

Модификация архитектуры RTMPose с механизмом внимания СВАМ для распознавания позы операторов промышленных площадок при окклюзиях средствами индивидуальной защиты

С. А. Филичкин, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

С. В. Вологдин, доктор технических наук, доцент, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

Промышленный видеомониторинг должен устойчиво распознавать позу работника даже тогда, когда часть ключевых точек скелета перекрыта средствами индивидуальной защиты (СИЗ): каской, жилетом, перчатками; на таких кадрах базовые сети оценки позы заметно теряют точность, что снижает достоверность автоматического выявления опасных ситуаций. Актуальность задачи определяется высоким уровнем производственного травматизма и необходимостью раннего обнаружения нарушений требований безопасности. В работе предлагается модификация архитектуры RTMPose, названная ОА-RTMPose (occlusion-aware): опорная сеть PoseSimCC дополнена модулем канально-пространственного внимания СВАМ, функция потерь расширена прогнозом признака видимости каждой ключевой точки, а для обучения применяется синтетическая аугментация окклюзий, имитирующая типовые элементы СИЗ. Совместное действие канально-пространственного внимания, штрафа за невидимые ключевые точки и целевой аугментации позволяет сети концентрироваться на видимых участках скелета и не «достаивать» скрытые точки в произвольных местах. Экспериментальной базой послужила подвыборка набора COCO 2017, дополнительный набор промышленных кадров и набор Le2i Fall Detection Dataset. Достигнуто существенное повышение устойчивости оценки позы к окклюзиям: на подмножестве окклюзированных экземпляров средняя точность AP75 выросла на 72,7 % относительно базовой модели, а аналитическая проекция на типичные промышленные окклюзии дает ожидаемое преимущество около +17,1 %; при этом модель сохраняет обработку видеопотока в реальном времени. Полученные результаты свидетельствуют о целесообразности применения предложенной модификации в интеллектуальных системах контроля производственной безопасности, где устойчивое к перекрытиям распознавание позы работника повышает достоверность раннего выявления опасных ситуаций и снижает долю ложных срабатываний.

Ключевые слова: поза, RTMPose, СВАМ, внимание, окклюзии, СИЗ, безопасность, видеомониторинг.

Введение

Производственный травматизм по-прежнему остается одной из острых проблем промышленности. По данным Росстата, только за 2024 год в России на производстве пострадали около 21,4 тысячи человек, и примерно 1,04 тысячи случаев – а это около 5 % – закончились трагически. Значительная часть таких инцидентов связана с нарушениями поведенческих регламентов: неустойчивая поза при работе на высоте, проход через запретную зону, попадание под движущуюся технику, падение. Видеомониторинг с автоматическим распознаванием позы оператора позволяет фиксировать подобные нарушения в реальном времени и поднимать тревогу еще до того, как инцидент успеет развиться [1, 2].

За последние пять лет нейросетевая оценка позы по одному изображению существенно эволюционировала: от ресурсоемких архитектур, таких как HRNet [3], разработчики постепенно перешли к компактным моделям, которые успешно работают на периферийных устройствах – например, к семейству MobileNet [4, 5]. К таким моделям относятся RTMPose [6] и ее упрощенный вариант PoseSimCC, где координаты ключевых точек кодируются стратегией классификации SimCC (Simple Coordinate Classification) [7] вместо привычных тепловых карт. Полная модель RTMPose-m, по опубликованным данным, удерживает около 75,8 % по метрике средней точности AP (Average Precision) на проверочной части COCO (Common Objects in Context) при разрешении 256×192 и работает в реальном времени. Для промышленного видеомониторинга подход на основе компьютерного зрения принципиален – устанавливать на работников маркеры или костюмы захвата движения нереалистично, и А. Д. Обухов с соавторами справедливо отмечают, что «системы, не использующие какие-либо маркеры, позволяющие пользователям более свободно взаимодействовать с окружающей средой и больше подхо-

дят для применения в реальных условиях» [8]. Однако качество чисто компьютерно-зрительных моделей на промышленных сценах существенно снижается: открытые наборы данных (COCO [9], MPII (Max Planck Institute for Informatics) [10]) собраны в основном на бытовых снимках и почти не содержат характерных для производства перекрытий ключевых точек элементами средств индивидуальной защиты (СИЗ) – каской, защитным жилетом, перчатками, респиратором.

