



Using domain adaptation for the human pose estimation task

Andrei Sergeevich **Tokarev**¹, Ilia Mikhailovich **Voronkov**²

^{1,2} Moscow Institute of Physics and Technology, Moscow, Russia

¹✉ tokarev.as@phystech.su

Abstract. The article studies domain adaptation algorithms for the task of recognizing key points on the human body for the purpose of using them in sports, when it is necessary to increase the accuracy of recognition and reduce the labor intensity of manual data labeling. The result of the work is an algorithm for iterative adaptation of the model on its own pseudo-labels. It is experimentally shown that the method allows obtaining a more effective final neural network model in comparison with conventional additional training. (*Linked article texts in English and in Russian*).

Key words and phrases: Keypoints, human pose estimation, unsupervised domain adaptation

2020 *Mathematics Subject Classification:* 68T05; 68T45

For citation: Andrei S. Tokarev, Ilia M. Voronkov. *Using domain adaptation for the human pose estimation task*. Program Systems: Theory and Applications, 2025, **16**:4(67), pp. 23–50. (*In English, in Russian*). https://psta.psiras.ru/read/psta2025_4_23-50.pdf

Introduction

In recent years, among the rapidly developing areas of machine learning and computer vision, human pose estimation (HPE) stands out, which is actively used in sports, based on the recognition of key points on the human body (Keypoint Detection). On its basis, a semi-automatic offside detection system has been created, which has been used in international football competitions since 2022.

A large number of sports make it possible to fully realize the potential of human pose estimation technology for solving complex dynamic problems based on video analytics, as well as during online competitions. For each individual problem statement, the accuracy of human pose estimation is of great importance, and it is also necessary to take into account the specifics of the sports. The effectiveness of the technology is unthinkable without a large number of high-quality labeled data sets for subsequent training of neural network models.

Among the abundance of video recordings of sporting events, very few are labeled by people, since this takes a lot of time. Unsupervised domain adaptation methods make it possible to do without data labeling, while improving the quality of the model. These methods bridge the gap between source and target data by equalizing data distributions, adapting features, and applying trained models to a new domain with minimal additional processing.

The described path makes models more flexible and able to work effectively in a variety of settings without requiring significant new data or reworking the architecture. Ultimately, using domain adaptation methods not only speeds up the implementation process, but also reduces development costs by reducing the need for a full training cycle for a new model.

1. Related work

1.1. Human pose estimation

Pose estimation models require defining the topology of the points that will make up the pose. Historically, the pose was defined by the dataset on which the model is trained, but since 2015, the most common has been the

17-point topology proposed by researchers at Microsoft [1]. All newer “skeleton” descriptions are extensions of this and specify the position of the limbs.

In the world of human body keypoint recognition, there are two approaches to solving problems today: top-down and bottom-up. The first approach requires an additional detector that localizes a person in the image. Then, the necessary keypoints are recognized in the localized area and a skeleton is assembled. This ensures high prediction accuracy, but makes the architecture vulnerable to errors at the level of human detection. The second approach, on the contrary, recognizes all keypoints in the image at once, and only then solves the problem of correctly distributing the points across the “skeletons”. Models using this approach require a high-resolution image as input, which increases the consumption of resources for their training.

Due to resource limitations, the study considered top-down models such as the Simple Coordinate Classification (SimCC) [2] algorithm with the ResNet neural network as the backbone, and the RTMPose [3] model. Both architectures are described in the MMPose project from OpenMMLab [4], which allowed universalization of the code for the training process and adaptation of different models.

1.2. Domain adaptation

Domain adaptation is a special case of the transfer learning problem: when the problem statement remains constant, and only the data set (domain) changes. Unsupervised domain adaptation is commonly used when only raw data without annotations is available. It allows a model trained on a large and universal domain to maintain high accuracy of predicted results and performance on a previously unknown dataset, avoiding the cost of additional training. This creates the ability to quickly configure a neural network to work on a highly specialized task, while saving enormous resources for analyzing and labeling new data.

Existing methods for adapting human pose estimation models to an unlabeled domain use additional neural networks, such as an adversarial regressor in Regressive Domain Adaptation [5] or style transfer models and the average teacher algorithm in Unified Framework [6]. This increases the computational complexity of the method and also requires manipulations

with internal layers of the architecture, which complicates software implementation.

There is also a simpler but effective method called Progressive Unsupervised Learning[7], which works without adding additional neural networks. The basic idea behind PUL is to iteratively train the model on the best of its predictions, resulting in improved performance accuracy. In such a case, the model predictions are called a pseudo-label or a set of pseudo-labels, and the best ones are selected by a chosen filtering function. Thus, the problem of the method is reduced to the correctness of choosing such a function so that the algorithm does not turn out to be computationally complex and the neural network demonstrates improved performance.

In the current study, we have shown that for the task of recognizing key points on the human body, there exists a filtering function that allows us to use the PUL method to adapt neural network models to new domains. The goal of the proposed function is to filter out noisy predictions using a measure of the model's confidence in its predictions and an adjustable threshold value. All predictions for which the average skeletal confidence crosses the threshold fall into a set of pseudo labels on which the model is further trained in the current iteration.

2. Experiment

The practical part of the study involved several steps such as data preparation, model training and then comparing their results across states. The first stage included data preparation: data collection, processing and partial labeling. Next, model adaptation and training was carried out based on the prepared data. After the training process was completed, the results were analyzed and compared to determine the effectiveness of the domain adaptation method compared to traditional model training.

2.1. Semi-automatic data labeling system

A semi-automated pose-markup system (<https://github.com/andrwtokar/pose-markup>) was created to annotate a set of videos from the target data domain. An example of the program's operation can be seen at Figure 1.



(a) Original image (b) Automatic marking (c) Adjusted markup

FIGURE 1. Example of the data labeling system operation.

The basis of the system was the idea of automatic markup using the BlazePose [8] model of recognizing key points on the human body with subsequent correction of the result by an expert. For the first part, the public API in Python provided by the MediaPipe project was used. Prediction results were corrected using image processing tools provided by OpenCV library.

A 17-point COCO [1] topology was chosen as the base topology. For each frame the annotation is written in JSON format with the possibility of further aggregation of the results of several frames into one JSON file.

Analyzing the data labeling performed by the system showed how much it reduces the amount of resources consumed. It was noticed that on average 6.5 points out of 17 were changed on the frame. And most often it was necessary to correct coordinates of the points of the hand, located farther from the camera, and the head. More detailed distribution of the changes made can be seen on Figure 2.

