УДК 004.8
DOI: 10.26102/2310-6018/2025.49.2.036

Оценка человеческих поз по видеопотоку

М.А. Потенко✉

Московский авиационный институт (национальный исследовательский университет), Москва, Российская федерация

Резюме. В статье представлено исследование системы оценки позы человеческого тела, основанной на использовании двух нейронных сетей. Предложенная система позволяет определять пространственное расположение 33 ключевых точек, соответствующих основным сочленениям тела человека (кисти, локти, плечи, стопы и др.), а также строить маску сегментации для точного выделения границ человеческой фигуры на изображении. Первая нейронная сеть реализует функции детектора объектов и базируется на архитектуре Single Shot Detector (SSD) с применением принципов Feature Pyramid Network (FPN). Данный подход обеспечивает эффективное объединение признаков различного уровня абстракции и позволяет обрабатывать входные изображения размерностью 224×224 для последующего определения положения людей на кадре. Особенностью реализации является использование информации из предыдущих кадров, что способствует оптимизации вычислительных ресурсов. Вторая нейронная сеть предназначена для выделения ключевых точек и построения маски сегментации. Она также основана на принципах многомасштабного анализа признаков FPN, что обеспечивает высокую точность локализации ключевых точек и границ объекта. Сеть оперирует изображениями размерностью 256×256, что позволяет достичь необходимой точности определения пространственных координат. Предложенная архитектура характеризуется модульностью и масштабируемостью, позволяя адаптировать систему под различные задачи, требующие разного количества контрольных точек. Результаты исследования имеют широкое практическое применение в таких областях, как компьютерное зрение, анимация, мультипликация, системы безопасности и другие направления, связанные с анализом и обработкой визуальной информации.

Ключевые слова: нейронные сети, сверточные нейронные сети, машинное обучение, компьютерное зрение, оценка позы человека, ключевые точки, сегментация изображений.

Для цитирования: Потенко М.А. Оценка человеческих поз по видеопотоку. *Моделирование, оптимизация и информационные технологии.* 2025;13(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1920 DOI: 10.26102/2310-6018/2025.49.2.036

Human pose estimation from video stream

M.A. Potenko✉

Moscow Aviation Institute (National Research University), Moscow, the Russian Federation

Abstract. The article presents a study of a human body pose estimation system based on the use of two neural networks. The proposed system allows determining the spatial location of 33 key points corresponding to the main joints of the human body (wrists, elbows, shoulders, feet, etc.), as well as constructing a segmentation mask for accurate delineation of human figure boundaries in an image. The first neural network implements object detection functions and is based on the Single Shot Detector (SSD) architecture with the application of Feature Pyramid Network (FPN) principles. This approach ensures the effective combination of features at different levels of abstraction and enables the processing of input images with a resolution of 224×224 for subsequent determination of people's positions in a frame. A distinctive feature of the implementation is the use of information from previous frames, which helps optimize computational resources. The second neural network is designed for key point detection and segmentation mask construction. It is also based on the principles of multi-scale feature analysis

using FPN, ensuring high accuracy in localizing key points and object boundaries. The network operates on images with a resolution of 256×256, which allows achieving the necessary precision in determining spatial coordinates. The proposed architecture is characterized by modularity and scalability, enabling the system to be adapted for various tasks requiring different numbers of control points. The research results have broad practical applications in fields such as computer vision, animation, cartoon production, security systems, and other areas related to the analysis and processing of visual information.

Keywords: neural networks, convolutional neural networks, machine learning, computer vision, human pose estimation, keypoints, image segmentation.

For citation: Potenko M.A. Human pose estimation from video stream. *Modeling, Optimization and Information Technology*. 2025;13(2). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1920 DOI: 10.26102/2310-6018/2025.49.2.036

Введение

Анализ позы человека является одной из востребованных задач компьютерного зрения, полезной для широкого применения в медицине, спорте, робототехнике, анимации и системах видеонаблюдения. Основная цель данной задачи заключается в точной локализации ключевых точек человеческого тела (суставов, лица и других анатомических элементов), что позволяет эффективно моделировать положение и движения человека на изображении или в видеопотоке.

Анализ научных публикаций показывает, что существующие методы сталкиваются с рядом существенных ограничений. Как отмечается в работах [1, 2], основными проблемами являются: необходимость обработки больших объемов данных с высокой частотой кадров, наличие перекрытий между объектами, изменение условий освещения, сложные фоны и быстрые движения. Эти факторы существенно снижают эффективность существующих подходов, особенно при работе с несколькими людьми в режиме реального времени.