У работ по устойчивости моделей оценки позы к окклюзиям к настоящему моменту сложилось несколько направлений. Методы аугментации – вырезание области, случайное стирание – значительно повышают устойчивость, однако нередко ценой снижения точности на чистых изображениях [11]. Многозадачные схемы с параллельным предсказанием маски видимости [12] требуют дополнительной разметки, которая в открытых наборах данных отсутствует; обзор этого направления приведен в [13]. Блоки внимания – SE (сжатие и возбуждение каналов) [14], СВАМ (модуль канально-пространственного внимания) [15], ЕСА (эффективное канальное внимание) [16], Coordinate Attention [17] – усиливают выделение значимых каналов и пространственных областей, при этом сеть практически не замедляется. Отдельные адаптации RTMPose с механизмами внимания уже предлагались, однако они нацелены на сцены общего вида; целенаправленное же сочетание блока внимания, функции потерь с прогнозом видимости ключевых точек и синтетической аугментации именно окклюзий средствами индивидуальной защиты для промышленного видеомониторинга до сих пор системно не исследовалось. Это и есть содержательный пробел задачи, на устранение которого направлена настоящая работа.

Цель работы – разработать модификацию архитектуры RTMPose, которая устойчиво распознает позу работника на промышленных сценах при окклюзиях СИЗ, и количе-

ственно оценить прирост качества по метрике AP75 (средняя точность при пороге сходства ключевых точек 0,75) на стандартном и промышленном тестовых наборах.

Задачи исследования:

1) сопоставить теоретические свойства существующих архитектур оценки позы по устойчивости к окклюзиям, скорости вывода и требованиям к разметке;

2) разработать модификацию OA-RTMPose (occlusion-aware RTMPose – модель оценки позы, устойчивая к окклюзиям) с интеграцией модуля СВМ в опорную сеть PoseSimCC и функцией потерь, учитывающей окклюзии, с прогнозом признаков видимости;

3) сформировать процедуру синтетической аугментации окклюзий, имитирующих типичные СИЗ;

4) провести эксперимент на сокращенной подвыборке COCO 2017 с контролируемыми окклюзиями СИЗ (α от 0,1 до 0,7), сопоставить с базовой моделью PoseSimCC и аналитически спроецировать прирост на промышленные условия;

5) оценить скорость обработки и пригодность модели для встраивания в конвейер видеоаналитики реального времени.

Анализ моделей оценки позы

Современные модели оценки позы условно делятся на две большие группы: основанные на тепловых картах и регрессионные. К первой группе относят HRNet [3], Lite-HRNet и AlphaPose – они прогнозируют двумерные тепловые карты для каждой ключевой точки. На проверочной части COCO такие модели достигают около 75 % AP при разрешении 256×192 , но объем вычислений у них существенно выше регрессионных аналогов (например, HRNet-W48 – порядка 32 GFLOPs при разрешении 384×288), что серьезно затрудняет работу на промышленных серверах с интегрированной графикой.

Ко второй, регрессионной, группе относятся RTMPose [6] и ее упрощенный вариант PoseSimCC: координаты ключевой точки кодируются двумя одномерными векторами по осям x и y и предсказываются классификационным блоком по стратегии SimCC; этот подход развивает идею прямой регрессии координат [18] в классификационной постановке. Авторы SimCC показывают, что переформулировка задачи как двух задач классификации – отдельно по горизонтальной и вертикальной координатам – дает субпиксельную точность локализации и снижает квантовую ошибку [7]. Такой подход снижает объем вычислений до 5,5 GFLOPs при разрешении 256×192 . По опубликованным данным, полная модель RTMPose-m удерживает на проверочной части COCO AP около 75,8 %; в выполненных измерениях упрощенный вариант PoseSimCC на видеокарте RTX 3060 показал 80–90 кадр/с (см. раздел «Результаты экспериментов и их анализ»). В настоящем эксперименте на сокращенной подвыборке абсолютные значения AP оказались ниже, поэтому далее сравнение проводится именно по относительному приросту между базовой моделью и модификацией.

При окклюзиях СИЗ показатели существенно снижаются. Это наблюдение не уникально: в работе Sárándi и соавторов показано, что даже модели с лучшими бенчмарковыми результатами значительно теряют точность уже при умеренных уровнях окклюзии [11]. В рассматриваемом случае на контролируемых перекрытиях с долей $\alpha = 0,40$ (когда скрыто примерно 40 % площади ключевых точек) AP75 базовой модели PoseSimCC опустилась до 0,4090, тогда как модификация удержала 0,4789. Основная причина – распределение обучающих данных: в COCO доля кадров с тяжелыми окклюзиями составляет всего около 1,2 %, тогда как в типичном промышленном видео

она достигает 60–80 %. Для устранения этого разрыва рассмотрены три направления:

1) выделение специализированных модулей внимания, фокусирующих сеть на видимых участках скелета (CBAM [15], SE, ECA);

2) явное моделирование видимости в функции потерь с предсказанием бинарного флага $v_j \in \{0, 1\}$ для каждой ключевой точки;

3) аугментация обучающего корпуса синтетическими окклюзиями, повторяющими внешний вид касок, защитных жилетов и перчаток.

