечивая гибкое и семантически осмысленное представление пространственных отношений.

Однако трансформерные подходы к моделированию скелетных данных имеют существенные ограничения. Во-первых, при представлении скелета в виде плоской последовательности они игнорируют врождённую графовую топологию тела, что заставляет модель неявно восстанавливать анатомические связи из данных и снижает её обобщающую способность. Во-вторых, глобальный характер механизма внимания часто приводит к недостаточному учёту локальных временных закономерностей – таких как плавность траекторий суставов или краткосрочные кинематические зависимости, – что критично для точного восстановления динамики движений

Основные проблемы в задаче 3D HPE

Неопределенность глубины – самая серьёзная проблема. Одно двумерное изображение не имеет прямых признаков глубины, а это означает, что несколько различных 3D-поз могут проецироваться на один и тот же двумерный силуэт.

Эмпирически доказано, что добавление канала глубины к RGB-изображению повышает точность модели. Карту глубины можно получить из монокулярного RGB-изображения с помощью специализированных нейросетей (например, MiDaS [16] и Qwen Image [17]), создав таким образом RGB-D вход для последующего восстановления 3D-позы.

Части человеческого тела могут быть скрыты одеждой, другими объектами сцены или даже другими частями тела (самоокклюзия). Эта частичная видимость затрудняет точную оценку полной трёхмерной позы. Кроме того, большое разнообразие форм и размеров человеческого тела ещё больше усложняет задачу, поскольку модели должны быть обобщены для различных анатомических особенностей. Одним из решений проблемы окклюзии является использование набора данных, собранного

на основе многоракурсного подхода. В такой системе поза, невидимая с одного угла обзора, с высокой вероятностью будет видна с другого, что позволяет обеспечить более надежные результаты за счет межракурсного вывода. Кроме того, извлечению дополнительной информации из изображения способствуют гибридные архитектуры, сочетающие графовые сверточные сети (GCN) и трансформеры. GCN эффективно кодируют локальную информацию о суставах скелета, в то время как трансформеры компенсируют ограниченность локального поля восприятия, обеспечивая глобальное понимание позы. Современные подходы стремятся к комплексному подходу [18-20], где GCN используются для начального кодирования топологических зависимостей, а трансформеры уточняют предсказания, что повышает устойчивость к сложным сценариям.

Многие модели хорошо работают на контролируемых лабораторных наборах данных, но испытывают трудности при применении к реальным изображениям с различным фоном, освещением и ракурсами и требуют большого объема размеченных 3D-данных, что ограничивает их применимость. Для решения этой проблемы можно использовать альтернативные стратегии обучения. Один из вариантов, обучение без учителя, которое исследует внутренние закономерности в данных, используя, например, геометрическую самосогласованность или состязательное обучение для восстановления 3D-позы без 3D-разметки. Самообучение, как подвид обучения без учителя, использует специальные задачи, такие как многовидовая реконструкция или циклы проекции, для получения 3D-позы. Обучение со слабым контролем и перенос обучения позволяют эффективно использовать ограниченные или слабо размеченные данные, уменьшая зависимость от дорогостоящей полной разметки.

Дальнейшие работы

В перспективе планируется разработка системы для оцифровки трехмерной позы пользователя на основе изображения с одной RGB-камеры и последующего переноса полученной позы на модель виртуального аватара. Ключевыми требованиями к системе являются высокая скорость обработки и точность определения позы, что обусловлено необходимостью работы в реальном времени. Кроме того, система должна быть универсальной и работать исключительно с одной стандартной RGB-камерой, без необходимости использования дополнительных датчиков, нескольких камер или специализированного оборудования.

Среди рассмотренных подходов – генеративный, гибридный и дискриминативный – предпочтение отдано последнему, из-за меньшей вычислительной сложности. Для компенсации возможного снижения точности будет использоваться метод подъема 2D-позы в 3D (2D-to-3D lifting). Эмпирически доказано, что этот подход превосходит по точности прямую регрессию, а также обеспечивает модульность системы: 2D-детектор позы может быть заменен без необходимости модификации модуля реконструкции 3D-позы.

