ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК: 004.932 **DOI 10.52928/2070-1624-2025-44-1-2-8**

ОЦЕНКА 3D-ПОЗЫ ЧЕЛОВЕКА НА ОСНОВЕ 2D КЛЮЧЕВЫХ ТОЧЕК

А. ДИН, канд. техн. наук, доц. О. В. НЕДЗЬВЕДЬ
(Белорусский государственный университет, Минск)

Предложена инновационная мало выборочная легковесная архитектура для решения задачи оценки 3D-позы человека на основе 2D ключевых точек. В рамках данного подхода введены специализированные обучаемые позиционные кодировки, предназначенные для задач трехмерной оценки позы, которые используются совместно с традиционными позиционными кодировками для представления входных данных. Архитектура метода включает многоуровневую обработку признаков и их адаптивное объединение с использованием механизма пространственного внимания, что позволяет усиливать релевантные признаки. Эксперименты, проведенные на стандартных тестовых наборах данных, подтвердили эффективность предложенного метода: достигнуто значение средней ошибки положения суставов (MPJPE) 42,1, что превосходит результаты существующих подходов.

***Ключевые слова**: оценка 3D-позы, обработка многоуровневых признаков, адаптивное объединение признаков, прогрессивное уточнение признаков, обучаемые позиционные кодировки.*

Введение. Оценка трехмерной позы человека на основе монокулярных видеопоследовательностей с использованием двумерных входных данных является одной из ключевых и при этом наиболее сложных задач в области компьютерного зрения. Данная задача требует решения ряда проблем, включая точное предсказание глубины, обеспечение временной согласованности [1], а также достижение устойчивости к частичным перекрытиям и сложным позам. Современные достижения в разработке пространственно-временных архитектур продемонстрировали значительные успехи в решении различных задач компьютерного зрения, предоставляя мощные инструменты для моделирования дальнодействующих зависимостей. Однако прямое применение подобных архитектур к задаче оценки трехмерной позы человека сопряжено с рядом существенных ограничений. К ним относятся недостаточная эффективность в захвате иерархических (многоуровневых) представлений признаков, неполноценное моделирование пространственно-временных взаимосвязей, а также отсутствие механизмов прогрессивного уточнения, направленных на повышение точности предсказаний.

На рисунке 1 представлена новая архитектура MAPS, которая интегрирует пространственно-временные методы с многоуровневым обучением признаков и механизмом прогрессивного уточнения. В рамках предложенного подхода введен адаптивный механизм слияния признаков, обеспечивающий динамический баланс их вклада на различных масштабах [3]. Кроме того, разработан инновационный модуль пространственно-временного внимания, эффективно захватывающий пространственные и временные зависимости. Проведенные исследования методом исключений подтвердили значимость каждого компонента архитектуры. Детальный анализ поведения модели в различных условиях, а также сравнение с современными методами демонстрируют превосходство предложенного подхода [4].

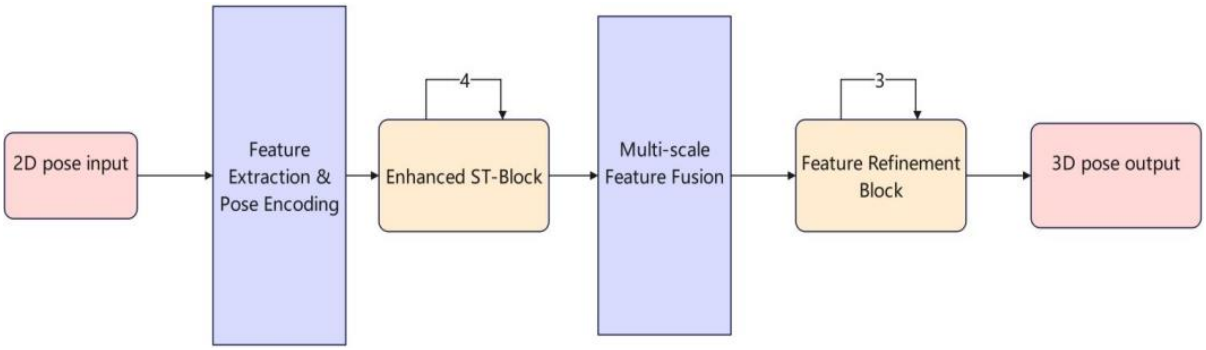

Рисунок 1. – Сетевая структура для оценки позы человека, преобразующая двумерные входные данные в трехмерную модель