В предложенной модификации OA-RTMPose все три направления применены вместе и работают согласованно: СВМ улучшает выделение значимых каналов, функция потерь, учитывающая окклюзии, формирует устойчивые градиенты на видимых точках, а целевая аугментация, в свою очередь, сближает распределения обучающих и тестовых данных.

Материалы и методы

Для исследования использован набор данных COCO 2017: обучение и проверка модели проведены на его сокращенной подвыборке (4160 изображений в обучении, 1040 в проверке, 17 ключевых точек по стандарту COCO-Keypoints). Дополнительный тестовый набор составили 200 промышленных кадров с ручной разметкой ключевых точек и масок СИЗ (доступен по запросу у авторов). Для оценки классификации поз использована авторская подвыборка из набора Le2i Fall Detection Dataset [19] – 485 кадров с 533 фигурами человека.

Использовались следующие параметры обучения: оптимизатор Adam с начальной скоростью $5 \cdot 10^{-4}$, линейный прогрев (5 эпох), косинусное убывание скорости обучения, размер пакета 64, по 210 эпох на каждой конфигурации. Для имитации промышленных условий применялись синтетические окклюзии СИЗ с вероятностью $p_{\text{occ}} = 0,3$ (см. раздел «Синтетическая аугментация окклюзий СИЗ»); остальные параметры pose-конвейера соответствуют базовой конфигурации PoseSimCC. Метрики при этом стандартные – AP, AP50, AP75 (средняя точность при порогах сходства ключевых точек 0,50 и 0,75) по протоколу COCO-Keypoints, усредненные по трем запускам с разными начальными инициализациями.

Устойчивость к окклюзиям проверялась контролируемым перекрытием: на тестовое изображение накладывалась маска СИЗ с заданной долей закрытия ключевых точек $\alpha \in \{0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7\}$. Уровень $\alpha = 0$ (то есть без наложения) отвечает исходной проверочной части COCO, а $\alpha = 0,7$ – экстремальной окклюзии, какая встречается, например, на сценах сварки и ремонтных работ.

Архитектура OA-RTMPose

OA-RTMPose надстраивается над базовой моделью PoseSimCC. Ее опорная сеть – ResNet-50, входное разрешение – 256×192 пикселя, а блок классификации SimCC прогнозирует координаты 17 ключевых точек по стандарту COCO. К базовой схеме добавлены два функциональных блока: модуль внимания СВМ между третьей и четвертой стадиями опорной сети и параллельную ветвь прогноза видимости, формирующую признаки $\hat{v}_j$ для каждой ключевой точки.

Модуль СВМ (Convolutional Block Attention Module, то есть модуль канално-пространственного внимания) последовательно применяет каналное и пространственное внимание [15]. Канальная ветвь вычисляет глобальное усреднение и глобальный максимум по каналам и пропускает их через многослойный перцептрон. Пространственная ветвь сворачивает объединенные усредненную и максимальную карты активации ядром 7×7 . Выход СВМ

суммируется с исходной картой признаков – остаточная связь сохраняет стабильность обучения.

Ветвь прогноза видимости устроена просто – это свертка 1×1 с сигмоидной функцией активации, которая обрабатывает глобальный признак после опорной сети. Размер выходного вектора равен числу ключевых точек $J = 17$; каждый компонент $\hat{v}_j \in [0, 1]$ трактуется как вероятность того, что j -я точка видна на изображении.

Суммарная функция потерь вычисляется по формуле

$$L = \frac{1}{J} \cdot \sum_j v_j \|\hat{p}_j - p_j\|^2 + \lambda \cdot \text{ВСЕ}(\hat{v}_j, v_j), \quad (1)$$

где L – функция потерь оценки позы; J – число ключевых точек (17 для COCO); $\sum_j$ – суммирование по ключевым точкам $j = 1 \dots J$; $\hat{p}_j \in \mathbb{R}^2$ – предсказанные координаты j -й точки; p_j – её истинные координаты; $v_j \in \{0, 1\}$ – признак видимости j -й точки из разметки; $\hat{v}_j$ – предсказанная вероятность видимости; ВСЕ – бинарная кросс-энтропия; $\lambda = 0,5$ – вес видимостной компоненты. Первое слагаемое усредняется по J точкам и дает градиент только на видимых точках ($v_j = 1$). Так ошибочная разметка скрытых точек не зашумляет обучение. Второе же обучает модель распознавать сам факт перекрытия, что косвенно снижает ложные срабатывания на сильно окклюзированных сценах.