The time of automatic markup of a video with 420 frames took about 2 minutes. Point correction took an average of 3 minutes per frame. In total, it took 24 hours of work to markup 420 frames of target data. If 2500 frames were marked up, it would have taken more than 120 hours.

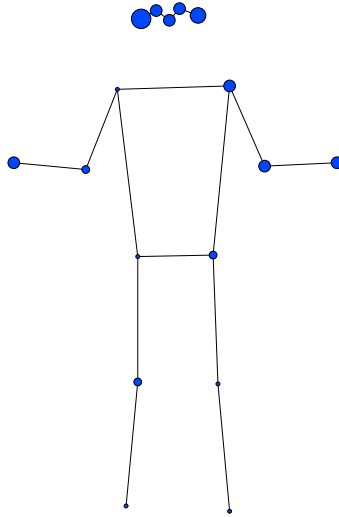


FIGURE 2. Schematic representation of changes in points made by the expert.

2.2. Data

Two datasets were chosen for the study: a universal dataset on which the model will be baseline trained, and a target dataset to which we will adapt the model.

The COCO dataset [1] was chosen as the first one. Due to its large size and limited resources for training, it was decided to reduce the volume. From more than 250 thousand poses, only those with all 17 keypoints consistent with the COCO topology were selected. As a result, a universal set consisting of 8 thousand annotated poses represented in 6 thousand images was obtained.

The small second dataset was created manually. For this purpose, two videos of single training sessions of boxer athletes of 2500 frames each were taken. The first one was used for the iterative method of model adaptation. From the second video, 420 frames were selected, which were further labeled using the pose-markup system created for this purpose, and used for model refinement and metrics extraction. An example of the resulting data can be seen in Figure 3.



FIGURE 3. Examples of images of the target domain.

2.3. Metrics

Pose estimation articles use average precision (AP) metric based on object keypoint similarity (OKS) metric as a basic metric. However, it requires knowledge of the standard deviation of keypoints for the selected dataset, which requires a large sample size and resources to collect statistics.

Among many similar metrics that evaluate the error between predicted and real points, the metric percentage of correct keypoints (PCK) stands out. The advantage of this metric is that it can be applied to both two-dimensional and three-dimensional coordinates. The error is represented as the Euclidean distance between points and compared to a threshold value. The mathematical formula is as follows:

$$(1) \quad PCK = \frac{\sum_{i=1}^n \text{bool}(d_i < \text{threshold} * \text{body_height})}{n},$$

where

- d_i is the distance between the predicted and the correct point,
- threshold is the threshold set by the researcher,
- body_height is the height of the rectangle the person is inside of,
- $\text{bool}(\ast)$ is the boolean condition, returns 1 if true and 0 otherwise,
- n is the sample size.

The PCK metric with thresholds of 0.05, 0.25 and 0.5 was used in the experiment.

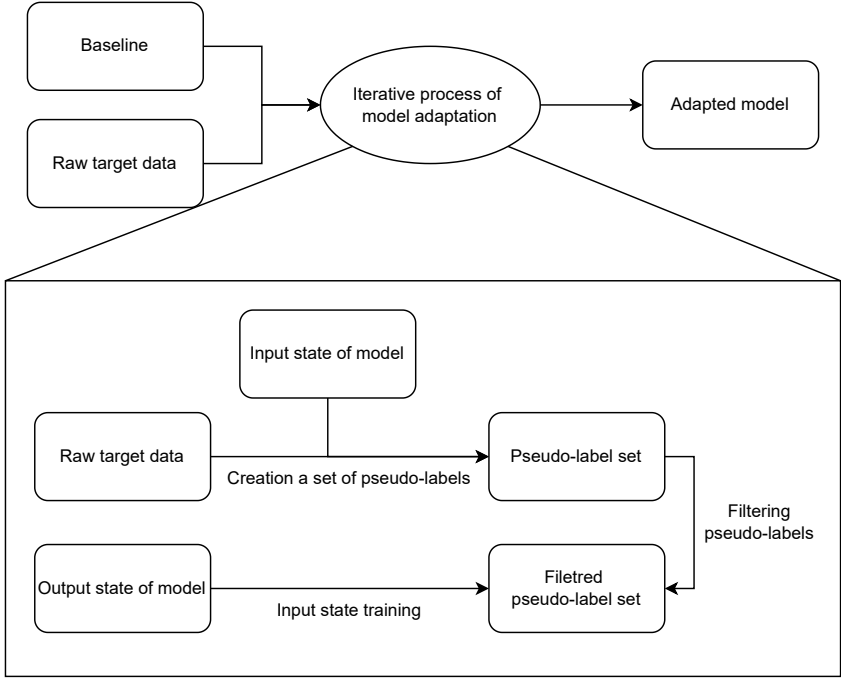


FIGURE 4. Model Adaptation Scheme

2.4. Results

The experiment compared three states of the algorithm:

- (1) Base case. When the model is trained only on the universal dataset.
- (2) Adapted state. We applied 10 iterations of the domain adaptation algorithm to the base model. The model adaptation scheme is presented in Figure 4.
- (3) Pre-trained state. The base model was additionally trained on the marked-up data from the target domain. The scheme of additional model training is presented in Figure 5.

The pseudolabel filtering function described above involves obtaining a confidence value for the predicted pose. This value is obtained by averaging the confidence value over all 17 topology points.

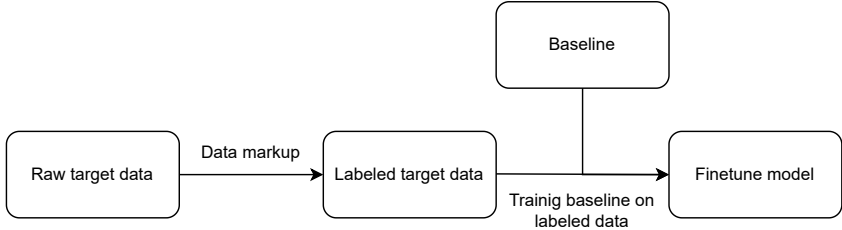


FIGURE 5. Scheme for additional model training

TABLE 1. Comparative statistics of multiple states for the SimCC algorithm.

