DOI 10.37882/2223-2966.2025.06.14

ОБУЧЕНИЕ РОБОТА-МАНИПУЛЯТОРА БЕЗОПАСНОМУ ОБХОДУ ПРЕПЯТСТВИЙ: ПЛАНИРОВАНИЕ, ИМИТАЦИЯ И ПОДКРЕПЛЕНИЕ

TRAINING ROBOTIC MANIPULATORS FOR SAFE OBSTACLE AVOIDANCE: PLANNING, IMITATION, AND REINFORCEMENT

Gao Tianci
Yang Bo
Rao Shengren

Summary. This paper proposes a comprehensive method for training a UR5 robotic manipulator to perform safe and efficient movements in environments with randomly placed obstacles. In the first stage, collision-free «expert» trajectories are automatically generated using the OMPL motion planning library. This eliminates the need for manual robot control and ensures high-quality demonstrations. In the second stage, a behavioral cloning approach is used to derive an initial policy capable of reproducing these demonstrations and avoiding collisions. Finally, reinforcement learning with the PPO algorithm is applied to fine-tune and improve the policy based on a reward function that accounts for positioning accuracy, collision penalties, and other factors. This strategy unites formal trajectory planning with the adaptability inherent to reinforcement learning methods. Experimental results in the simulator show that the proposed three-stage framework — *planner–imitation learning–reinforcement learning* — achieves higher goal accuracy and a lower collision rate compared to classical approaches trained from scratch.

Keywords: motion planning (ompl), imitation learning, reinforcement learning, safe obstacle avoidance, collision-free trajectory, industrial robotics.

Гао Тяньцы
Аспирант, Московский государственный технический университет им. Н.Э. Баумана
gaotianci0088@gmail.com
Ян Бо
Аспирант, Московский государственный технический университет им. Н.Э. Баумана
yangbo.123@hotmail.com
Жао Шэнжэнь
Аспирант, Московский государственный технический университет им. Н.Э. Баумана
raoshengren@gmail.com

Аннотация. В данной работе предлагается комплексный метод обучения манипулятора UR5 безопасным и эффективным перемещениям в условиях случайно расположенных препятствий. На первом этапе с использованием библиотеки планирования движений OMPL автоматически генерируются коллизионно-свободные «экспертные» траектории, что устраняет необходимость ручного управления роботом и обеспечивает высокую надёжность демонстраций. Затем путём поведенческого клонирования формируется начальная политика, способная воспроизводить данные демонстрации и избегать столкновений. На заключительном этапе используется обучение с подкреплением (алгоритм PPO) для донастройки и совершенствования полученной политики с учётом целевой функции вознаграждения (точность позиционирования, штрафы за столкновения и т.д.). Такой подход позволяет объединить формальное планирование безопасных траекторий с адаптивностью, присущей методам обучения с подкреплением. Результаты экспериментов в симуляторе показывают, что предложенная трёхэтапная схема «планировщик — имитационное обучение — подкрепляющее обучение» обеспечивает более высокую точность достижения цели и меньшую частоту столкновений по сравнению с классическими методами, основанными на обучении с нуля.

Ключевые слова: планирование движений (OMPL), имитационное обучение, обучение с подкреплением, безопасный обход препятствий, коллизионно-свободная траектория, промышленная робототехника.

Введение

Актуальность исследования

Современные универсальные манипуляторы находят широкое применение в промышленности благодаря сочетанию высокой точности и гибкости. Однако традиционное ручное программирование в условиях быстро меняющихся производственных задач требует значительных ресурсов и не всегда оказывается эффективным при наличии препятствий. В связи с этим возрастает потребность в методах автоматизированного построения и адаптации траекторий, которые обеспечивают безопасность (отсутствие столкновений) и высокую точность конечного позиционирования.