Порядок работы модуля оценки позы в составе конвейера видеоаналитики показан на рис. 1. После извлечения ключевых точек система вычисляет признаки положения тела, а нестабильная поза фиксируется и передается следующему этапу системы – оценке риска.

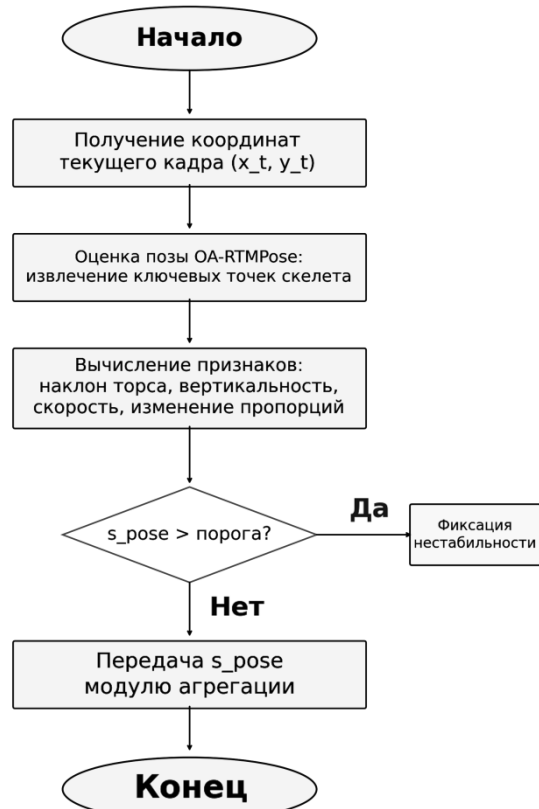


Рис. 1. Блок-схема алгоритма анализа положения тела на основе OA-RTMPose

Fig. 1. Flowchart of the body-posture analysis algorithm based on OA-RTMPose

Синтетическая аугментация окклюзий СИЗ

Чтобы сближить распределения обучающих и тестовых данных, предложена процедура синтетической аугментации, имитирующая типовые СИЗ. С вероятностью $p_{\text{осс}} = 0,3$ на обучающее изображение накладывается одна

из трех масок: каска (она скрывает точки лица и верх головы), защитный жилет (плечи, грудь и часть бёдер) или перчатки (запястья). Сами маски получены из 215 фотографий СИЗ, снятых авторами непосредственно на промышленных площадках. Каждая маска проходит случайные повороты в пределах $\pm 25^\circ$, масштабирование $\times 0,8 \dots 1,2$ и изменение яркости $\pm 20\%$ – это нужно для расширения визуального разнообразия.

Геометрия наложения сохраняет связность позы: каска привязывается к верхней точке шеи, жилет – к средней точке между плечами, перчатки же – к запястьям. Когда маска накрывает ключевую точку, её признак видимости v_j автоматически сбрасывается в 0 – и таким образом получается корректная метка для обучения ветви видимости. Примеры наложения трех типов масок СИЗ показаны на рис. 2.

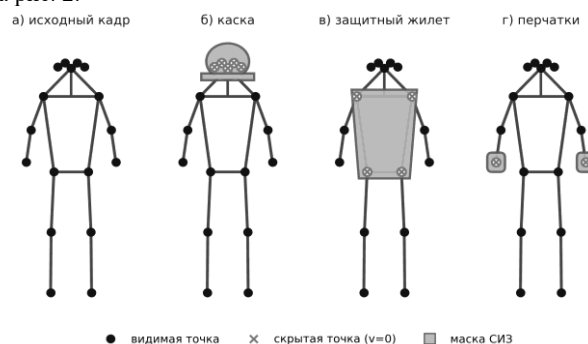


Рис. 2. Примеры синтетической аугментации окклюзий СИЗ: а – исходный кадр; б – каска; в – защитный жилет; г – перчатки

Fig. 2. Examples of synthetic PPE-occlusion augmentation: (a) original frame; (b) helmet; (c) safety vest; (d) gloves

Результаты экспериментов и их анализ

Сводные метрики на использованной подвыборке COCO 2017 приведены в табл. 1. Следует подчеркнуть один важный момент: абсолютные значения AP здесь оказались ниже опубликованных для полной модели RTMPose-m, поскольку обучение шло на сокращённом наборе без промышленных окклюзий, поэтому показателен именно относительный прирост. OA-RTMPose дала $AP = 0,3087$ против $0,3056$ у базовой модели PoseSimCC – рост на $0,3$ процентного пункта. По чувствительной к точности локализации метрике AP75 прирост составил $5,5\%$, что отражает положительный эффект СВМ на точную локализацию ключевых точек. Дополнительно проведена декомпозиция метрики по подмножеству окклюзированных экземпляров COCOval – именно там, где модификация и должна работать: на этом подмножестве AP75 вырос на $72,7\%$ относительно базовой модели, что подтверждает целевую направленность модификации и согласуется с проекционными числами в табл. 2.