Особый интерес представляют исследования в области многоуровневых архитектур анализа признаков [3, 4] и механизмов внимания (attention mechanisms) [5], которые позволяют фокусироваться на наиболее значимых областях изображения. Также важным направлением является использование информации с предыдущих кадров видеопотока для минимизации вычислительных затрат за счет ограничения области поиска (ROI – Region of Interest) вокруг ранее обнаруженных объектов [6].

Проблема заключается в недостаточной эффективности существующих методов для одновременного обеспечения высокой скорости и точности выделения ключевых точек нескольких людей в реальном времени.

Целью данного исследования является разработка гибридного подхода к выделению ключевых точек нескольких человек по видеопотоку, сочетающего высокую скорость обработки с точностью локализации.

Материалы и методы

Для представления человеческого тела в представленном исследовании используется топология, состоящая из 33 ключевых точек. Данная топология соответствует другим топологиям, таким как COCO [7] и MPII [1], но включает дополнительные точки на руках, ногах и лице. Такая топология позволяет точно определять углы поворота отдельных частей тела, что особенно полезно для анализа движений в спорте, биомеханике и анимации. Следует отметить, что чем больше ключевых точек необходимо найти, тем больше нагрузка на вычислительные ресурсы. Поэтому в случаях, когда скорость работы важнее объема доступной информации, то

рекомендуется использовать топологию COCO из 17 точек или топологию MPII из 14 точек. Используемая топология представлена на Рисунке 1.

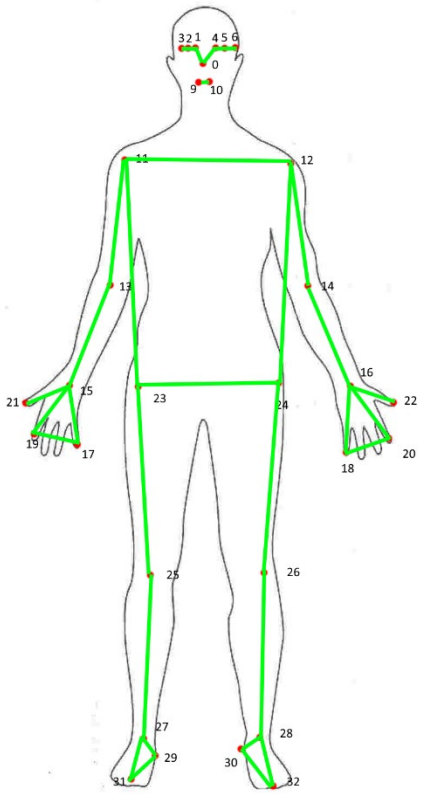

Рисунок 1 – Схема человеческого тела
Figure 1 – Scheme of human body

Обработка видеопотока проходит в несколько этапов:

1. На самом первом берется отдельный кадр видеопотока.
2. Полученный кадр преобразуется к размеру 224×224. При этом соотношение сторон изображения остается неизменным – с этой целью недостающие области изображения заполняются пустым пространством (черный цвет).
3. Выполняется нормализация изображения к диапазону [−1, 1], что улучшает сходимость модели нейронной сети.
4. Специальный детектор выполняет поиск областей (ограничивающих рамок), содержащих представления людей.
5. Выделение ключевых точек и маски сегментации из рамок.

Между соседними кадрами видеопотока, как правило, изменения минимальны. Поэтому поиск людей на изображении стоит выполнять в небольших областях относительно их предыдущего местоположения. Такие области называются ROI (Region of Interest). Эти области строятся вокруг ограничивающих рамок с предыдущего кадра и уточняются после выделения ключевых точек.

Полный проход детектора делается на первом кадре. На втором кадре ROI формируются вокруг обнаружений с первого кадра. На третьем кадре ROI предсказывается на основе предыдущих кадров. Если на каком-то из кадров предсказание ROI оказывается некорректным, то детектор приходится перезапускать полностью. В худшем случае детектор срабатывает на каждом кадре, а в лучшем – только один раз. Визуализация схемы работы представлена на Рисунке 2.