	PCK@0.05	PCK@0.25	PCK@0.5
Baseline	0.024	0.429	0.758
Adapted model	0.037	0.445	0.792
Finetune model	0.0062	0.411	0.76

This method on the first iterations discarded more than 75 percent of the data used for adaptation. At 10 iterations, the number of noisy poses was close to zero, indicating that the quality of the model prediction improved with increasing number of iterations.

The iterative process was stopped after 10 steps. If more steps were selected, the iterative process could be stopped either when a pre-selected accuracy was reached or when there was no change in the filtered set of pseudo-labels.

Neural networks were trained on the Google Colaboratory platform using Tesla T4 GPUs with 15 GB of memory. Metrics were evaluated locally on a laptop with an NVidia MX250 GPU with 2 GB of memory.

The table shows the results obtained for the SimCC algorithm with confidence threshold equal to 0.6. The results in Table 1 show an improvement of the PCK metric not only relative to the baseline state, but also relative to the additionally trained state. It should be noted that the total time spent on adaptation was no more than 80 minutes, which is much less than the time spent on markup the same data.

TABLE 2. Comparative statistics of several states for the RTMPose model.

	PCK@0.05	PCK@0.25	PCK@0.5
Baseline	0.282	0.882	0.976
Adapted model	0.145	0.652	0.709
Finetune model	0.076	0.709	0.922

TABLE 3. RTMPose model prediction results without considering fuzzy key points.

	PCK@0.25	PCK@0.5
Baseline	0.944	0.988
Adapted model	0.964	1.0

The baseline version of RTMPose showed much better results than the previous model. The Table 2 shows that among the whole experiment, the baseline gives the best results.

To explain the reasons for the errors, let us draw a histogram of the distribution of correct predictions by topological points. As can be seen in Figure 6 and Figure 7, the number of correct coordinates of the upper limb points drops to almost zero and introduces a large error into the metric value. The quality of prediction for ankle points also drops. The described points of boxers are more often in motion than others and are blurred on some frames of the video recording.

In the current version of the filtering function, these points may have entered the training sample at the selected iteration due to the filtering mechanism and averaging the confidence over all points to obtain pose confidence.

To compare the performance of the algorithm without considering blurring, we exclude these 6 points from the sample. As can be seen in Table 3, the adapted model shows higher prediction quality relative to the base case. In the future, in order to overcome the errors introduced by the blurred parts of the image, it is necessary to modify the filtering function.

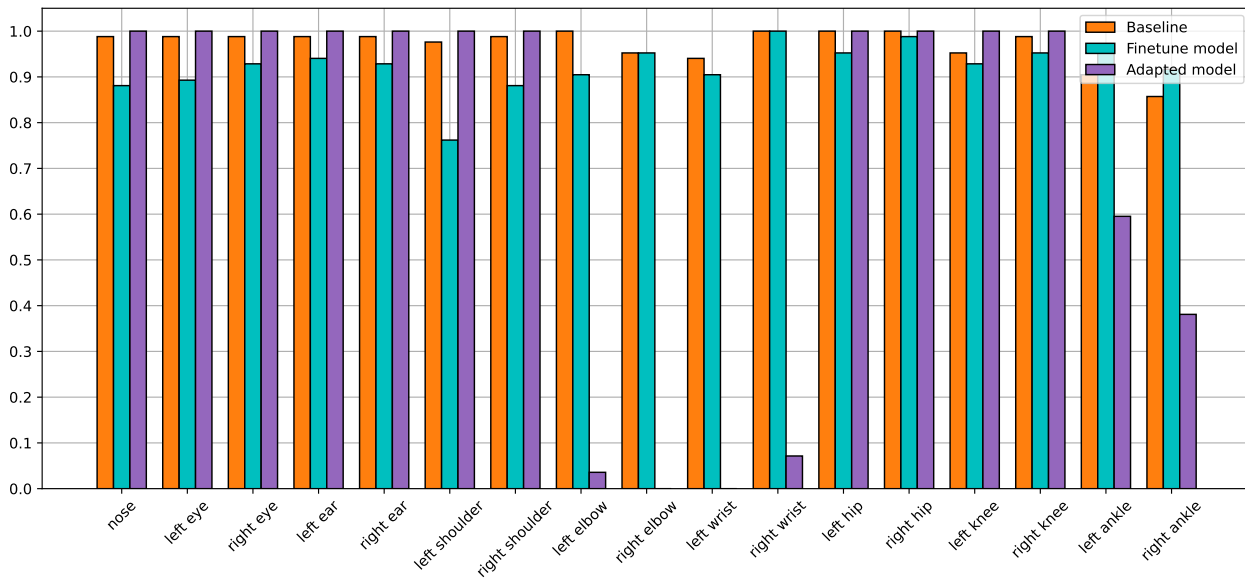


FIGURE 6. Distribution of correctly predicted points over the whole topology for the RTMPose model, related PCK@0.5

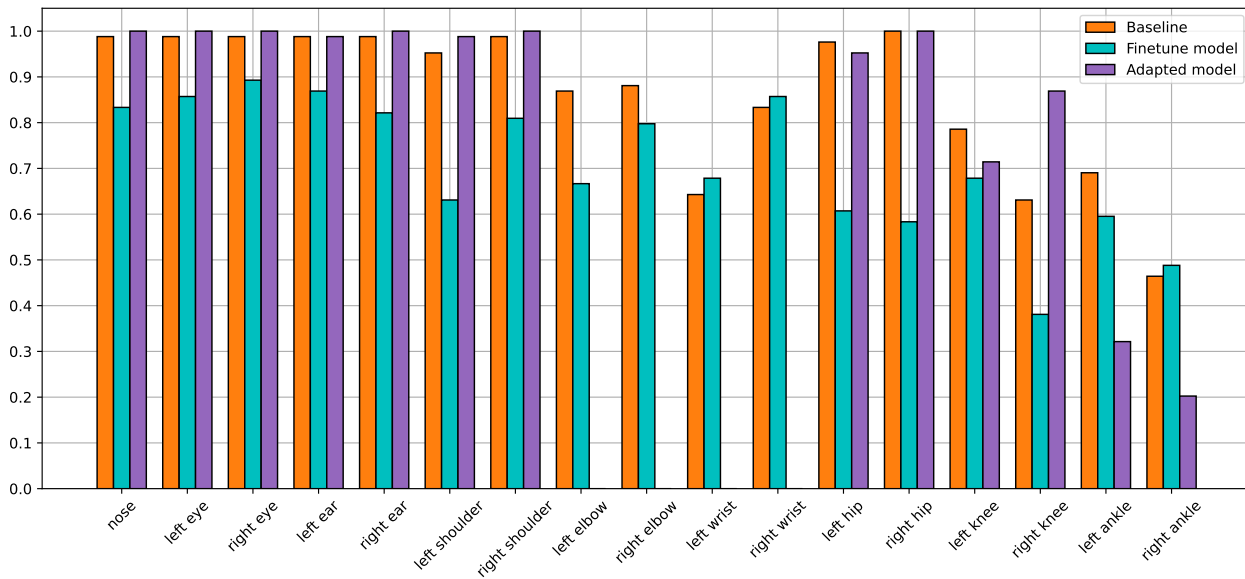


FIGURE 7. Distribution of correctly predicted points over the whole topology for the RTMPose model, related PCK@0.25