Сочетание обучаемого параметрического кодирования с позиционным кодированием. В рамках данного эксперимента в задаче оценки трехмерной позы человека применяются обучаемые параметрические кодировки, предназначенные для извлечения структурной информации о скелете человека. Этот подход позволяет модели эффективно захватывать сложные характеристики скелетной структуры, повышая ее информативность. Параллельно с этим используются позиционные кодировки, которые предоставляют ключевым точкам явную пространственную информацию об их местоположении. Это позволяет модели лучше понимать относительное расположение узлов скелета и временные зависимости между последовательными кадрами. Комбинация указанных методов обеспечивает более точное моделирование структуры человеческого тела и его динамики, что приводит к повышению точности предсказаний и улучшению способности модели к обобщению. Математически это может быть выражено формулой [5]

$$F'(x, y, i) = F(x, y, i) + PE(x, y, i) + P(x, y, i). \quad (1)$$

Функция $PE(x, y, i)$ с помощью синусов и косинусов обеспечивает уникальное и плавное кодирование позиционной информации для элементов, расположенных в различных позициях последовательности или пространства. Данный подход позволяет сохранить непрерывность и однозначность представления позиционных данных. $P(x, y, i)$ представляет собой обучаемую позиционную кодировку, способную интегрировать информацию о структурных связях между суставами скелета или временных зависимостях между кадрами. Совместное использование данных методов способствует повышению точности модели и ее способности к обобщению, что подтверждается исследованиями [6].

Модуль многомасштабного объединения признаков (Multi-scale Feature Fusion Module) представляет собой архитектурный компонент, предназначенный для повышения способности модели к обработке сложных данных посредством интеграции признаков, извлечённых на различных масштабах.

На рисунке 2 показано, как модуль многомасштабного объединения признаков в задаче оценки позы человека обеспечивает эффективное извлечение информации о ключевых точках тела на различных пространственных и временных уровнях. Это позволяет модели учитывать как локальные, так и глобальные зависимости в данных, что способствует улучшению качества моделирования позы человека и повышению точности предсказаний [9].

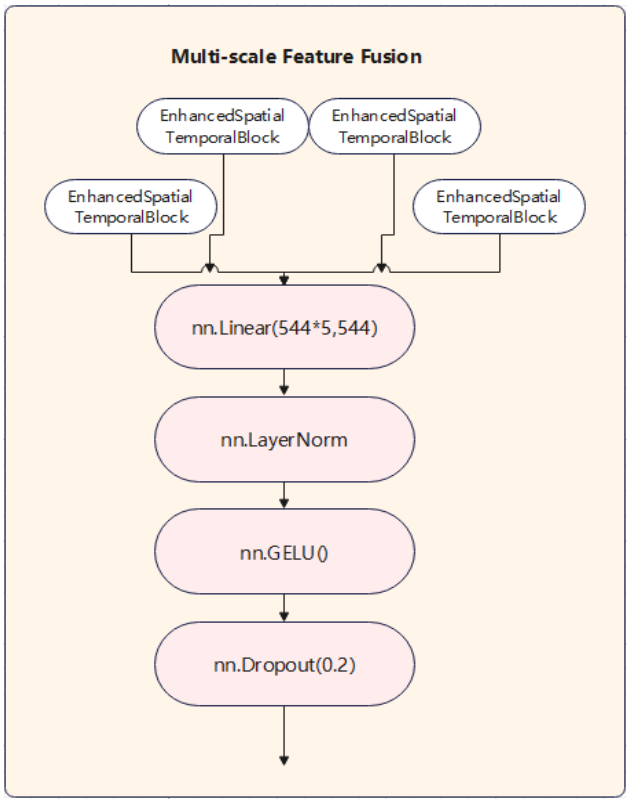

Рисунок 2. – Модуль многомасштабного объединения признаков

Входные данные представляют собой набор признаков, извлеченных на различных этапах или масштабах обработки. Каждый этап обеспечивает захват уникальных локальных и глобальных последовательных признаков, которые впоследствии объединяются для формирования многомасштабного представления.