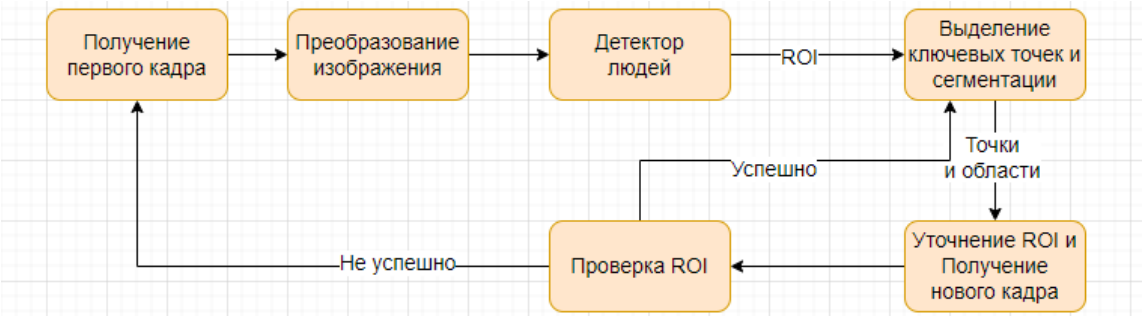

Рисунок 2 – Схема работы
Figure 2 – Scheme of work

Детектор людей

Детектор людей, используемый в представленном исследовании, реализован в виде сверточной нейронной сети, базирующейся на архитектуре Single Shot Detector (SSD) [7]. Основным преимуществом этой архитектуры является ее способность эффективно обрабатывать изображения для задач обнаружения объектов в реальном времени. В основе работы детектора лежит принцип Feature Pyramid Network (FPN) [8], который позволяет объединять высокоуровневые семантические признаки с низкоуровневыми пространственными характеристиками. Такой подход обеспечивает надежное обнаружение объектов разных масштабов – от крупных до мелких, что особенно важно при анализе изображений с людьми, находящимися на различных расстояниях от камеры. На вход детектора подаются нормализованные изображения размером 224×224 пикселей. Выбор данного разрешения обусловлен несколькими факторами. Во-первых, размер 224×224 является компромиссом между точностью обнаружения и вычислительной сложностью: он достаточно велик для сохранения значимых деталей, необходимых для корректного обнаружения людей, и в то же время мал для обеспечения высокой скорости обработки. Во-вторых, это стандартное разрешение, широко используемое в современных сверточных нейронных сетях, таких как ResNet [9] и VGG [10], что упрощает интеграцию с предобученными моделями. Перед подачей в сеть значения пикселей нормализуются до диапазона $[-1, 1]$, что улучшает сходимость модели во время обучения за счет стабилизации градиентов. В процессе обработки изображения детектор генерирует набор ограничивающих рамок, называемых якорями. Каждая рамка сопровождается значением достоверности, которое представляет собой вероятность наличия человека внутри данной области. Однако многие из этих якорей могут перекрываться или иметь низкую достоверность, что делает их анализ вычислительно затратным и избыточным. Для решения этой проблемы применяется метод немаксимального подавления (Non-Maximum Suppression, NMS) [11] с минимальным порогом перекрытия IoU (Intersection over Union), установленным на уровне 0,3 (выбрано как значение оптимальное для баланса между точностью и производительностью: более высокий порог может привести к чрезмерному удалению полезных рамок, а более низкий – к сохранению избыточных обнаружений). Метрика IoU (Intersection over Union) вычисляется как отношение площади их пересечения к площади объединения. На основе этой метрики соседние якоря с высокой достоверностью объединяются в одну ограничивающую рамку, что позволяет уменьшить количество финальных областей для анализа. После выполнения NMS полученные рамки масштабируются обратно к исходному разрешению изображения, чтобы координаты соответствовали реальным размерам объектов. Результатом работы детектора является набор ограничивающих рамок, каждая из которых задается координатами верхней левой и нижней правой границ прямоугольника. Эти рамки

представляют собой области изображения, содержащие потенциальных людей, и служат входными данными для последующих этапов анализа, таких как выделение ключевых точек и сегментация.

Выделение ключевых точек и сегментации

Архитектура сети по выделению ключевых точек и маски сегментации основана на двух ключевых принципах:

1. Генерация тепловых карт (карт достоверностей). Тепловая карта – это матрица, которая ставится в соответствие исходному изображению с определенным масштабированием. Для входного изображения размером 256×256×3 генерируются тепловые карты размером 64×64×1. Такой размер является компромиссом между точностью и производительностью. Однако для задач, требующих повышенной точности (например, локализации глаз), размер тепловой карты можно увеличить до 128×128 или выше. Значениями тепловой карты являются вероятности присутствия определенного сустава в соответствующей позиции. Например, для модели с 33 ключевыми точками будет сгенерировано 33 тепловые карты. При этом карты используются только для обучения нейронной сети, но не подаются в ее выходные данные, поскольку там нужны уже не карты, а ключевые точки.