Conclusion

As a result of this study, the effectiveness of Progressive Unsupervised Learning’s teacherless domain adaptation method based on generation and filtering of pseudo-labeled data for pose estimation models has been demonstrated. The results obtained can be put into practice in real-time sports systems. By selecting the filtering function, this method not only saves a lot of resources previously spent on labeling, but also shows an improvement in prediction quality compared to other approaches for training a similar model.

The proposed approach of utilizing domain adaptation significantly reduces the time to develop a task-specific solution. Therefore, it is possible to further investigate more sophisticated filtering functions that will improve the performance of the model.

References

- [1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. Lawrence Zitnick. “Microsoft COCO: Common objects in context”, Computer Vision – ECCV 2014 (Zurich, Switzerland, September 6–12, 2014), Lecture Notes in Computer Science, vol. **8693**, Springer, Cham, 2014, ISBN 978-3-319-10602-1, pp. 740–755.  [↑](#)^{25, 27, 28}
- [2] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, S.-T. Xia. “SimCC: A simple coordinate classification perspective for human pose estimation”, Computer Vision – ECCV 2022 (Tel Aviv, Israel, October 23–27, 2022), Lecture Notes in Computer Science, vol. **13666**, Springer, Cham, 2022, ISBN 978-3-031-20068-7, pp. 89–106.  [↑](#)²⁵
- [3] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, K. Chen. *RTMPose: Real-time multi-person pose estimation based on MMPose*, 2023, 11 pp. [arXiv:2303.07399](#) [cs.CV] [↑](#)²⁵
- [4] MMPose Contributors. *OpenMMLab pose estimation toolbox and benchmark*, <https://github.com/open-mmlab/mmpose>, 2020. [↑](#)²⁵
- [5] J. Jiang, Y. Ji, X. Wang, Y. Liu, J. Wang, M. Long. “Regressive domain adaptation for unsupervised keypoint detection”, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2021 (Nashville, TN, USA, June 20–25, 2021), IEEE, 2021, ISBN 978-1-6654-4509-2, pp. 6776–6785.  [↑](#)²⁵
- [6] D. Kim, K. Wang, K. Saenko, M. Betke, S. Sclaroff. “A unified framework for domain adaptive pose estimation”, Computer Vision – ECCV 2022 (Tel Aviv, Israel, October 23–27, 2022), Lecture Notes in Computer Science, vol. **13693**, Springer, Cham, 2022, ISBN 978-3-031-19827-4, pp. 603–620.  [↑](#)²⁵

- [7] Q. Wu, J. Wan, A. B. Chan. “Progressive unsupervised learning for visual object tracking”, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021* (Nashville, TN, USA, June 20–25, 2021), IEEE, 2021, ISBN 978-1-6654-4509-2, pp. 2993–3002. [doi](#) ^{↑26}
- [8] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann. *BlazePose: On-device real-time body pose tracking*, 2020, 4 pp. [arXiv](#) ^{↑27} [2006.10204](#) [cs.CV] ^{↑27}

Received	20.12.2024;
approved after reviewing	16.01.2025;
accepted for publication	07.04.2025;
published online	30.08.2025.

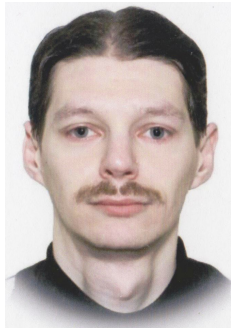
Information about the authors:



Andrei Sergeevich Tokarev

Master’s degree in Moscow Institute of Physics and Technology. Research interests: application of neural networks for image and video analysis, data processing technologies.

e-mail: tokarev.as@phystech.su



Ilia Mikhailovich Voronkov

Senior lecturer at Moscow Institute of Physics and Technology. Research interests: development and application of neural networks in practical problems, analysis and processing of images, video, text data, parametrized data, including for use in robotics problems, in cybersecurity problems.

e-mail: voronkov.im@phystech.su

The authors contributed equally to this article.

The authors declare no conflicts of interests.



Использование доменной адаптации для задачи распознавания ключевых точек на теле человека

Андрей Сергеевич **Токарев**^{1✉}, Илья Михайлович **Воронков**²

^{1,2} Московский физико-технический институт, Москва, Россия

[✉] tokarev.as@phystech.su

Аннотация. В статье проводится исследования алгоритмов доменной адаптации для задачи распознавания ключевых точек на теле человека с целью применения в спорте, когда требуется повысить точность распознавания и сократить трудоемкость ручной разметки данных. Результатом работы является алгоритм итеративной адаптации модели на собственных псевдометках. Экспериментально показано, что метод позволяет получить более эффективную итоговую нейросетевую модель в сравнении с обычным дообучением.

(Связанные тексты статьи на английском и на русском языках)

Ключевые слова и фразы: Ключевые точки, оценка позы человека, доменная адаптация без учителя

Для цитирования: Токарев А. С., Воронков И. М. *Использование доменной адаптации для задачи распознавания ключевых точек на теле человека* // Программные системы: теория и приложения. 2025. Т. 16. № 4(67). С. 23–50. (Англ.+русс.) https://psta.psiras.ru/read/psta2025_4_23-50.pdf

Введение

В последние годы среди стремительно развивающихся направлений машинного обучения и компьютерного зрения выделяется активно применяемая в спорте оценка позы человека (Human Pose Estimation), основанная на распознавании ключевых точек на теле человека (Keypoint Detection). На ее основе создана система полуавтоматического определения офсайда, которая применяется на международных футбольных соревнованиях с 2022 года.