В качестве базовой архитектуры предлагается гибридное решение, сочетающее графовые сверточные сети (GCN) и трансформеры. Для обеспечения высокой производительности планируется применение ряда оптимизаций, таких как квантование весов, использование формата ONNX и выполнение вычислений на графических ускорителях (GPU). Это позволит сохранить высокую скорость работы при увеличении объема извлекаемой релевантной информации.

Для дальнейшего повышения точности оценки позы в входные данные планируется добавить карту глубины в качестве дополнительного канала изображения. Карта глубины будет извлекаться из исходного RGB-изображения с помощью отдельно обученной нейронной сети.

Выводы

В работе приведено математическое описание задачи 3D HPE: описаны входные (для изображения и видео), выходные данные для прямой регрессии, подъема и SMPL задачи. Было рассмотрено три подхода в задаче 3D HPE: 1) генеративный (с использованием модели); 2) дискриминативный (без использования модели); 3) гибридный. Можно сделать вывод, что актуальными являются гибридный и дискриминативный подход, а генеративный сейчас чаще используется в составе гибридного. Были рассмотрены разновидности дискриминативного подхода: прямая регрессия, подъем из 2D в 3D и тепловые карты, из которых наиболее точным является второй подход. Также описаны две наиболее эффективные архитектуры нейронных сетей для задачи 3D HPE: трансформер, графовая сеть и основанный на них гибридный подход.

Рассмотрены основные проблемы 3D HPE и предложены решения некоторых проблем. 1) Поскольку один силуэт может соответствовать разным 3D-конфигурациям, это создает неопределенность глубины. Решением служит использование RGB-D данных, где карта глубины генерируется нейросетями. 2) Скрытие частей тела одеждой или объектами, а также анатомическое разнообразие людей. Для борьбы с этим применяют многоракурсные системы, где поза, невидимая с одного ракурса, видна с другого. Также эффективны гибридные архитектуры, сочетающие графовые сверточные сети (GCN) для анализа локальных связей и трансформеры для глобального контекста. 3) Модели, обученные на лабораторных данных, часто плохо обобщаются на реальные условия. Чтобы снизить зависимость от дорогой 3D-разметки, используют альтернативные стратегии: обучение без учителя (геометрическая самосогласованность, состязательные сети), самообучение, слабый контроль и перенос обучения.

В результате можно сделать следующие выводы. 1) Генеративный подход позволяет обеспечить анатомическую достоверность позы и не требует обучающих данных, но вычислительно затратен и требует наличие точной трехмерной модели. Дискриминативный – его полная противоположность. 2) Подъем из 2D в 3D также обеспечивает модульность системы, позволяя заменять 2D-детектор при необходимости на другой, без замены всей остальной части системы. 3) Достичь высокой скорости обработки можно и без упрощения системы или замены подхода. Для этого можно воспользоваться квантованием весов нейронной сети и ее экспортом в формат ONNX.

Список литературы

1. El Kaid A., Baïna K. A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation / A. El Kaid, K. Baïna // Journal of Imaging. – 2023. – P. 8-15. – DOI: 10.3390/jimaging9120275
2. Киселев Ю.В., Богомолов И.А., Розалиев В.Л., Баклан В.А. Анализ подходов, методов и решений для детектирования позы человека. выбор инструмента для задачи определения эмоционального состояния человека по его позе // Современные наукоемкие технологии. – 2023. – №6. – С. 41-47. – DOI: https://doi.org/10.17513/snt.39629
3. Шергин И.А., Рыжов А.П. Проблема оценки позы человека: задачи, методы, решения. // Интеллектуальные системы. Теория и приложения. – 2024. – С. 69-84.
4. Gong W., Zhang X., Gonzàlez J., Sobral A., Bouwmans T., Tu C., Zahzah E. Human Pose Estimation from Monocular Images: A Comprehensive Survey / W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu // Sensors . – 2016. – P. 13-18. – DOI: 10.3390/s16121966
5. Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J. SMPL: A Skinned Multi-Person Linear Model / M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black. – 2015. – P.1-10. – DOI: 10.1145/3596711.3596800
6. Zheng C., Wu W., Chen C., Yang T., Zhu S., Shen J., Kehtarnavaz N., Shah M. Deep Learning-Based Human Pose Estimation: A Survey – 2023. – URL: https://arxiv.org/pdf/2012.13392 (дата обращения 15.09.2025).
7. Zhou L., Meng X., Liu Z., Wu M., Gao Z., Wang P. Human Pose-based Estimation, Tracking and Action Recognition with Deep Learning: A Survey / L. Zhou, X. Meng, Z. Liu, M. Wu, Z. Gao, P. Wang – 2023. – P. 4-10. – DOI: arXiv.2310.13039v1