Обзор существующих методов

Для решения задачи планирования движения манипуляторов традиционно рассматривают несколько подходов:

Планирование в конфигурационном пространстве (RRT, PRM, реализованные в OMPL) — данные алгоритмы позволяют формировать формально безопасные траектории, однако в динамичных условиях могут оказаться вычислительно затратными.

Обучение с подкреплением (ОС) даёт роботу возможность самостоятельно осваивать стратегию на основе механизма проб и ошибок, однако обучение «с нуля» связано с повышенным риском столкновений на ранних этапах.

Имитационное обучение (ИО), в частности поведенческое клонирование (ПК), использует «экспертные» демонстрации для быстрого старта, но в значительной степени зависит от качества и достаточности этих демонстраций.

Основная идея и вклад

В настоящей работе предлагается объединить преимущества формального планировщика и методов обучения. Подход включает три ключевых шага:

1. Генерация демонстрационных данных с помощью библиотеки OMPL, исключающая ручное управление роботом.
2. Имитирующее обучение (ПК) для формирования безопасной стартовой политики на основе «экспертных» траекторий.
3. Донастройка политики методом обучения с подкреплением (PPO), что дополнительно повышает точность позиционирования и адаптивность стратегии.

Проведённые эксперименты показывают, что предлагаемая трёхэтапная схема «планировщик — имитация — подкрепление» обеспечивает более высокие показатели точности и безопасности по сравнению с подходами, не использующими демонстрационные данные.

Связанные работы

В области робототехники и управления манипуляторами обычно выделяют три ключевых направления исследований: планирование движений, обучение с подкреплением и имитационное обучение. За последние годы каждое из этих направлений претерпело существенное развитие, особенно в задачах, связанных с манипуляторами, имеющими высокое число степеней свободы.

1. *Методы планирования движений*

Классические алгоритмы, такие как RRT и PRM, остаются одним из наиболее распространённых способов построения коллизионно-свободных траекторий для роботов-манипуляторов [1], [2]. Открытые библиотеки (например, OMPL [3]) предоставляют гибкий инструментарий для поиска безопасного пути в конфигурационном пространстве с учётом геометрических и кинематических ограничений. Эти методы обычно отличаются высокой надёжностью и объяснимостью, однако в динамических или плохо определённых сценариях вычислительная сложность может существенно возрастать.

2. *Обучение с подкреплением и имитационное обучение*

Обучение с подкреплением (ОС) основано на механизме проб и ошибок с целью максимизации итогового вознаграждения. Такие алгоритмы, как PPO [4] и SAC [5], успешно применяются в непрерывных и высоко размерных задачах. Тем не менее при обучении «с нуля» в реальных роботизированных системах нередко возникают повышенные риски столкновений на ранних этапах и значительные затраты ресурсов.

Имитационное обучение (ИО) использует демонстрации эксперта и тем самым сокращает объём неэффективных действий. Самый простой вариант — поведенческое клонирование (ПК), при котором модель обучается воспроизводить «состояние–действие» на основе предоставленных экспертных данных. Это позволяет быстро получить первоначальную стратегию, однако качество результата сильно зависит от полноты и разнообразия демонстраций. Кроме того, при недостаточном охвате возможных ситуаций возникает риск смещения политики.

3. *Позиционирование данной работы*

С учётом перечисленных особенностей в настоящей работе предлагается объединить методы формального планирования движений и имитационного обучения:

1. **Генерация «экспертных» траекторий** манипулятора с использованием OMPL без ручного управления.
2. **Поведенческое клонирование** для формирования начальной безопасной политики.
3. **Дальнейшая оптимизация** с помощью обучения с подкреплением (PPO) для повышения точности и адаптивности.

Таким образом, комбинация «планировщик + имитация + подкрепление» обеспечивает как безопасность и высокое качество стартовой политики, так и гибкость методов ОС. Эксперименты в среде со случайно размещёнными препятствиями демонстрируют преимущество предлагаемого подхода по точности и эффективности по сравнению с альтернативами, не использующими демонстрационные данные.