Полученные признаки проходят через полносвязный слой, где выполняется их проекция в заданное пространство размерностей. Для повышения стабильности процесса обучения применяется операция нормализации [10]. На заключительном этапе применяется метод Dropout, направленный на улучшение обобщающей способности модели и минимизацию риска переобучения. Математическое описание данного процесса может быть представлено следующей формулой:

$$F_{fusion} = \sum_{i=1}^{n} \alpha_i F_i, \tag{2}$$

где F_i – объединенные признаки;

α_i – весовой коэффициент для i-го признака;

n – количество масштабов признаков.

Применение модуля многомасштабного объединения признаков позволяет выполнить проекцию объединенных признаков в общее пространство эмбеддинговой размерности. Объединенные признаки включают как временную, так и пространственную информацию, извлеченную на различных масштабах, что обеспечивает более полное и детализированное представление данных.

Функция SE-модуля (Squeeze-and-Excitation Module) заключается в повышении способности модели к селективному выделению наиболее информативных признаков по канальной размерности. Данный модуль адаптивно регулирует весовые коэффициенты, определяющие значимость каждого канала, что позволяет усиливать релевантные признаки и ослаблять менее значимые. В контексте задачи оценки позы человека применение SE-модуля особенно эффективно для фокусировки на характеристиках, связанных с 17 ключевыми суставами, что способствует более точному выделению пространственных и структурных особенностей скелета [7].

SE-модуль (Squeeze-and-Excitation Module) реализует операцию глобального усредненного объединения (global average pooling) по пространственному измерению входных признаков, что позволяет агрегировать глобальную контекстную информацию. Полученные данные затем обрабатываются через два последовательных полносвязных слоя, которые формируют нелинейные зависимости и вычисляют веса для каждого канала. На заключительном этапе размерность выходных данных восстанавливается до исходной глубины признаков, что обеспечивает согласованность с входными данными. Математическое описание данного процесса может быть представлено следующей формулой [8]:

$$s_c = \text{sigmoid}(W_2 \cdot \text{Re}\,LU(W_1 \cdot g(X))). \tag{3}$$

SE-модуль, интегрированный перед многослойным перцептроном MLP и остаточным соединением, повышает способность модели к селективному выделению информативных признаков между каналами. Здесь $g(X)$ – операция глобального усредненного объединения (global average pooling), которая агрегирует пространственную информацию, а W_1 и W_2 – весовые матрицы двух полносвязных слоев, отвечающих за вычисление значимости каналов. Такая архитектура позволяет эффективно усиливать релевантные признаки и подавлять шумовые компоненты.

Многослойный перцептрон MLP выполняет нелинейное преобразование входных признаков, что позволяет усилить способность модели к представлению сложных характеристик данных. Архитектура MLP включает два полносвязных слоя: первый слой увеличивает размерность признаков в четыре раза по сравнению с исходной, что способствует повышению способности модели к репрезентации сложных зависимостей и закономерностей в данных, а второй слой восстанавливает размерность до начального значения. Для обеспечения нелинейности и улучшения обобщающей способности модели применяется функция активации GELU (Gaussian Error Linear Unit) и метод регуляризации Dropout. Математически данный процесс можно описать выражением

$$MLP(x) = W_2 \cdot \text{Dropout}(GELU(W_1 \cdot x)), \tag{4}$$

где W_1 и W_2 представляют собой весовые матрицы линейного отображения, предназначенные для дальнейшего извлечения признаков из высокоразмерного пространства после применения механизма усиления внимания к каналам (SEModule).

Метод исключений. Для оценки вклада каждого модуля в общую производительность модели оценки позы человека было проведено несколько исключающих экспериментов. В рамках исследования последовательно исключались следующие компоненты: модуль уточнения признаков (Feature Refinement Block), модуль многомасштабного объединения признаков (Multi-scale Feature Fusion), блок двух типов кодирования (Two Encodings) и усовершенствованный пространственно-временной блок (Enhanced ST-Block). В таблице для каждого исключения фиксировалось значение средней ошибки положения суставов (MPJPE).