8. Knap P. Human Modelling and Pose Estimation Overview / P. Knap . – 2024. – P. 1-4. – DOI: arXiv. 2406.19290v1
9. Cao Z., Hidalgo G., Simon T., Wei S., Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh. – 2017. – P. 1-5. – DOI: arXiv:1812.08008
10. Pishchulin L., Insafutdinov E., Tang S., Andres B., Andriluka M., Gehler P., Schiele B. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation / L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele. – 2015. – P.1-15. – DOI: arXiv:1511.06645
11. Bazarevsky V., Grishchenko I., Raveendran K., Zhu T., Zhang F., Grundmann M. BlazePose: On-device Real-time Body Pose tracking / V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann. – 2020. – P. 1-4. – DOI: arXiv:2006.10204
12. Shahjahan M., Hamza B. Flexible graph convolutional network for 3D human pose estimation / M. Shahjahan, B. Hamza. – 2024. – P. 1-7. – DOI: arXiv:2407.19077
13. Azizi N., Possegger H., Rodolà E., Bischof H. 3D Human Pose Estimation Using Möbius Graph Convolutional Networks / N. Azizi, H., Possegger, E. Rodolà. – 2022. – P. 1-13. – DOI: arXiv:2203.10554
14. Yang Z., Loo J. PyCAT4: A Hierarchical Vision Transformer-based Framework for 3D Human Pose Estimation – 2025. – URL: https://arxiv.org/pdf/2508.02806 (дата обращения 10.09.2025)
15. Zheng C., Zhu S., Mendieta M., Yang T., Chen C., Ding Z. 3D Human Pose Estimation with Spatial and Temporal Transformers / C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding. – 2021. – P. 1-13. – DOI: arXiv:2103.10455v3
16. Birkl R., Wofk D., Muller M. MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation – 2023. – URL: https://arxiv.org/pdf/2307.14460 (дата обращения 13.09.2025).
17. Wu C., Li J., Zhou J., Lin J., Gao K., Yan K. Qwen-Image Technical Report / C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan. – 2025. – P.1-23. – DOI: arXiv:2508.02324
18. Aouaidjiaa K., Lia A., Zhanga W., Zhanga C., 3D Human Pose Estimation via Spatial Graph Order Attention and Temporal Body Aware Transformer / K. Aouaidjiaa, A. Lia, W. Zhanga, C. Zhanga. – 2025. – P. 1-16. – DOI: arXiv.2505.01003v1
19. Mehraban S., Adeli V., Taati B. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network / S. Mehraban, V. Adeli, B. Taati. – 2023. – P. 1-11. – DOI: arXiv:2310.16288
20. Fu Y., Huang C., Li J., Kong H., Tian Y., Li H., Zhang Z. HDiffTG: A Lightweight Hybrid Diffusion-Transformer-GCN Architecture for 3D Human Pose Estimation / Y. Fu, C. Huang, J. Li, H. Kong, Y. Tian, H. Li, Z. Zhang. – 2025. – P. 1-8. – DOI: arXiv:2505.04276