Методология

Ниже представлена пошаговая схема предлагаемых методов обучения манипулятора безопасному и эффективному взаимодействию в условиях случайно расположенных препятствий. На рис. 1 показаны основные этапы — от генерации препятствий до итоговой настройки управляющей стратегии.

Основные стадии можно описать следующим образом:

1. **Создание случайных конфигураций среды:** В симуляторе генерируются различные препятствия и целевые положения конечного звена.
2. **Сбор демонстрационных данных с помощью планировщика движений (OMPL):** Алгоритм автоматического поиска пути формирует допустимые, коллизионно-свободные траектории.
3. **Предварительное обучение (поведенческое клонирование):** На основе полученных демонстраций создаётся начальная политика, способная воспроизводить «экспертные» траектории.
4. **Донастройка с подкреплением (PPO):** Политика уточняется с учётом заданной функции вознаграждения, включая штрафы за столкновения и отклонения от цели.
5. **Периодическая оценка и сохранение модели:** Система регулярно оценивает качество стратегии и сохраняет лучшие параметры.

Модель и вознаграждение

Пусть s_t — состояние (углы в суставах, относительная позиция до цели, показатели расстояния до препятствий и пр.), a_t — действие (приращения по суставам). Ищется стратегия $\pi_\theta(a|s)$, максимизирующая сумму вознаграждений.

Обычно вознаграждение включает положительный вклад за приближение к цели и существенный штраф за столкновения или выход за физические ограничения.

Сбор демонстрационных данных

При помощи OMPL генерируется последовательность состояний и действий $\{(s_i, a_i)\}$, гарантированно свободных от коллизий (см. рис. 2). Данный набор служит базой для последующего имитационного обучения.

Имитационное обучение (поведенческое клонирование)

Обучается параметрическая функция $\mu_\phi(\mathrm{s})$, минимизирующая среднеквадратичную ошибку:

$$L_{\text{demo}}(\theta) = \mathbb{E}_{(s,a)\sim D}\left\|a - \pi_\theta(s)\right\|^2. \quad (1)$$

где (s,a) берутся из набора «экспертных» демонстраций D. В результате формируется политика, способная воспроизводить демонстрационные траектории и избегать коллизий.

Донастройка с помощью PPO

Поскольку исходная стратегия может оказаться субоптимальной (избыточные движения, удлинённые тра-