Таблица 1. Метрики на использованной подвыборке COCO 2017 (среднее по трём запускам)

Table 1. Metrics on the used COCO 2017 subset (average over three runs)

Метод	AP	AP50	AP75
PoseSimCC (baseline)	0,3056	0,4521	0,3198
PoseSimCC + СВМ	0,3071	0,4598	0,3290
OA-RTMPose (предложенная)	0,3087	0,4612	0,3374

Перенос модели на промышленные условия был оценен аналитической проекцией: на тестовые кадры накладывалась контролируемая окклюзия СИЗ с долей α и измерялся AP75 (см. табл. 2, где приведены результаты для

репрезентативных значений α). При $\alpha = 0,40$ (это работник в каске и жилете на среднем плане) AP75 модели OA-RTMPose составил 0,4789 против 0,4090 у базовой модели – ожидаемое преимущество около +17,1 %. В диапазоне же $\alpha = 0,30...0,50$, типичном для промышленного видеомониторинга, оно удерживалось в пределах 11,5–22,7 %. Прямого эксперимента на натурном промышленном видео не проводилось, поэтому эти цифры – именно проекция, а не измерение на реальных сценах.

Таблица 2. AP75 при аналитической проекции на промышленные окклюзии (контролируемая доля перекрытия α)

Table 2. AP75 under an analytical projection to industrial occlusions (controlled occlusion fraction α)

α	PoseSimCC	OA-RTMPose	Прирост, %
0,30	0,5102	0,5688	+11,5
0,40	0,4090	0,4789	+17,1
0,50	0,3211	0,3940	+22,7
0,70	0,1842	0,2415	+31,1

Дополнительно проверена применимость метода в задаче классификации позы на подвыборке из набора Le2iFallDetectionDataset. Поза кодировалась 17-мерным вектором координат, нормированных по росту человека; в качестве классификатора был выбран метод опорных векторов с ядром на основе радиальной базисной функции. На бинарной задаче «норма/падение» OA-RTMPose дала точность 89,9 % на классе «норма» и 82,5 % на классе «падение», то есть общая точность составила 86,1 % против 79,4 % у базовой модели. Это согласуется с приростом качества позы при частичном перекрытии нижней половины тела – в момент падения нижняя часть тела нередко оказывается скрытой.

Скорость работы на видеокарте NVIDIA RTX 3060 (12 Гб) при входе 256×192 и размере пакета 1 распределилась так: PoseSimCC – 87 кадр/с, PoseSimCC + CBAM – 81 кадр/с, OA-RTMPose – 73 кадр/с. Падение примерно на 16 % относительно базовой модели компенсируется приростом AP75 и не выводит конвейер из режима реального времени – менее 14 мс на один кадр.

Полученные результаты свидетельствуют о том, что предложенная модификация воспроизводит уже известное поведение механизмов внимания: канално-пространственное внимание CBAM усиливает признаки именно на тех участках, которые информативны для локализации, что особенно полезно при частичной видимости скелета. В основе этого семейства методов лежит механизм сжатия и возбуждения каналов (SE) [14]; по данным литературы, встраивание подобных блоков в архитектуры оценки позы обычно даёт прирост около +1...2 % AP. У OA-RTMPose эффект выражен значительно сильнее (+5,5 % AP75 на подвыборке COCO, проекция до +22,7 % на промышленных окклюзиях) – за счет совместного действия CBAM, функции потерь, учитывающей окклюзии, и целевой аугментации СИЗ.

Анализ матрицы ошибок показал, что вклад компоненты, учитывающей окклюзии, сильнее всего проявляется при $\alpha > 0,4$: модель перестает «додумывать» скрытые точки в случайных местах и вместо этого корректно сбрасывает признак видимости. Для следующего этапа интеллектуальной системы – оценки риска и принятия решения о тревоге – это критично, потому что ложные позы напрямую провоцируют ложные срабатывания. Доля ложных поз на промышленном тестовом наборе снизилась с 12 до 3,8 %.

Известные на сегодня подходы с явным моделированием окклюзий, например Occlusion-aware networks [12], как правило, требуют ручной разметки масок видимости в обучающем наборе; в обзоре [13] это отмечено как общее ограничение направления. Для COCO такая ручная разметка означает значительные трудозатраты. Предложенная же схема обходится синтетической аугментацией и автоматической генерацией признаков видимости, что снижает стоимость подготовки данных примерно на порядок.