Таблица. – Результаты проведения исключающего эксперимента

Исключаемый модуль	Значение MPJPE
Без исключений (полная модель)	42,1
Feature Refinement Block	69,3
Multi-scale Feature Fusion	68,5
Two Encodings	60,1
Enhanced ST-Block	77,6

Модуль уточнения признаков (Feature Refinement Block) в задаче оценки позы человека реализует многослойную оптимизацию признаков, направленную на постепенное улучшение качества представления эмбеддингов суставов. Данный модуль использует линейные преобразования для захвата высокоразмерных зависимостей между суставами, а также применяет нормализацию на уровне слоев (LayerNorm) для повышения стабильности и согласованности признаков. Функция активации GELU обеспечивает усиление нелинейной выразительности модели, а механизм Dropout способствует снижению риска переобучения за счет регуляризации [11].

Концепция поэтапной оптимизации позволяет моделировать сложные пространственно-временные зависимости между суставами, обеспечивая устойчивость модели при обработке данных из множества кадров. В результате работы модуль формирует высококачественные представления признаков, что создает надежную основу для точного прогнозирования положения ключевых точек тела. Данный подход представляет собой эффективный метод повышения точности оценки позы человека за счет улучшения качества извлечения и обработки признаков.

Оценка позы человека требует точного моделирования сложных пространственно-временных взаимосвязей между суставами. Модуль уточнения признаков обеспечивает воспроизведение этих зависимостей посредством поэтапных преобразований и применения нелинейных функций активации [2]. На каждом уровне модуля выполняется обновление и оптимизация признаков, что способствует их постепенному приближению к целевому представлению. Такой подход позволяет достичь высокой точности в предсказании трехмерных координат суставов. Благодаря прогрессивной оптимизации признаков, предсказание положения ключевых точек в конечном прогнозе становится более точным. Математическое описание данного процесса может быть представлено формулой [12]

$$X_t = X_{t-1} + F_{refine}(X_{t-1}). \tag{5}$$

В формуле X_t представляет собой результат уточнения признаков X_{t-1}. Данный процесс реализуется посредством рекурсивного обновления, или пошагового уточнения, где функция F_{refine} выполняет поэтапное улучшение признаков. Процесс продолжается до достижения состояния сходимости или исчерпания заданного максимального числа итераций *T*. Основная цель данного подхода заключается в постепенном преобразовании начального приближенного результата в высококачественное итоговое представление, что обеспечивает повышение точности и устойчивости модели.

Обучение модели и результаты. Эксперимент проводился на наборе данных Human3.6M. В рамках эксперимента была определена оптимальная комбинация параметров, включающая скорость обучения $L_r = 0,004,$ количество блоков Enhanced ST-Block, равное 4, и значение параметра Feature Refinement Block, равное 3. Данная конфигурация продемонстрировала наилучшую производительность модели. Обучение проводилось на тренировочном наборе данных, состоящем из 10 000 образцов, с выполнением 150 итераций. Для анализа динамики обучения были построены графики функции потерь на тренировочном наборе и кривой средней ошибки положения суставов (MPJPE) на валидационном наборе данных (рисунок 3).

Результаты эксперимента показали, что на 145-й итерации модель достигла наилучшей производительности, при этом значение средней ошибки положения суставов (MPJPE) на валидационном наборе данных снизилось до 42,1, что представляет собой наилучший результат в рамках проведенного исследования. В процессе обучения наблюдалось постепенное снижение значения функции потерь, а кривая MPJPE на валидационном наборе данных демонстрировала устойчивую тенденцию к снижению, что свидетельствует о высокой сходимости модели и ее способности к обобщению при выбранной конфигурации параметров. Финальные результаты подтвердили эффективность предложенной конфигурации, предоставив надежное решение для задачи оценки позы человека.