Большое количество видов спорта позволяет в полной мере раскрыть потенциал технологии оценки позы человека для решения сложных динамических задач на основе видеоаналитики, а также в ходе соревнований в режиме онлайн. Для каждой отдельной постановки задачи имеет большое значение точность оценки позы человека, а также необходимо учитывать специфику видов спорта. Эффективность технологии немыслима без большого количества качественно размеченных наборов данных для последующего обучения нейросетевых моделей.

Среди обилия видеозаписей спортивных мероприятий крайне мало размеченных людьми, так как это требует достаточно много времени. Методы доменной адаптации без учителя позволяют обойтись без разметки данных, при этом улучшать качество работы модели. Эти методы позволяют преодолеть разрыв между исходными и целевыми данными путем выравнивания распределений данных, адаптации признаков и применения обученных моделей к новому домену с минимальной дополнительной обработкой. Таким образом, модели становятся более гибкими и способны эффективно работать в различных условиях, не требуя значительного объема новых данных или переработки архитектуры. В конечном итоге использование методов доменной адаптации не только ускоряет процесс внедрения, но и снижает затраты на разработку, поскольку уменьшает необходимость в полном цикле обучения новой модели.

1. Смежные исследования

1.1. Оценка позы человека

Для моделей оценки позы необходимо определить топологию точек, которые и будут эту позу составлять. Исторически поза определялась датасетом, на котором модель обучается, но с 2015 года наиболее распространенной стала 17 точечная топология, предложенная исследователями

из Microsoft [1]. Все более новые описания «скелетов» являются ее расширениями и уточняют положение конечностей.

На сегодняшний день в мире распознавания ключевых точек на теле человека существуют два подхода к решению задач: нисходящий (top-down) и восходящий (bottom-up). Для первого подхода понадобится дополнительный детектор, который локализует человека на изображении. Потом на локализованной области распознаются необходимые ключевые точки и собирается скелет. Это обеспечивает высокую точность предсказания, но делает архитектуру уязвимой к ошибкам на уровне обнаружения человека. Второй подход, наоборот, распознает сразу все ключевые точки на изображении, а уже после этого решает задачу корректного распределения точек по «скелетам». Модели, использующие данный подход, требуют на вход изображение высокого разрешения, что увеличивает потребление ресурсов для их обучения.

Из-за ограниченности в ресурсах в исследовании рассматривались модели top-down подхода, такие как алгоритм Simple Coordinate Classification (SimCC) [2] с нейронной сетью ResNet, в качестве магистральной части, и модель RTMPose [3]. Обе архитектуры описаны в проекте MMPose от OpenMMLab [4], что позволило универсализировать код для процесса обучения и адаптации разных моделей.

1.2. Доменная адаптация

Доменная адаптация является частным случаем задачи переноса обучения (transfer learning): когда постановка задачи остается постоянной, а изменяется только набор данных (домен). При наличии только сырых данных без аннотаций принято использовать доменную адаптацию без учителя. Она позволяет модели, обученной на большом и универсальном домене, сохранить высокую точность предсказанных результатов и производительность на неизвестном ранее датасете, избегая затрат на дополнительное обучение. Таким образом создается возможность быстрой настройки нейронной сети для работы над узкоспециализированной задачей, сохраняя при этом колоссальные ресурсы на анализ и разметку новых данных.

Существующие методы адаптации моделей оценки позы человека на неразмеченном домене используют дополнительные нейронные сети, такие как состязательный перепесор в Regressive Domain Adaptation [5] или модели переноса стиля и алгоритм среднего учителя в Unified Framework [6]. Это повышает вычислительную сложность метода, а также

требует манипуляций со внутренними слоями архитектуры, что усложняет программную реализацию.

Также есть более простой, но эффективный метод Progressive Unsupervised Learning [7], который работает без добавления дополнительных нейронных сетей. Основная идея PUL состоит в итеративном дообучении модели на лучших из своих предсказаний и позволяет повысить точность работы. В таком случае предсказания модели называются псевдо-разметкой или набором псевдометок, а отбор лучших осуществляется выбранной функцией фильтрации, а задача сводится к выбору такой функции, чтобы алгоритм не оказался вычислительно сложным, и нейросеть демонстрировала улучшение результатов работы.

Ниже мы показываем, что для задачи распознавания ключевых точек на теле человека существует функция фильтрации, которая позволяет использовать метод PUL для адаптации нейросетевых моделей к новым доменам. Функция позволяет отсеивать шумные предсказания с помощью показателя уверенности модели в своих предсказаниях и регулируемого порогового значения. Все предсказания, для которых средняя уверенность по скелету переходит порог, попадают в набор псевдометок, на которых в текущей итерации модель дообучается.

2. Эксперимент

Практическая часть исследования включала несколько этапов, таких как подготовка данных, обучение моделей и последующее сравнение их результатов в различных состояниях. На первом этапе осуществлялась подготовка данных, которая включала сбор, обработку и частичную разметку. Далее на основе подготовленных данных проводились адаптация и обучение моделей. После завершения процесса обучения проведены анализ и сравнение полученных результатов, чтобы определить эффективность метода доменной адаптации относительно обычного обучения модели.

2.1. Система полуавтоматической разметки данных

Для аннотирования набора видеозаписей из целевого домена данных была создана система полуавтоматической разметки «pose-markup» (<https://github.com/andrwtokar/pose-markup>). Пример работы программы можно увидеть на рисунке 1.



Рисунок 1. Пример работы системы полуавтоматической разметки данных

Основой системы стала идея автоматической разметки с помощью модели распознавания ключевых точек на теле человека BlazePose [8] с последующей корректировкой результата экспертом. Для первой части использовалось публичное API на языке Python, предоставленное проектом MediaPipe. Правка результатов предсказаний производилась с помощью инструментов обработки изображений, предоставляемых библиотекой OpenCV.

В качестве базовой топологии была выбрана 17-точечная топология COCO [1]. Для каждого кадра аннотация записывается в формате JSON с возможностью последующей агрегации результатов нескольких кадров в один JSON файл.

Анализ разметки данных, выполненной системой, показал насколько это сокращает количество затрачиваемых ресурсов. Было замечено, что в среднем на кадре изменялось 6.5 точек из 17. Причем чаще всего приходилось править координаты точек руки, расположенной дальше от камеры, и головы. Более детальное распределение произведенных изменений можно увидеть на рисунке 2.