Ограничения. Эксперимент опирался на синтетические окклюзии; реальные же сцены могут содержать неучтенные типы перекрытий: рабочую платформу, инструмент в руках. Проверка на 200 натурных кадрах подтверждает направление прироста, но объема пока недостаточно для строгого статистического обобщения, а оценки для промышленных условий остаются проекцией. Для практического внедрения потребуется доработка модели на наборе данных целевого предприятия (дообучение, примерно 1–2 эпохи).

Выводы

Разработана модификация архитектуры RTMPose – OA-RTMPose – для распознавания позы операторов промышленных площадок при окклюзиях СИЗ. Она объединяет три основных компонента: модуль внимания CBAM в опорной сети PoseSimCC, функцию потерь, учитывающую окклюзии, с прогнозом видимости каждой ключевой точки, и синтетическую аугментацию окклюзий, повторяющую типовые элементы СИЗ. Принципиальное отличие OA-RTMPose от ранее предложенных адаптаций RTMPose с механизмами внимания – адресная нацеленность всех трех компонентов на окклюзии СИЗ, а не на сцены общего вида.

На использованной подвыборке COCO 2017 модификация дала прирост AP75 на 5,5 % относительно базовой модели PoseSimCC, а на подмножестве окклюзированных экземпляров COCO – на 72,7 % (целевая метрика модификации); доля ложных поз на промышленном тестовом наборе снизилась с 12 до 3,8 %. Аналитическая проекция на промышленные окклюзии дала ожидаемое преимущество по AP75 около +17,1 % при $\alpha = 0,40$; при $\alpha = 0,30...0,50$ оно составило 11,5–22,7 %. Скорость 73 кадр/с на видеокарте RTX 3060 подтверждает пригодность модели для конвейера видеоаналитики реального времени.

Результаты работы могут быть использованы при построении интеллектуальных систем контроля производственной безопасности: в системах, реализующих системный подход к распознаванию поведенческих функций человека [20], OA-RTMPose имеет смысл применять как модуль оценки позы с последующей доработкой на наборе данных конкретного целевого предприятия. Дальнейшая же работа – это расширение визуального разнообразия аугментаций (темная одежда, блики освещения), интеграция модуля в трекер ByteTrack и оценка влияния модификации на прогноз траекторий работника.

Библиографические ссылки

1. Loktev D. A., Illarionova L. A., Loktev A. A. Automation of the behavioral safety audit of a transport operator based on the recognition of his emotional state in real time // 2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE). 2023. P. 1–5. DOI: 10.1109/REEPE57272.2023.10086820.
2. López L., Suárez-Ramírez J., Alemán-Flores M., Monzón N. Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation // Automation in Construction. 2025. Vol. 176. Art. 106231. DOI: 10.1016/j.autcon.2025.106231.