В идентичных экспериментальных условиях (с использованием 10 000 обучающих и 200 валидационных образцов) было проведено обучение моделей LSTM, MHFormer, MyModel и PoseFormer на протяжении 150 итераций. После завершения обучения для каждой модели были построены графики кривых средней ошибки положения суставов (MPJPE) на валидационном наборе данных, что позволило провести сравнительный анализ их производительности (рисунок 4).

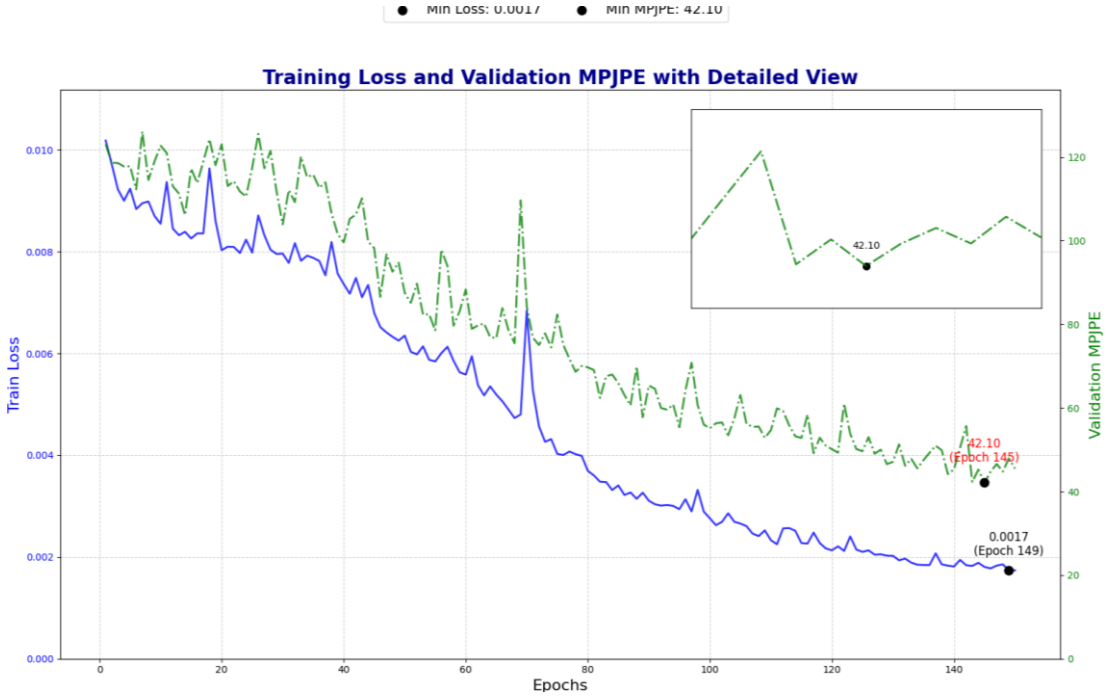

Рисунок 3. – Потери модели в данном эксперименте и MPJPE на валидационном наборе данных