Время автоматической разметки видеозаписи с 420 кадрами заняло порядка 2 минут. Корректировка точек занимала в среднем 3 минуты на один кадр. Суммарно на разметку 420 кадров целевых данных было потрачено 24 часа работы. При разметке 2500 кадров это заняло бы более 120 часов. Хотя система и автоматизировала часть работы, мы все равно получаем большие затраты ресурсов на разметку данных.

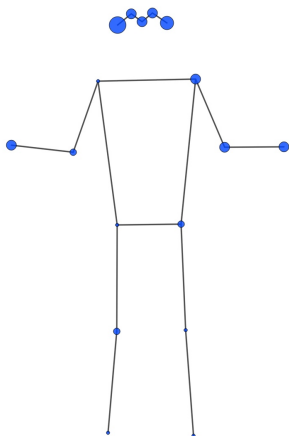


РИСУНОК 2. Схематическое представление изменений в точках, внесенных экспертом

2.2. Данные

Для проведения исследования было выбрано два набора данных: универсальный, на котором базово будет обучена модель, и целевой, к которому мы будем адаптироваться.

В качестве первого из них был выбран датасет COCO [1]. Из-за его большого объема и ограниченности в ресурсах для обучения было принято решение уменьшить объем. Из более чем 250 тысяч поз были выбраны только те, в которых представлены все 17 ключевых точек, согласующихся с топологией COCO. В итоге получен универсальный набор, состоящий из 8 тысяч аннотированных поз, представленных на 6 тысячах изображений.

Второй набор данных был создан вручную в небольшом объеме. Для этого было снято два видео одиночных тренировок спортсменов боксеров по 2500 кадров каждое. Первое использовалось для итеративного метода адаптации модели. Из второго видео были выбраны 420 кадров, которые в дальнейшем были размечены с помощью созданной для этого системы pose-markup и использовались для дообучения модели и снятия метрик. Пример полученных данных можно видеть на рисунке 3.



Рисунок 3. Примеры изображений целевого домена

2.3. Метрики

В качестве базовой метрики в статьях про оценку позы используют метрику average precision (AP) на основе метрики object keypoint similarity (OKS). Однако для нее необходимо знать стандартное отклонение ключевых точек для выбранного датасета, что требует большого объема выборки и ресурсов для сбора статистических показателей.

Из множества похожих метрик, которые оценивают ошибку между предсказанной и реальной точкой, выделяется метрика percentage of correct keypoints (РСК). Преимущество данной метрики в том, что она может быть применима как в случае двумерных координат, так и в случае трехмерных. Ошибка представляется как евклидово расстояние между точками и сравнивается с пороговой величиной. Математическая формула выглядит следующим образом:

$$(1) \quad PCK = \frac{\sum_{i=1}^n \text{bool}(d_i < \text{threshold} * \text{body_height})}{n},$$

где

d_i — расстояние между предсказанной и правильной точкой,

threshold — порог, задаваемый исследователем,

body_height — высота прямоугольника, внутри которого находится человек,

$\text{bool}(\ast)$ — логическое условие, возвращает 1, если оно верно и 0 в ином случае,

n — размер выборки.

В рамках эксперимента использовалась метрика РСК с пороговыми значениями равными 0.05, 0.25 и 0.5.

2.4. Полученные результаты

В рамках эксперимента производилось сравнение трех состояний алгоритма:

- (1) Базовый случай. Когда модель обучена только на универсальном датасете.
- (2) Адаптированное состояние. К базовой модели применили 10 итераций алгоритма доменной адаптации. Схема адаптации моделей представлена на рисунке 4.

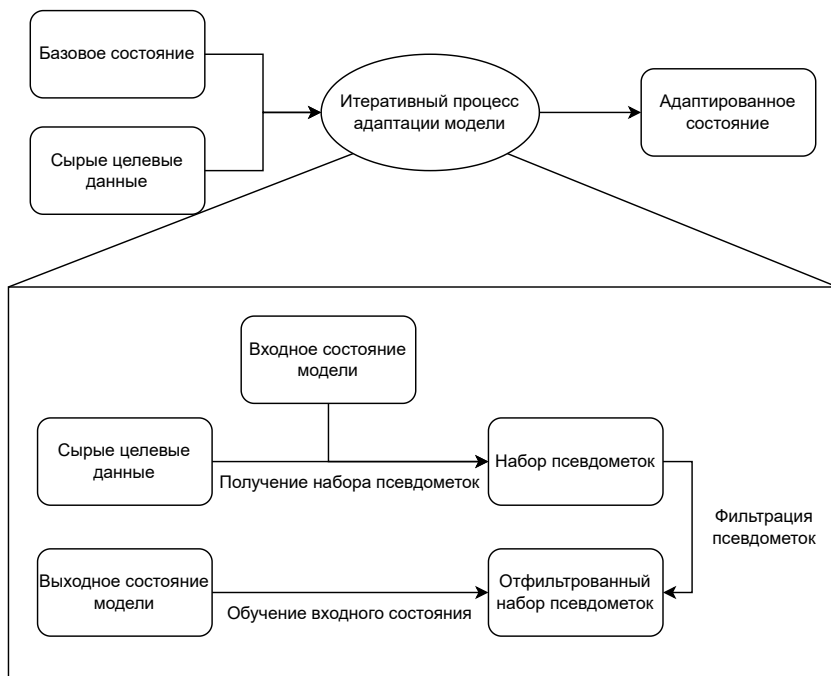


Рисунок 4. Схема адаптации моделей

- (3) Дообученное состояние. Базовую модель дополнительно обучили на размеченных данных из целевого домена. Схема дополнительного обучения моделей представлена на рисунке 5.

Описанная выше функция фильтрации псевдометок подразумевает получение значения уверенности в предсказанной позе. Это значение получается путем усреднения значения уверенности по всем 17 точкам топологии.

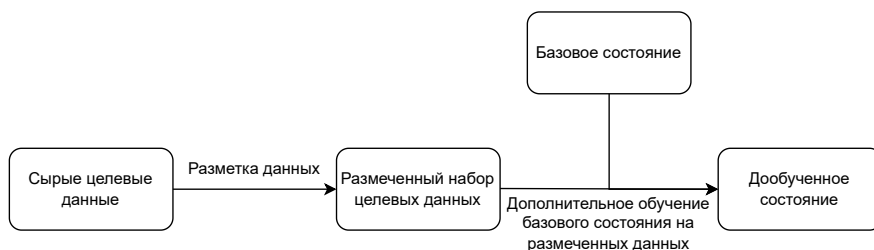


Рисунок 5. Схема дополнительного обучения моделей

Данный способ на первых итерациях отсеивал более 75 процентов данных, используемых для адаптации. На 10 итерациях количество шумных поз было близким к нулю, что говорит об улучшении качества предсказания модели с увеличением количества итераций.