2. Объединение признаков различных уровней абстракции. Как и современные архитектуры для классификации изображений, представленная модель использует последовательность сверточных слоев для извлечения признаков. Этот процесс преобразует низкоуровневые значения пикселей в высокоуровневые признаки, несущие информацию об изображении. Однако такой подход имеет недостаток: потеря пространственной точности. Чтобы решить эту задачу, применяются идеи Feature Pyramid Network (FPN) [8]. Согласно им, высокоуровневые признаки объединяются с низкоуровневыми детализированными признаками, что обеспечивает как точное определение наличия сустава, так и его локализацию.

В рамках архитектуры нейросети узел «Блок» обозначает повторяющуюся последовательность слоев, предназначенную для извлечения и объединения признаков. Каждый блок состоит из нескольких связок, каждая из которых включает: Depthwise-свёртки для получения мелких деталей изображения (например, края или текстуры) и обычные сверточные слои для их объединения для получения более сложных признаков. Каждая связка вычисляется и объединяется с результатами предыдущей связки с использованием операции сложения слоев, таким образом накапливаются различные признаки изображения. Первая связка в блоке объединяется с переданными в блок данными с применением операции MaxPooling. Такая структура позволяет сохранять пространственную информацию.

Результатом работы нейронной сети являются: маска сегментации, которая генерируется с тем же разрешением, что и исходное изображение (256×256). Для оптимизации вычислений маска предварительно вычисляется с уменьшенным вдвое разрешением (128×128) и затем расширяется до исходного размера. Тридцать три трёхмерные точки, представляют собой координаты ключевых точек тела в трёхмерном пространстве. При необходимости архитектура может быть дополнена вычислением параметра видимости для каждой ключевой точки. Это позволяет учитывать случаи, когда человек частично или полностью отсутствует на изображении. Для этого к признакам из последнего вычисленного блока применяется сигмоидная активация, и результаты преобразуются в вектор размерности 1×1, содержащий булево значение видимости.

Для обучения нейронной сети был выбран оптимизатор Adam [12], так как позволяет справляться с зашумлением в данных (ошибкам и неточностям при обучении).

В качестве функции потерь для выделения ключевых точек использовался алгоритм Huber Loss [13], также из-за его устойчивости к зашумлению. Поскольку задача сегментации также является задачей бинарной классификации (оценка принадлежности точки изображения к фону или телу), для сегментации была выбрана функция потерь Binary Crossentropy (Бинарная Кроссэнтропия) [14]. Данная функция штрафует модель за ошибки в каждом пикселе отдельно, что полезно для сегментации, где точность на уровне пикселей напрямую влияет на качество результата. Обучение проводится в 3 этапа: сначала фиксируются все веса, кроме тех, что нужны для генерации тепловых карт и сеть обучается выделять необходимые для них признаки. Затем фиксируются все слои кроме тех, в которых происходит выделение ключевых точек. На последнем этапе аналогично, но уже с выделением маски сегментации. Сеть можно обучать в один проход, но при таком подходе в ходе экспериментов результаты плохо сходились и сеть выдавала низкую точность. Для успешного схождения обучения понадобилось 230 эпох.

Визуализация архитектуры представлена на Рисунке 3.