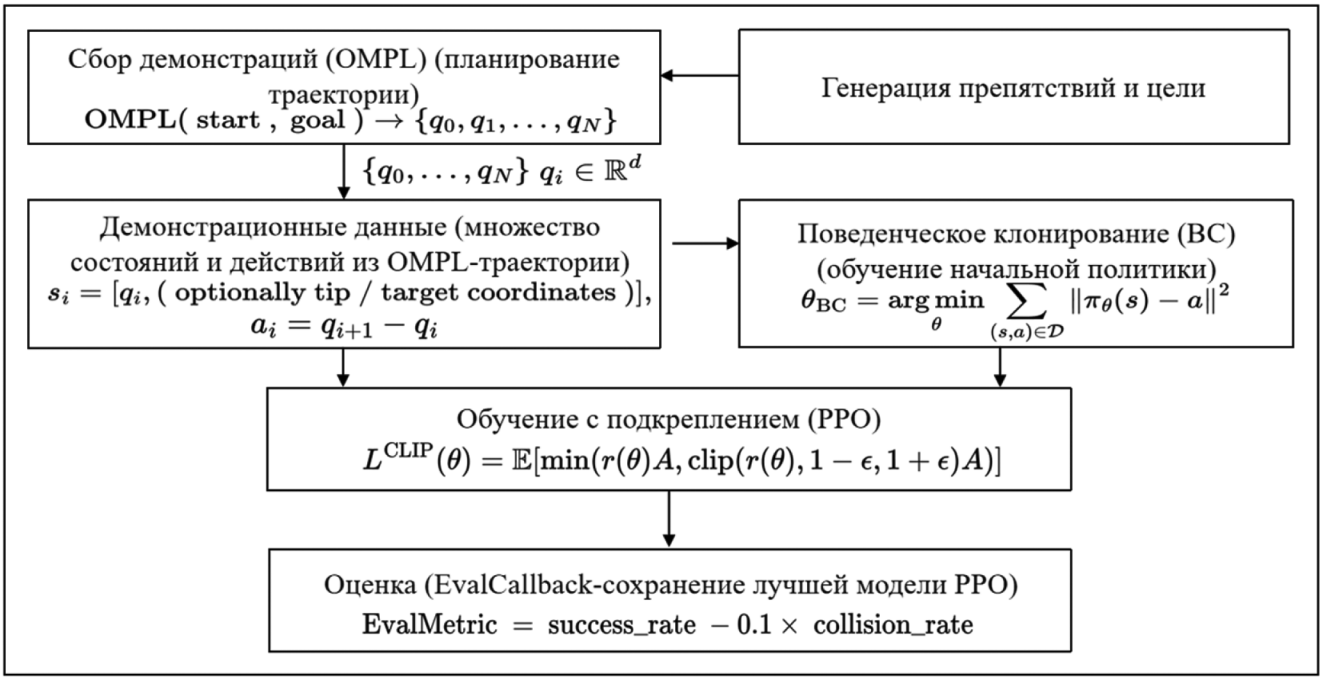

Рис. 1. Общая структура обучения

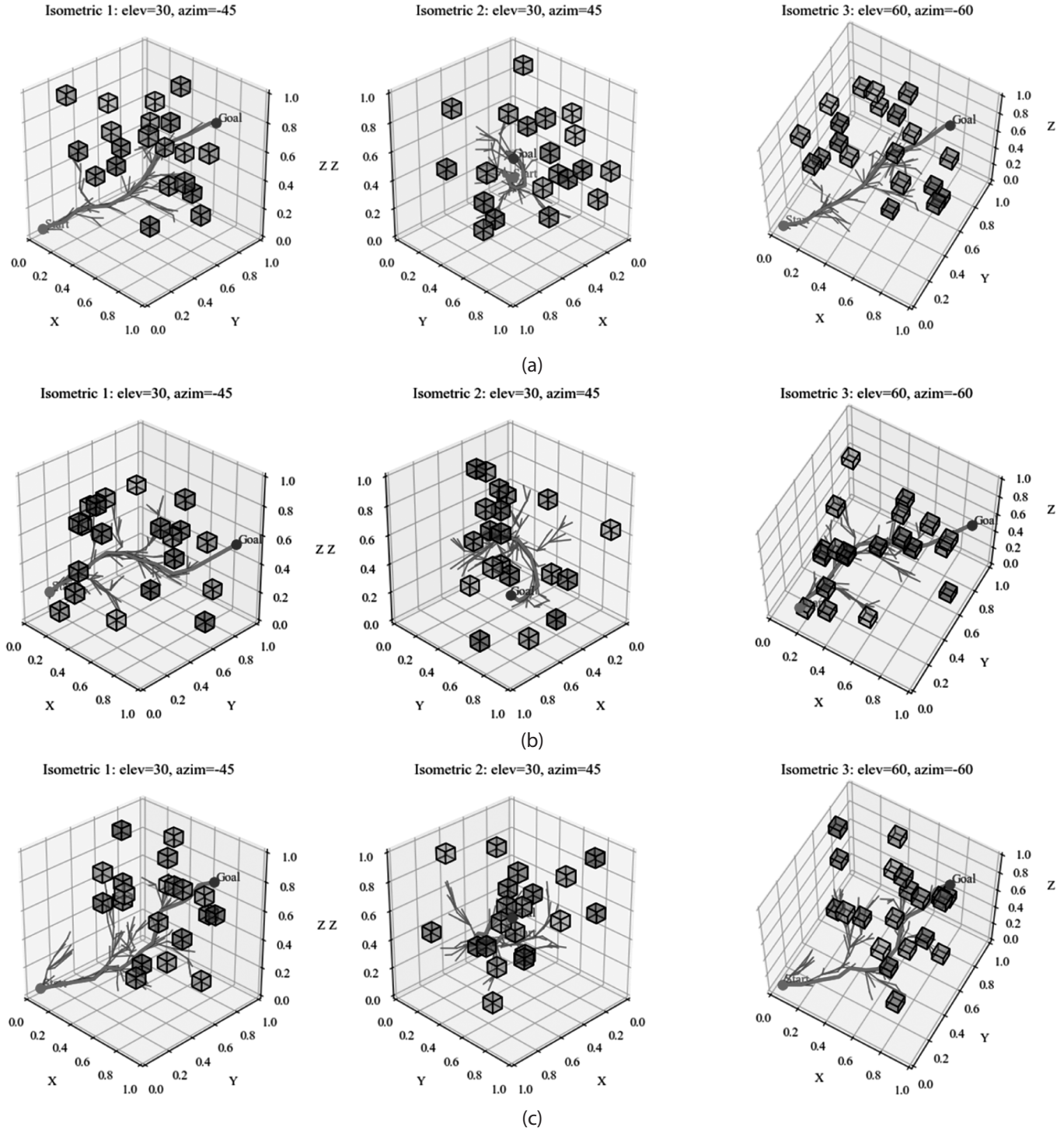

Рис. 2. Пример трёхмерной сцены с препятствиями и построенной траекторией

ектории и т.п.), следующим этапом используется обучение с подкреплением (PPO):

$$L_{reinforce}(\theta) = \\ = \mathbb{E}_t\left[min\left(r_t(\theta)\hat{A}_t, clip\left(r_t(\theta), 1-\epsilon, 1+\epsilon \right)\hat{A}_t \right)\right] \quad (2)$$

где $r_t(\theta)$ — отношение вероятностей действия, $\hat{A}_t$ — оценка преимущества, а ϵ — малый гиперпараметр (например, 0.1 или 0.2). Такой подход обеспечивает плавное обновление параметров и снижает риск деградации уже сформированной политики.

Эксперименты

Среда и настройки

В экспериментальной среде (см. рис. 3) манипулятор UR5 (6 степеней свободы) функционирует в рабочем

пространстве, представленном в виде сетки возможных позиций. В задаче моделируются случайно расположенные препятствия (например, кубы различных размеров и цветов), усложняющие поиск безопасной траектории. Для генерации «экспертных» демонстрационных данных используется планировщик движений OMPL (алгоритмы RRT/PRM), который автоматически формирует допустимые, коллизионно-свободные маршруты.

Для управления манипулятором была выбрана нейросеть со скрытыми слоями по 64 нейрона в каждом (функция активации ReLU). На вход подаются параметры состояния (положение и размеры препятствий, текущие углы в суставах, координаты целевой точки и т.д.), а на выходе формируются приращения по суставам. На рис. 4 схематически представлено, как данная сеть сочетается с двумя этапами обучения: **поведенческим клонированием (ПК)** и **обучением с подкреплением (PPO)**.

Методы сравнения и метрики

В экспериментах сравнивались следующие методы:

1. Предлагаемый (ПК + PPO), где «экспертные» демонстрации служат основой для быстрой инициализации, а затем осуществляется дообучение с подкреплением (на графиках обозначен светло-серым).
2. PPO с модифицированными эвристиками (AEP).
3. Стандартный PPO (без начальных демонстраций).
4. A2C, DDPG, SAC, TD3 — распространённые алгоритмы обучения с подкреплением.