References

1. El Kaid A., Baïna K. A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation / A. El Kaid, K. Baïna // Journal of Imaging. – 2023. – P. 8-15. – DOI: 10.3390/jimaging9120275
2. Kiselev Yu.V., Bogomolov I.A., Rozaliev V.L., Baklan V.A. ANALYSIS OF APPROACHES, METHODS AND SOLUTIONS FOR DETECTING HUMAN POSTURE. CHOOSING A TOOL FOR THE TASK OF DETERMINING A PERSON'S EMOTIONAL STATE BY THEIR POSTURE // Modern high-tech technologies. – 2023. – No. 6. – pp. 41-47. – DOI: https://doi.org/10.17513/snt.39629
3. Shergin I.A., Ryzhov A.P. The problem of assessing human posture: tasks, methods, solutions. // Intelligent systems. Theory and applications. – 2024. – Pp. 69-84.
4. Gong W., Zhang X., Gonzàlez J., Sobral A., Bouwmans T., Tu C., Zahzah E. Human Pose Estimation from Monocular Images: A Comprehensive Survey / W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu // Sensors . – 2016. – P. 13-18. – DOI: 10.3390/s16121966
5. Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J. SMPL: A Skinned Multi-Person Linear Model / M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black. – 2015. – P.1-10. – DOI: 10.1145/3596711.3596800
6. Zheng C., Wu W., Chen C., Yang T., Zhu S., Shen J., Kehtarnavaz N., Shah M. Deep Learning-Based Human Pose Estimation: A Survey – 2023. – URL: https://arxiv.org/pdf/2012.13392 (accessed 15.09.2025).
7. Zhou L., Meng X., Liu Z., Wu M., Gao Z., Wang P. Human Pose-based Estimation, Tracking and Action Recognition with Deep Learning: A Survey / L. Zhou, X. Meng, Z. Liu, M. Wu, Z. Gao, P. Wang – 2023. – P. 4-10. – DOI: arXiv.2310.13039v1
8. Knap P. Human Modelling and Pose Estimation Overview / P. Knap . – 2024. – P. 1-4. – DOI: arXiv. 2406.19290v1
9. Cao Z., Hidalgo G., Simon T., Wei S., Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh. – 2017. – P. 1-5. – DOI: arXiv:1812.08008

10. Pishchulin L., Insafutdinov E., Tang S., Andres B., Andriluka M., Gehler P., Schiele B. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation / L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele. – 2015. – P.1-15. – DOI: arXiv:1511.06645
11. Bazarevsky V., Grishchenko I., Raveendran K., Zhu T., Zhang F., Grundmann M. BlazePose: On-device Real-time Body Pose tracking / V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann. – 2020. – P. 1-4. – DOI: arXiv:2006.10204
12. Shahjahan M., Hamza B. Flexible graph convolutional network for 3D human pose estimation / M. Shahjahan, B. Hamza. – 2024. – P. 1-7. – DOI: arXiv:2407.19077
13. Azizi N., Possegger H., Rodolà E., Bischof H. 3D Human Pose Estimation Using Möbius Graph Convolutional Networks / N. Azizi, H., Possegger, E. Rodolà. – 2022. – P. 1-13. – DOI: arXiv:2203.10554
14. Yang Z., Loo J. PyCAT4: A Hierarchical Vision Transformer-based Framework for 3D Human Pose Estimation – 2025. – URL: https://arxiv.org/pdf/2508.02806 (accessed 09/10/2025)
15. Zheng C., Zhu S., Mendieta M., Yang T., Chen C., Ding Z. 3D Human Pose Estimation with Spatial and Temporal Transformers / C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding. – 2021. – P. 1-13. – DOI: arXiv:2103.10455v3
16. Birkl R., Wofk D., Muller M. MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation – 2023. – URL: https://arxiv.org/pdf/2307.14460 (accessed 09/13/2025).
17. Wu C., Li J., Zhou J., Lin J., Gao K., Yan K. Qwen-Image Technical Report / C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan. – 2025. – P.1-23. – DOI: arXiv:2508.02324
18. Aouaidjiaa K., Lia A., Zhanga W., Zhanga C., 3D Human Pose Estimation via Spatial Graph Order Attention and Temporal Body Aware Transformer / K. Aouaidjiaa, A. Lia, W. Zhanga, C. Zhanga. – 2025. – P. 1-16. – DOI: arXiv.2505.01003v1
19. Mehraban S., Adeli V., Taati B. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network / S. Mehraban, V. Adeli, B. Taati. – 2023. – P. 1-11. – DOI: arXiv:2310.16288
20. Fu Y., Huang C., Li J., Kong H., Tian Y., Li H., Zhang Z. HDiffTG: A Lightweight Hybrid Diffusion-Transformer-GCN Architecture for 3D Human Pose Estimation / Y. Fu, C. Huang, J. Li, H. Kong, Y. Tian, H. Li, Z. Zhang. – 2025. – P. 1-8. – DOI: arXiv:2505.04276

RESUME

O. A. Lyamtsev, I. I. Maksimenko

Overview and problems of using deep learning for three-dimensional estimation of human pose from a single image

The article is devoted to the problems and directions of deep learning in the problem of three-dimensional assessment of human posture. The paper provides a mathematical description of the problem of three-dimensional assessment of posture, three approaches are considered: 1) generative; 2) discriminative; 3) hybrid. The discriminative approach is considered in more detail and three more approaches are highlighted: 1) direct regression; 2) ascent from 2D to 3D; 3) heat maps.