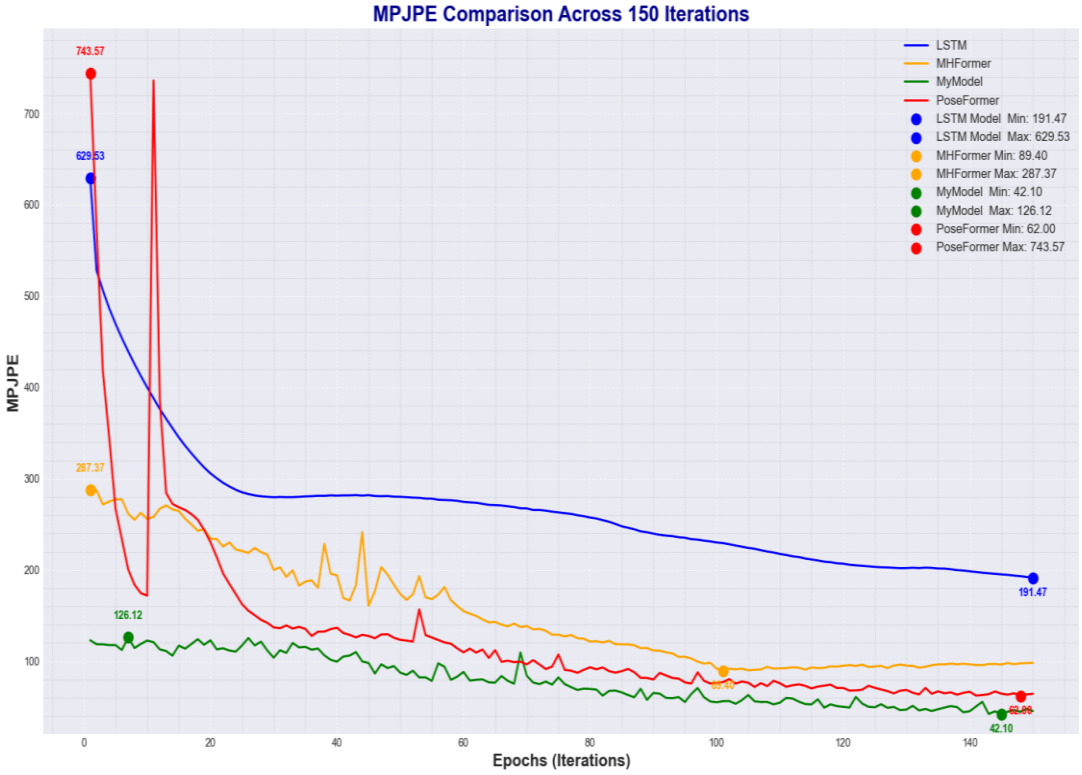

Рисунок 4. – Сравнение MPJPE различных моделей

Результаты эксперимента продемонстрировали, что модель PoseFormer достигла наилучшей производительности, показав минимальное значение ошибки. Модель MyModel заняла второе место, значительно превзойдя по точности модели LSTM и MHFormer. Полученные данные подтверждают высокую эффективность MyModel в задаче оценки позы человека, что подчеркивает ее практическую значимость и потенциал для дальнейшего применения.

Для оценки производительности моделей в задаче предсказания координат 17 суставов человека были выбраны три архитектуры: LSTM, PoseFormer и MyModel. Тестирование проводилось на 200 валидационных образцах, после чего была построена гистограмма, отражающая среднее значение ошибки положения суставов (MPJPE) для каждой модели по всем 17 суставам. Дополнительно для визуализации результатов предложенной модели MyModel была использована линейная диаграмма, демонстрирующая значение MPJPE для каждого отдельного сустава (рисунок 5).