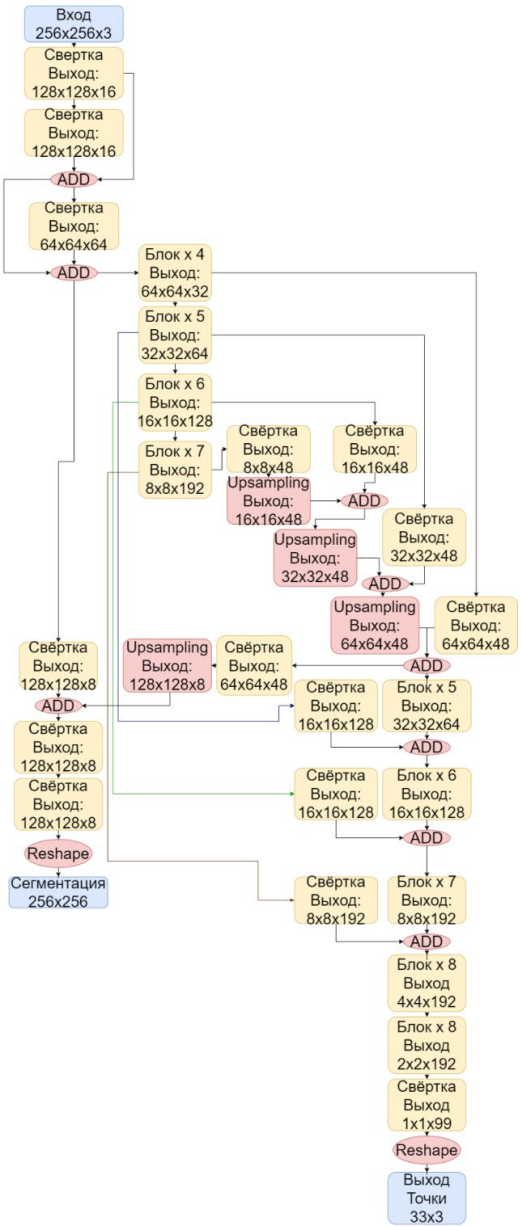

Рисунок 3 – Архитектура нейронной сети
Figure 3 – Neural network architecture

Результаты

Для оценки точности определения ключевых точек использовалась стандартная метрика mAP (mean Average Precision), основанная на показателе PCK (Percentage of Correct Key-points). Метрика PCK характеризует долю точек, расстояние которых между предсказанными и истинными координатами ключевых точек лежит меньше установленного значения. В данном исследовании установлено пороговое значение PCK, равное 0,7 при допустимой погрешности локализации точки в пределах 15 % от размеров тела человека. Это означает, что для достижения положительной оценки больше 70 % предсказанных точек должны отклоняться от истинных позиций максимум на 15 % относительно высоты и ширины человеческого тела.

Оценка качества сегментации также выполнялась с использованием метрики mAP, однако в качестве базового показателя применялся коэффициент IoU (Intersection over Union). Данная метрика вычисляется как отношение площади пересечения предсказанной и истинной масок сегментации к их объединенной площади. Для признания результата успешным установлен пороговый уровень IoU ≥ 0,7, что означает минимальное требование в 70 % совпадения реальных и предсказанных границ объектов.

По результатам тестирования на наборе данных [15] получены следующие показатели точности:

- определение ключевых точек: 86 %;
- сегментация: 77 %;
- обнаружение людей (детектор): 92 %.

Анализ производительности детектора показал, что полный перезапуск процедуры обнаружения объектов происходит в среднем через 10–15 кадров при частоте видеопотока 25 кадров в секунду. Частота вызова детектора напрямую зависит от динамики движения объектов: при увеличении скорости перемещения людей требуется более частый запуск детектора для поддержания точности трекинга. Однако это приводит к повышенному потреблению вычислительных ресурсов.

Важно отметить существенное ограничение предложенного подхода – задержка при обнаружении новых объектов в кадре. При использовании описанной стратегии интервал между последовательными запусками детектора может достигать 400–600 мс, что может быть критичным для некоторых приложений реального времени. Для повышения чувствительности к появлению людей в кадре можно увеличить частоту вызова детектора за счет дополнительной вычислительной нагрузки, особенно в сценах с интенсивным движением объектов.

Заключение

В данной статье представлен гибридный подход к выделению ключевых точек человеческого тела по видеопотоку, сочетающий высокую скорость обработки с точностью локализации. Предложенный метод решает проблему эффективности существующих подходов, особенно при работе с несколькими людьми в режиме реального времени. Использование многоуровневой архитектуры нейронной сети, основанной на принципах Feature Pyramid Network (FPN) и механизмах внимания, позволяет надежно обнаруживать объекты разных масштабов. Оптимизация вычислительных затрат достигается за счет применения ROI (Region of Interest) для ограничения области поиска новых объектов. Результаты тестирования демонстрируют высокую точность локализации ключевых точек (86 %), сегментации (77 %) и обнаружения людей (92 %). Однако в исследовании выявлен ряд ограничений, таких как задержка при обнаружении новых людей в кадре, которая может достигать 400–600 мс, что является критичным для некоторых приложений реального времени. Частота вызова

детектора зависит от динамики движения объектов, что увеличивает вычислительную нагрузку в сценах с интенсивным движением.