The paper considers the key problems of three-dimensional assessment of human posture and suggests ways to solve them. The main difficulty is the uncertainty of depth in 2D images, since different 3D poses can be projected into identical silhouettes. To increase accuracy, RGB to RGB-D conversion is used with the addition of a depth map generated by neural networks. Another problem is occlusion, when body parts are hidden, as well as the anatomical diversity of people. The solution is multi-camera systems that capture the pose from different angles, as well as hybrid architectures that combine graph convolutional networks to analyze local joint connections and transformers for a global understanding of the scene.In addition, models that are effective in the laboratory often turn out to be unsuitable for real images due to the need for a large amount of labeled 3D data. An alternative is unsupervised learning strategies that explore the internal patterns of data, as well as self-learning, weak control, and learning transfer that reduce dependence on expensive markup.

The generative method creates anatomically accurate poses, but requires a lot of calculations, whereas the discriminative method is its exact opposite. The upgrade from 2D to 3D makes the system modular, making it easy to replace the 2D detector. High processing speed can be achieved without changing the architecture, using quantization of neural network weights and export to ONNX format.

РЕЗЮМЕ

О. А. Лямцев, И. И. Максименко

Обзор и проблемы использование глубокого обучения для трехмерной оценки позы человека по одному изображению

Статья посвящена проблемам и направлениям глубокого обучения в задаче трехмерной оценки позы человека. В работе дано математическое описание задаче трехмерной оценке позы, рассмотрено три подхода: 1) генеративный; 2) дискриминативный; 3) гибридный. Рассмотрен более подробно дискриминативный подход и выделены еще три подхода: 1) прямая регрессию; 2) подъём из 2D в 3D; 3) тепловые карты.

В работе рассмотрены ключевые проблемы трёхмерной оценки позы человека и предложены пути их решения. Основной сложностью является неопределенность глубины на 2D-изображениях, поскольку разные 3D-позы могут проецироваться в идентичные силуэты. Для повышения точности применяется преобразование RGB в RGB-D с добавлением карты глубины, генерируемой нейросетями. Другая проблема – окклюзия, когда части тела скрыты, а также анатомическое разнообразие людей. Решением служат многокамерные системы, фиксирующие позу с разных ракурсов, а также гибридные архитектуры, сочетающие графовые сверточные сети для анализа локальных связей суставов и трансформеры для глобального понимания сцены.Кроме того, модели, эффективные в лабораторных условиях, часто оказываются непригодными для реальных изображений из-за необходимости в большом объеме размеченных 3D-данных. Альтернативой являются стратегии обучения без учителя, исследующие внутренние закономерности данных, а также самообучение, слабый контроль и перенос обучения, снижающие зависимость от дорогостоящей разметки.

Генеративный метод создает анатомически точные позы, но требует больших вычислений, тогда как дискриминативный – его полная противоположность. Подъем из 2D в 3D делает систему модульной, позволяя легко заменять 2D-детектор. Высокую скорость обработки можно достичь без изменения архитектуры, используя квантование весов нейросети и экспорт в формат ONNX.

Лямцев Олег Алексеевич – стажер-исследователь ФГБНУ "Институт прикладной математики и механики", ул. Розы Люксембург, 74, Донецк, 283048, Россия, бакалавр кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», ул. Университетская, 24, Донецк, 283001, Россия, gelo2003@yandex.ru. *Область научных интересов*: компьютерное зрение, машинное обучение, нейронные сети. Число научных публикаций – 5.

Максименко Игорь Иванович – заведующий отделом теории управляющих систем ФГБНУ "Институт прикладной математики и механики", ул. Розы Люксембург, 74, Донецк, 283048, Россия, доцент кафедры компьютерных технологий физико-технического факультета ФГБОУ ВО «ДонГУ», ул. Университетская, 24, Донецк, 283001, Россия, igor.maksimenko_1967@mail.ru. *Область научных интересов*: теория автоматов, теория графов, системы искусственного интеллекта. Число научных публикаций – более 40.

Статья поступила в редакцию 22.09.2025