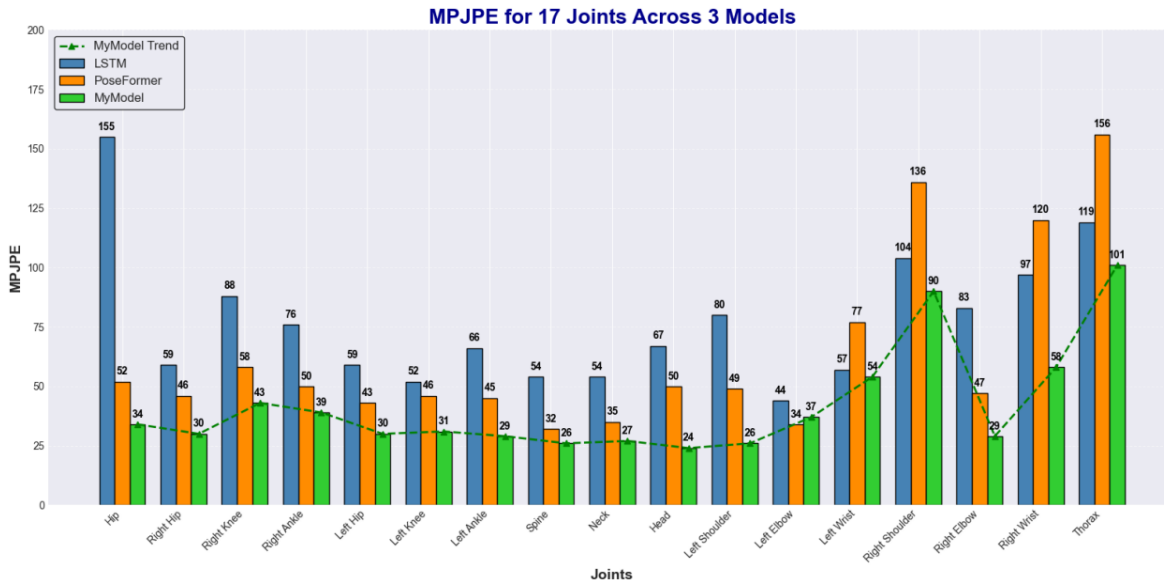

Рисунок 5. – Сравнение точности определения координат скелетных суставов в различных моделях

Результаты эксперимента демонстрируют, что средняя ошибка положения суставов (MPJPE) модели MyModel на всех суставах значительно ниже по сравнению с другими моделями. Особенно заметно превосходство в точности предсказания для сложных суставов, таких как запястья и лодыжки. Кроме того, распределение ошибок предложенной модели является более равномерным, что указывает на ее устойчивость и способность эффективно обрабатывать разнородные данные. Эти результаты свидетельствуют о значительных преимуществах модели MyModel в захвате сложных структурных зависимостей суставов человека и повышении точности преобразования из 2D в 3D, что имеет важное значение для разработки высокоточных методов оценки позы человека.

Заключение. В данной работе предложена комплексная архитектура для трехмерной оценки позы человека, направленная на решение ключевых задач в данной области. Основной научный вклад состоит во внедрении модуля многомасштабного объединения признаков, что улучшает извлечение контекстной информации и эффективно анализирует позы разной сложности; разработке усовершенствованного пространственно-временного блока с SE-модулем, что повышает качество моделирования связей между суставами и временных зависимостей; реализации прогрессивной стратегии уточнения с обучаемым позиционным кодированием, улучшающей качество признаков и точность предсказаний, особенно в сложных сценариях. Эксперименты на наборах данных, включая Human3.6M (MPJPE = 42,1), продемонстрировали высокую точность модели, ее устойчивость к окклюзиям и способность к обобщению в реальных условиях. Полученные результаты свидетельствуют, что сочетание многомасштабной обработки, механизмов внимания и учета неопределенности открывает перспективы для применения предлагаемой архитектуры в таких областях, как медицина, взаимодействие человека с машиной и захват движений.

ЛИТЕРАТУРА

1. 3D human pose estimation in video with temporal convolutions and semi-supervised training / D. Pavllo, C. Feichtenhofer, D. Grangier et al. // CVPR. – 2019. – DOI: 10.48550/arXiv.1811.11742.
2. Zhang T., Huang B., Wang Y. Object-occluded human shape and pose estimation from a single-color image // CVPR. – 2020. – DOI: 10.1109/CVPR42600.2020.00740.
3. Semantic Graph Convolutional Networks for 3D Human Pose Regression / L. Zhao, X. Peng, Y. Tian et al. // CVPR. – 2019. – DOI: 10.48550/arXiv.1904.03345.
4. Pavlakos G., Zhou X., Daniilidis K. Ordinal depth supervision for 3D human pose estimation // CVPR. – 2018. – DOI: 10.48550/arXiv.1805.04095.