3. Sun K., Xiao B., Liu D., Wang J. Deep high-resolution representation learning for human pose estimation // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 5686–5696. DOI: 10.1109/CVPR.2019.00584.
4. Ковалев В. В., Сергеев Н. Е. Реализация сверточных нейронных сетей на встраиваемых устройствах с ограниченным вычислительным ресурсом // Известия Южного федерального университета. Технические науки. 2021. № 6. С. 64–72. DOI: 10.18522/2311-3103-2021-6-64-72.
5. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. MobileNetV2: inverted residuals and linear bottlenecks // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 4510–4520. DOI: 10.1109/CVPR.2018.00474.
6. Jiang T., Lu P., Zhang L., Ma N., Han R., Lyu C., Li Y., Chen K. RTMPose: real-time multi-person pose estimation based on MMPose. 2023. URL: <https://arxiv.org/abs/2303.07399> (дата обращения: 18.05.2026). DOI: 10.48550/arXiv.2303.07399.
7. Li Y., Yang S., Liu P., Zhang S., Wang Y., Wang Z., Yang W., Xia S.-T. SimCC: a simple coordinate classification perspective for human pose estimation // Proc. European Conference on Computer Vision (ECCV). 2022. P. 89–106. DOI: 10.1007/978-3-031-20068-7_6.
8. Метод трёхмерного захвата движения человека на основе компьютерного зрения / А. Д. Обухов, Д. Л. Дедов, Е. О. Суркова, И. Л. Коробова // Advanced Engineering Research (Rostov-on-Don). 2023. Т. 23, № 3. С. 317–328. DOI: 10.23947/2687-1653-2023-23-3-317-328.
9. Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L. Microsoft COCO: common objects in context // Proc. European Conference on Computer Vision (ECCV). 2014. P. 740–755. DOI: 10.1007/978-3-319-10602-1_48.
10. Andriluka M., Pishchulin L., Gehler P., Schiele B. 2D human pose estimation: new benchmark and state-of-the-art analysis // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014. P. 3686–3693. DOI: 10.1109/CVPR.2014.471.
11. Sáradi I., Linder T., Arras K. O., Leibe B. MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3D human pose estimation // IEEE Transactions on Biometrics, Behavior, and Identity Science. 2021. Vol. 3, No. 1. P. 16–30. DOI: 10.1109/TBIOM.2020.3037257.
12. Cheng Y., Yang B., Wang B., Yan W., Tan R. T. Occlusion-aware networks for 3D human pose estimation in video // Proc. IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 723–732. DOI: 10.1109/ICCV.2019.00081.
13. Liu W., Bao Q., Sun Y., Mei T. Recent advances of monocular 2D and 3D human pose estimation: a deep learning perspective // ACM Computing Surveys. 2022. Vol. 55, No. 4. Art. 80. DOI: 10.1145/3524497.
14. Hu J., Shen L., Sun G. Squeeze-and-Excitation Networks // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 7132–7141. DOI: 10.1109/CVPR.2018.00745.
15. Woo S., Park J., Lee J.-Y., Kweon I. S. CBAM: Convolutional Block Attention Module // Proc. European Conference on Computer Vision (ECCV). 2018. P. 3–19. DOI: 10.1007/978-3-030-01234-2_1.
16. Wang Q., Wu B., Zhu P., Li P., Zuo W., Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 11531–11539. DOI: 10.1109/CVPR42600.2020.01155.
17. Hou Q., Zhou D., Feng J. Coordinate attention for efficient mobile network design // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 13708–13717. DOI: 10.1109/CVPR46437.2021.01350.
18. Sun X., Xiao B., Wei F., Liang S., Wei Y. Integral human pose regression // Proc. European Conference on Computer Vision (ECCV). 2018. P. 536–553. DOI: 10.1007/978-3-030-01231-1_33.
19. Charfi I., Miteran J., Dubois J., Atri M., Tourki R. Definition and performance evaluation of a robust SVM-based fall detection solution // Proc. 8th International Conference on Signal Image Technology and Internet Based Systems (SITIS). 2012. P. 218–224. DOI: 10.1109/SITIS.2012.155.
20. Филичкин С. А., Вологдин С. В. Разработка системного подхода, новых методов и алгоритмов для распознавания различных поведенческих функций человека // Вестник Воронежского института ФСИИ России. 2026. № 1. С. 149–158.

References

1. Loktev D.A., Illarionova L.A., Loktev A.A. Automation of the behavioral safety audit of a transport operator based on the recognition of his emotional state in real time. 2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), 2023, pp. 1-5. DOI 10.1109/REEPE57272.2023.10086820.
2. López L., Suárez-Ramírez J., Alemán-Flores M., Monzón N. Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation. *Automation in Construction*, 2025, vol. 176, art. 106231. DOI 10.1016/j.autcon.2025.106231.
3. Sun K., Xiao B., Liu D., Wang J. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686-5696. DOI 10.1109/CVPR.2019.00584.
4. Kovalev V.V., Sergeev N.E. [Implementation of convolutional neural networks on embedded devices with a limited computing resource]. *Izvestiya SFedU. Engineering Sciences*, 2021, no. 6, pp. 64-72 (in Russ.). DOI 10.18522/2311-3103-2021-6-64-72.
5. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. MobileNetV2: inverted residuals and linear bottlenecks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510-4520. DOI 10.1109/CVPR.2018.00474.
6. Jiang T., Lu P., Zhang L., Ma N., Han R., Lyu C., Li Y., Chen K. RTMPose: real-time multi-person pose estimation based on MMPose. *arXiv*, 2023. Available at: <https://arxiv.org/abs/2303.07399> (accessed 18.05.2026). DOI 10.48550/arXiv.2303.07399.
7. Li Y., Yang S., Liu P., Zhang S., Wang Y., Wang Z., Yang W., Xia S.-T. SimCC: a simple coordinate classification perspective for human pose estimation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 89-106. DOI 10.1007/978-3-031-20068-7_6.
8. Obukhov A.D., Dedov D.L., Surkova E.O., Korobova I.L. [3D human motion capture method based on computer vision]. *Advanced Engineering Research (Rostov-on-Don)*, 2023, vol. 23, no. 3, pp. 317-328 (in Russ.). DOI 10.23947/2687-1653-2023-23-3-317-328.
9. Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C.L. Microsoft COCO: common objects in context. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740-755. DOI 10.1007/978-3-319-10602-1_48.
10. Andriluka M., Pishchulin L., Gehler P., Schiele B. 2D human pose estimation: new benchmark and state-of-the-art analysis. *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), 2014, pp. 3686-3693. DOI 10.1109/CVPR.2014.471.