Таким образом, предложенный подход демонстрирует значительный потенциал для решения задачи анализа позы человека в реальном времени.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Andriluka M., Pishchulin L., Gehler P., Schiele B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, 23–28 June 2014, Columbus, OH, USA*. IEEE; 2014. P. 3686–3693. https://doi.org/10.1109/CVPR.2014.471
2. Newell A., Yang K., Deng J. Stacked Hourglass Networks for Human Pose Estimation. In: *Computer Vision – ECCV 2016: 14th European Conference: Proceedings: Part VIII, 11–14 October 2016, Amsterdam, The Netherlands.* Cham: Springer; 2016. P. 483–499. https://doi.org/10.1007/978-3-319-46484-8_29
3. Zhao Zh.-Q., Zheng P., Xu Sh.-T., Wu X. Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*. 2019;30(11):3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865
4. Zhang F., Zhu X., Ye M. Fast Human Pose Estimation. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15–20 June 2019, Long Beach, CA, USA.* IEEE; 2019. P. 3512–3521. https://doi.org/10.1109/CVPR.2019.00363
5. Guo M.-H., Xu T.-X., Liu J.-J., et al. Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media.* 2022;8(3):331–368. https://doi.org/10.1007/s41095-022-0271-y
6. Liu W., Anguelov D., Erhan D., et al. SSD: Single Shot MultiBox Detector. In: *Computer Vision – ECCV 2016: 14th European Conference: Proceedings: Part I, 11–14 October 2016, Amsterdam, The Netherlands.* Cham: Springer; 2016. P. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
7. Lin T.-Yi, Maire M., Belongie S., et al. Microsoft COCO: Common Objects in Context. In: *Computer Vision – ECCV 2014: 13th European Conference: Proceedings: Part V, 06–12 September 2014, Zurich, Switzerland.* Cham: Springer; 2014. P. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
8. Lin T.-Yi, Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature Pyramid Networks for Object Detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21–26 July 2017, Honolulu, HI, USA.* IEEE; 2017. P. 936–944. https://doi.org/10.1109/CVPR.2017.106
9. He K., Zhang X., Ren Sh., Sun J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27–30 June 2016, Las Vegas, NV, USA.* IEEE; 2016. P. 770–778. https://doi.org/10.1109/CVPR.2016.90
10. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv. URL: https://doi.org/10.48550/arXiv.1409.1556 [Accessed 25th March 2025].
11. Neubeck A., Van Gool L. Efficient Non-Maximum Suppression. In: *18th International Conference on Pattern Recognition (ICPR'06), 20–24 August 2006, Hong Kong, China.* IEEE; 2006. P. 850–855. https://doi.org/10.1109/ICPR.2006.479
12. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization. In: *3rd International Conference on Learning Representations, ICLR 2015, 07–09 May 2015, San Diego, CA, USA*. 2015. https://doi.org/10.48550/arXiv.1412.6980

13. Charbonnier P., Blanc-Féraud L., Aubert G., Barlaud M. Two Deterministic Half-Quadratic Regularization Algorithms for Computed Imaging. In: *Proceedings of 1st International Conference on Image Processing, 13–16 November 1994, Austin, TX, USA.* IEEE; 1994. P. 168–172. https://doi.org/10.1109/ICIP.1994.413553
14. Goodfellow I., Bengio Yo., Courville A. *Deep Learning*. Cambridge: MIT Press; 2016. 800 p.
15. Потенко М.А. Применение синтетических данных в обучении нейронных сетей для оценки поз человека. В сборнике: *Экспериментальные и теоретические исследования в современной науке: сборник статей по материалам CVIII международной научно-практической конференции, 25 декабря 2024 года, Новосибирск, Россия.* Новосибирск: Сибирская академическая книга; 2024. С. 11–17. Potenko M. Application of Synthetic Data in Training Neural Networks for Human Pose Estimation. In: *Eksperimental'nye i teoreticheskie issledovaniya v sovremennoi nauke: sbornik statei po materialam CVIII mezhdunarodnoi nauchno-prakticheskoi konferentsii, 25 December 2024, Novosibirsk, Russia.* Novosibirsk: Sibirskaya akademicheskaya kniga; 2024. P. 11–17. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Потенко Максим Алексеевич, аспирант, Московский авиационный институт (национальный исследовательский университет), Москва, Российская федерация.
e-mail: potenkog@gmail.com
ORCID: 0009-0008-5222-2664

Potenko M. Alexeevich, Postgraduate, Moscow Aviation Institute (National Research University), Moscow, the Russian Federation.

Статья поступила в редакцию 22.04.2025; одобрена после рецензирования 20.05.2025; принята к публикации 04.06.2025.

The article was submitted 22.04.2025; approved after reviewing 20.05.2025; accepted for publication 04.06.2025.