Остановка итеративного процесса производилась после фиксированного числа из 10 шагов. Если бы шагов было выбрано больше, то останавливать итеративный процесс можно либо по достижении заранее выбранной точности, либо при отсутствии изменений в отфильтрованном наборе псевдометок.

Обучение нейронных сетей производилось на платформе Google Colaboratory с использованием графического ускорителя Tesla T4 с 15 Гб памяти. Оценка метрики производилась локально на ноутбуке с графическим ускорителем NVidia MX250 с 2 Гб памяти.

В таблице представлены результаты, полученные для алгоритма SimCC с порогом уверенности равным 0.6. По результатам в таблице 1

Таблица 1. Сравнительная статистика нескольких состояний для алгоритма SimCC

	PCK@0.05	PCK@0.25	PCK@0.5
Baseline	0.024	0.429	0.758
Adapted model	0.037	0.445	0.792
Finetune model	0.0062	0.411	0.76

видно улучшение метрики PCK не только относительно базового состояния, но и относительно дополнительно обученного. При этом стоит заметить, что суммарно на адаптацию было потрачено не более 80 минут, что значительно меньше, чем на разметку тех же данных.

Базовая версия RTMPose показала результаты значительно лучше предыдущей модели. Из таблицы 2 видно, что среди всего эксперимента,

Таблица 2. Сравнительная статистика нескольких состояний для модели RTMPose

	PCK@0.05	PCK@0.25	PCK@0.5
Baseline	0.282	0.882	0.976
Adapted model	0.145	0.652	0.709
Finetune model	0.076	0.709	0.922

Baseline дает наилучшие результаты.

Чтобы объяснить причины ошибок, построим гистограмму распределения верных предсказаний по точкам топологии. Рисунки 6 и 7 показывают, что количество правильных координат точек верхних конечностей падает практически до нуля и вносит большую ошибку в значение метрики. Также падает качество предсказания для точек лодыжек. Описанные точки у боксеров чаще других находятся в движении и являются размытыми на некоторых кадрах видеозаписи.

В текущей версии функции фильтрации эти точки могли попасть в обучающую выборку на выбранной итерации из-за механизма фильтрации и усреднения уверенности по всем точкам для получения уверенности в позе.

Чтобы сравнить работу алгоритма без учета размытия, исключим данные 6 точек из результатов. Как видно из таблицы 3, адаптированная

Таблица 3. Результаты предсказания модели RTMPose без учета размытых ключевых точек

	PCK@0.25	PCK@0.5
Baseline	0.944	0.988
Adapted model	0.964	1.0

модель показывает более высокое качество предсказаний относительно базового случая. В дальнейшем, чтобы побороть ошибки, вносимые размытыми частями изображения, необходимо модифицировать функцию фильтрации.

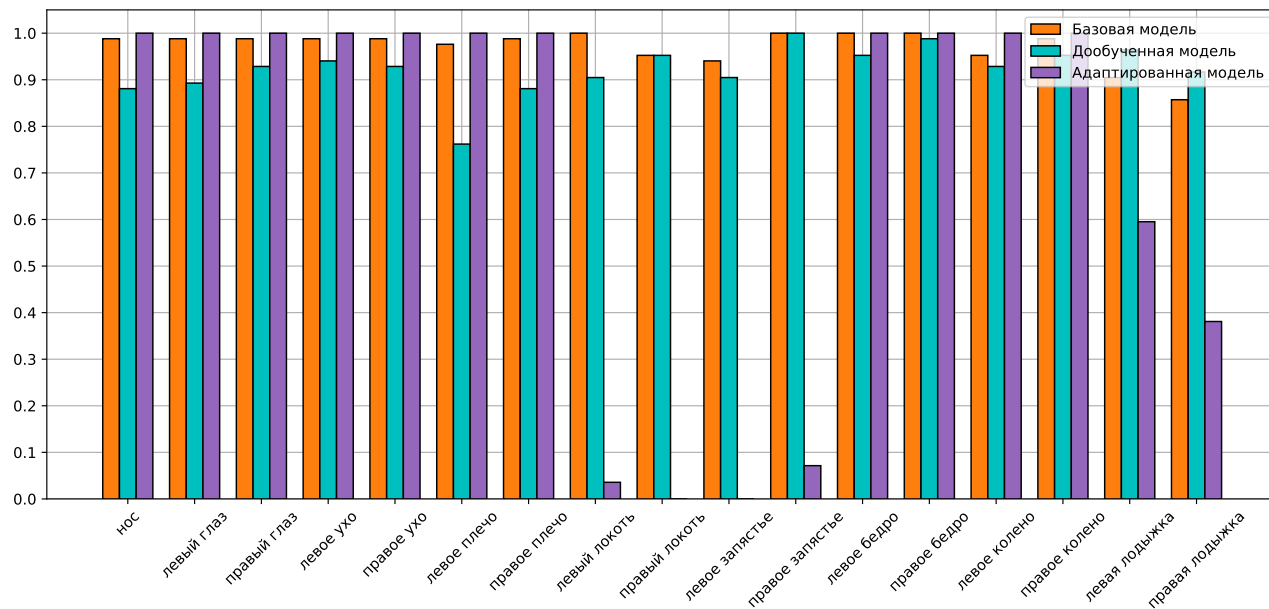


Рисунок 6. Распределение корректно предсказанных точек по всей топологии для модели RTMPose, аналог PCK@0.5