5. Artzi Y., Zettlemoyer L. Weakly supervised learning of semantic parsers for mapping instructions to actions // Trans. Assoc. Comput. Linguist. – 2013. – Vol. 1. – P. 49–62. – DOI: 10.1162/tacl_a_00209.
6. Unite the people: Closing the loop between 3D and 2D human representations / Lassner, C., Romero, J., Kiefel, M. et al. // CVPR. – 2017. – DOI: 10.48550/arXiv.1701.02468.
7. Shape and Pose Estimation for Closely Interacting Persons Using Multi-View Images / Li, K., Jiao, N., Liu, Y. et al. // Computer Graphics Forum. – 2018. – Vol. 37, iss. 7. – P. 361–371. – DOI: 10.1111/cgf.13574.
8. End-to-end learning for self-driving cars / M. Bojarski, D. Del Testa, D. Dworakowski et al. // CVPR. – 2016. – DOI: 10.48550/arXiv.1604.07316.
9. Cross-modal self-attention network for referring image segmentation / L. Ye, M. Rochan, Z. Liu et al. // CVPR. – 2019. – DOI: https://doi.org/10.48550/arXiv.1904.04745.
10. Yeh R. A., Hu Y.-T., Schwing A. G. Chirality nets for human pose regression // CVPR. – 2019. – DOI: 10.48550/arXiv.1911.00029.
11. SRNet: Improving generalization in 3D human pose estimation with a split-and-recombine approach / A. Zeng, X. Sun, F. Huang et al. // ECCV. – 2020. – DOI: 10.48550/arXiv.2007.09389.
12. Liang J., Lin M. C. Shape-aware human pose and shape reconstruction using multi-view images // ICCV. – 2019. – DOI: https://doi.org/10.48550/arXiv.1908.09464.

REFERENCES

1. Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. *CVPR*. DOI: 10.48550/arXiv.1811.11742.
2. Zhang, T., Huang, B., & Wang, Y. (2020). Object-occluded human shape and pose estimation from a single-color image. *CVPR*. DOI: 10.1109/CVPR42600.2020.00740.
3. Zhao, L., Peng, X., Tian, Y., Kapadia, M., & Metaxas, D. N. (2019). Semantic Graph Convolutional Networks for 3D Human Pose Regression. *CVPR*. DOI: 10.48550/arXiv.1904.03345.
4. Pavlakos, G., Zhou, X., & Daniilidis, K. (2018). Ordinal depth supervision for 3D human pose estimation. *CVPR*. DOI: 10.48550/arXiv.1805.04095.
5. Artzi, Y., & Zettlemoyer, L. (2013). Weakly supervised learning of semantic parsers for mapping instructions to actions. *Trans. Assoc. Comput. Linguist.*, (1), 49–62. DOI: 10.1162/tacl_a_00209.
6. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M. J., & Gehler, P. V. (2017). Unite the people: Closing the loop between 3D and 2D human representations. *CVPR*. DOI: 10.48550/arXiv.1701.02468.
7. Li, K., Jiao, N., Liu, Y., Wang, Y., & Yang, J. (2018). Shape and pose estimation for closely interacting persons using multi-view images. *Computer Graphics Forum*, *37*(7), 361–371. DOI: 10.1111/cgf.13574.
8. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., … Zieba, K. (2016). End-to-end learning for self-driving cars. *CVPR*. DOI: 10.48550/arXiv.1604.07316.
9. Ye, L., Rochan, M., Liu, Z., & Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. CVPR. DOI: 10.48550/arXiv.1904.04745.
10. Yeh, R. A., Hu, Y.-T., & Schwing, A. G. (2019). Chirality nets for human pose regression. *CVPR*. DOI: 10.48550/arXiv.1911.00029.
11. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., & Lin, S. (2020). SRNet: Improving generalization in 3D human pose estimation with a split-and-recombine approach. *ECCV*. DOI: 10.48550/arXiv.2007.09389.
12. Liang, J., & Lin, M. C. (2019). Shape-aware human pose and shape reconstruction using multi-view images. *ICCV*. DOI: 10.48550/arXiv.1908.09464.

Поступила 27.03.2025

HUMAN 3D POSE ESTIMATION BASED ON 2D KEYPOINTS

A. DING, O. V. NEDZVED
(Belarusian State University, Minsk)

In the presented work, innovative low-sampling lightweight architecture is proposed to solve the task of 3D human pose estimation based on 2D key points. The approach introduces specialized trainable pose encodings designed for 3D pose estimation tasks, which are used in conjunction with traditional pose encodings to represent the input data. The architecture of the method includes multilevel feature processing and their adaptive association using a spatial attention mechanism, which allows to enhance relevant features. Experiments conducted on standard test datasets confirmed the effectiveness of the proposed method: a mean joint position error (MPJPE) value of 42.1 was achieved, which exceeds the results of existing approaches.

Keywords: *3D pose estimation, multi-level feature processing, adaptive feature aggregation, progressive feature refinement, learnable positional encodings.*