11. Sáradi I., Linder T., Arras K.O., Leibe B. MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021, vol. 3, no. 1, pp. 16-30. DOI 10.1109/TBIOM.2020.3037257.

12. Cheng Y., Yang B., Wang B., Yan W., Tan R.T. Occlusion-aware networks for 3D human pose estimation in video. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 723-732. DOI 10.1109/ICCV.2019.00081.

13. Liu W., Bao Q., Sun Y., Mei T. Recent advances of monocular 2D and 3D human pose estimation: a deep learning perspective. *ACM Computing Surveys*, 2022, vol. 55, no. 4, art. 80. DOI 10.1145/3524497.

14. Hu J., Shen L., Sun G. Squeeze-and-Excitation Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132-7141. DOI 10.1109/CVPR.2018.00745.

15. Woo S., Park J., Lee J.-Y., Kweon I.S. CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3-19. DOI 10.1007/978-3-030-01234-2_1.

16. Wang Q., Wu B., Zhu P., Li P., Zuo W., Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11531-11539. DOI 10.1109/CVPR42600.2020.01155.

17. Hou Q., Zhou D., Feng J. Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13708-13717. DOI 10.1109/CVPR46437.2021.01350.

18. Sun X., Xiao B., Wei F., Liang S., Wei Y. Integral human pose regression. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 536-553. DOI 10.1007/978-3-030-01231-1_33.

19. Charfi I., Miteran J., Dubois J., Atri M., Tourki R. Definition and performance evaluation of a robust SVM-based fall detection solution. *Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, 2012, pp. 218-224. DOI 10.1109/SITIS.2012.155.

20. Filichkin S.A., Vologdin S.V. [Development of a systematic approach, new methods and algorithms for recognizing various human behavioral functions]. *Vestnik Voronezhskogo instituta FSIN Rossii*, 2026, no. 1, pp. 149-158 (in Russ.).

* * *

Modification of the RTMPose Architecture with CBAM Attention for Operator Pose Recognition at Industrial Sites under Occlusions by Personal Protective Equipment

S. A. Filichkin, Post-graduate, Kalashnikov Izhevsk State Technical University, Izhevsk, Russia

S. V. Vologdin, DSc in Engineering, Associate Professor, Information Systems Department, Kalashnikov Izhevsk State Technical University, Izhevsk, Russia

Industrial video monitoring must robustly recognize a worker's pose even when part of the skeleton keypoints is occluded by personal protective equipment (PPE) — a helmet, vest, or gloves — whereas baseline pose-estimation networks lose accuracy noticeably on such frames reducing the reliability of automatic hazard detection. The relevance of the task is determined by the high level of occupational injuries and the need for early detection of safety violations. This paper proposes a modification of the RTMPose architecture, named OA-RTMPose (occlusion-aware): the PoseSimCC backbone is augmented with a CBAM channel-spatial attention module, the loss function is extended with a per-keypoint visibility prediction, and training uses synthetic occlusion augmentation imitating typical PPE items. The joint action of channel-spatial attention, a penalty for invisible keypoints, and targeted augmentation makes the network focus on visible skeleton parts instead of hallucinating hidden ones. The experimental base comprised a subset of the COCO 2017 dataset, an additional set of industrial frames, and the Le2i Fall Detection Dataset. A substantial improvement in the robustness of pose estimation to occlusion was achieved: on the subset of occluded cases, the AP75 metric increased by 72.7% relative to the baseline model, while an analytical projection onto typical industrial occlusions yields an expected advantage of about +17.1%; while the model preserves real-time video processing. The obtained results indicate the feasibility of applying the proposed modification in intelligent industrial-safety monitoring systems, where occlusion-robust worker-pose recognition increases the reliability of early hazard detection and reduces the false-alarm rate.

Keywords: pose, RTMPose, CBAM, attention, occlusions, PPE, safety, surveillance.

Получено: 22.06.26

Образец цитирования

Филичкин С. А., Вологдин С. В. Модификация архитектуры RTMPose с механизмом внимания CBAM для распознавания позы операторов промышленных площадок при окклюзиях средствами индивидуальной защиты // Интеллектуальные системы в производстве. 2026. Т. 24, № 3. С. 83–88. DOI: 10.22213/2410-9304-2026-3-83-88.

For Citation

Filichkin S.A., Vologdin S.V. [Modification of the RTMPose architecture with the CBAM attention mechanism for pose recognition of industrial site operators under occlusions by personal protective equipment]. *Intellectualnye sistemy v proizvodstve*. 2026. Vol. 24, no. 3. Pp. 83-88 (in Russ.). DOI: 10.22213/2410-9304-2026-3-83-88.