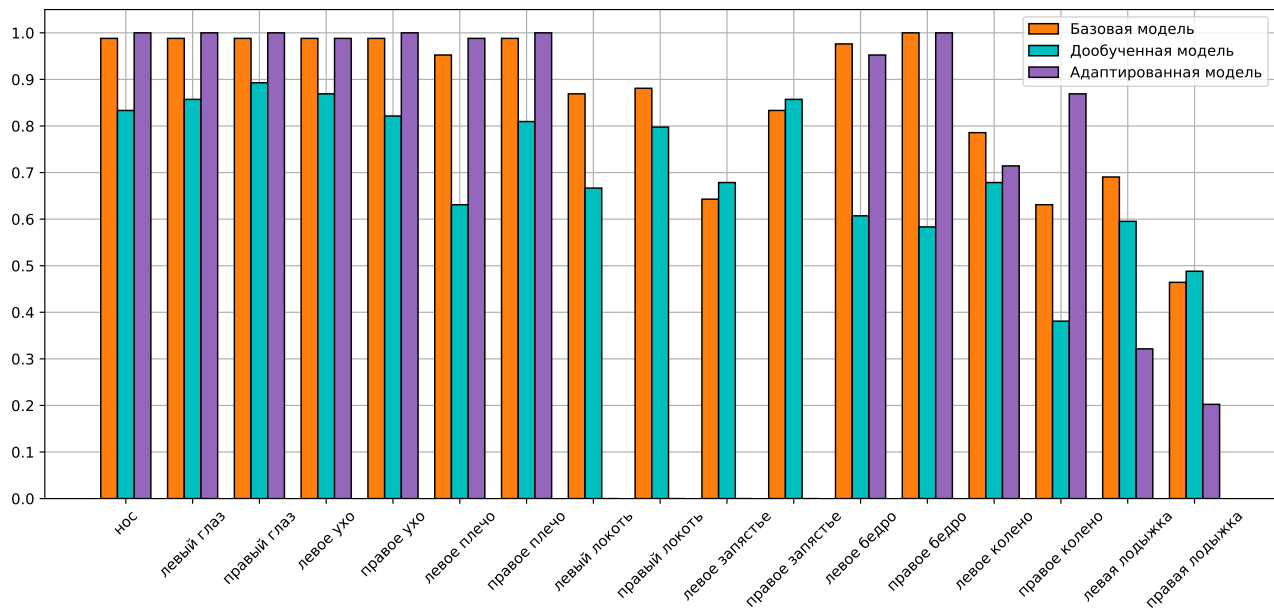


Рисунок 7. Распределение корректно предсказанных точек по всей топологии для модели RTMPose, аналог PCK@0.25

Заключение

В результате проведенного исследования была показана эффективность метода доменной адаптации без учителя Progressive Unsupervised Learning, основанного на генерации и фильтрации псевдоразмеченных данных, для моделей оценки позы. Полученные результаты можно применять на практике в спортивных системах реального времени. За счет выбора функции фильтрации данный метод не только экономит много ресурсов, ранее затрачиваемых на разметку, но и показывает улучшение качества предсказания в сравнении с другими подходами по обучению аналогичной модели.

Предложенный подход использования доменной адаптации существенно сокращает время на разработку решения задач под конкретные цели. Поэтому возможно продолжать исследования более сложных функций фильтрации, которые повысят эффективность модели.

Список использованных источников

- [1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. Lawrence Zitnick *Microsoft COCO: Common objects in context*, Computer Vision — ECCV 2014 (Zurich, Switzerland, September 6–12, 2014), Lecture Notes in Computer Science.– vol. **8693**, Cham: Springer.– 2014.– ISBN 978-3-319-10602-1.– Pp. 740–755.  ^{↑39, 41, 42}
- [2] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, S.-T. Xia *SimCC: A simple coordinate classification perspective for human pose estimation*, Computer Vision — ECCV 2022 (Tel Aviv, Israel, October 23–27, 2022), Lecture Notes in Computer Science.– vol. **13666**, Cham: Springer.– 2022.– ISBN 978-3-031-20068-7.– Pp. 89–106.  ^{↑39}
- [3] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, K. Chen *RTMPose: Real-time multi-person pose estimation based on MMPose*.– 2023.– 11 pp. [arXiv:2303.07399](https://arxiv.org/abs/2303.07399) [cs.CV] ^{↑39}
- [4] MMPose Contributors *OpenMMLab pose estimation toolbox and benchmark*, <https://github.com/open-mmlab/mmpose>.– 2020. ^{↑39}
- [5] J. Jiang, Y. Ji, X. Wang, Y. Liu, J. Wang, M. Long *Regressive domain adaptation for unsupervised keypoint detection // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2021 (Nashville, TN, USA, June 20–25, 2021).– IEEE.– 2021.– ISBN 978-1-6654-4509-2.– Pp. 6776–6785.  ^{↑39}
- [6] D. Kim, K. Wang, K. Saenko, M. Betke, S. Sclaroff *A unified framework for domain adaptive pose estimation*, Computer Vision — ECCV 2022 (Tel Aviv, Israel, October 23–27, 2022), Lecture Notes in Computer Science.– vol. **13693**, Cham: Springer.– 2022.– ISBN 978-3-031-19827-4.– Pp. 603–620.  ^{↑39}

- [7] Q. Wu, J. Wan, A. B. Chan *Progressive unsupervised learning for visual object tracking* // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021 (Nashville, TN, USA, June 20–25, 2021).– IEEE.– 2021.– ISBN 978-1-6654-4509-2.– Pp. 2993–3002. doi ↑40
- [8] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann *BlazePose: On-device real-time body pose tracking.*– 2020.– 4 pp. arXiv:2006.10204 [cs.CV] ↑41

Поступила в редакцию 20.12.2024;
одобрена после рецензирования 16.01.2025;
принята к публикации 07.04.2025;
опубликована онлайн 30.08.2025.

Рекомендовал к публикации

к.т.н. Е. П. Куршев

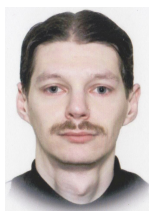
Информация об авторах:



Андрей Сергеевич Токарев

Магистр МФТИ. Научные интересы: применение нейросетей для анализа изображений и видео, технологии обработки данных.

e-mail: tokarev.as@phystech.su



Илья Михайлович Воронков

старший преподаватель кафедры интеллектуальных информационных систем и технологий МФТИ. Научные интересы: разработка и применение нейросетей в практических задачах, анализ и обработка изображений, видео, текстовых данных, параметических данных, в том числе для применения в задачах робототехники, в задачах кибербезопасности.

e-mail: voronkov.im@phystech.su

Вклад авторов: А.С. Токарев – 75% (идея, программное обеспечение, валидация, формальный анализ, проведение экспериментов, сбор материала, курирование данных, написание черновой версии, доработка и редактирование, визуализация); И.М. Воронков – 25% (идея, методология, доработка и редактирование, наставничество).

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*