Для оценки эффективности использовались следующие ключевые метрики:

- Успешность: доля эпизодов, в которых манипулятор достиг цели без столкновений.
- Частота столкновений: процент эпизодов, завершившихся коллизией.
- Средняя награда и скорость сходимости: оценка по кривой вознаграждения (reward) в процессе обучения.

Результаты

Успешность. На рис. 5 приводятся результаты при различных допусках по точности конечного позиционирования (10–100 мм). Предлагаемый метод демонстрирует наивысшую успешность, особенно при более жёстких требованиях.

Кривые обучения. На рис. 6 показано, что стратегия, инициализированная через демонстрации, достигает высоких значений награды быстрее, чем методы, обучающиеся «с нуля».

Частота столкновений

Согласно данным таблицы 1, комбинация «ПК + PPO» даёт существенно более низкую вероятность коллизий, поскольку робот перенимает «экспертные» навыки обхода препятствий уже на ранних этапах обучения. Это значительно повышает безопасность и стабильность поведения в дальнейшем.

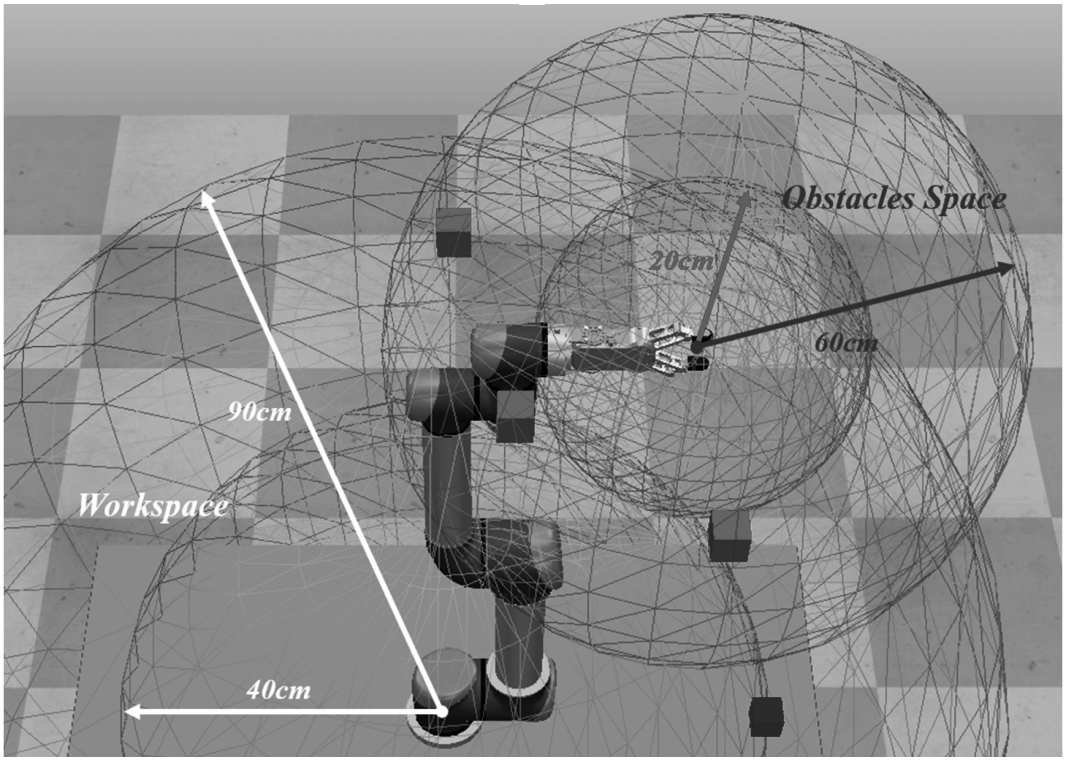

Рис. 3. Пример рабочего пространства манипулятора UR5 с сеткой возможных позиций и случайными препятствиями

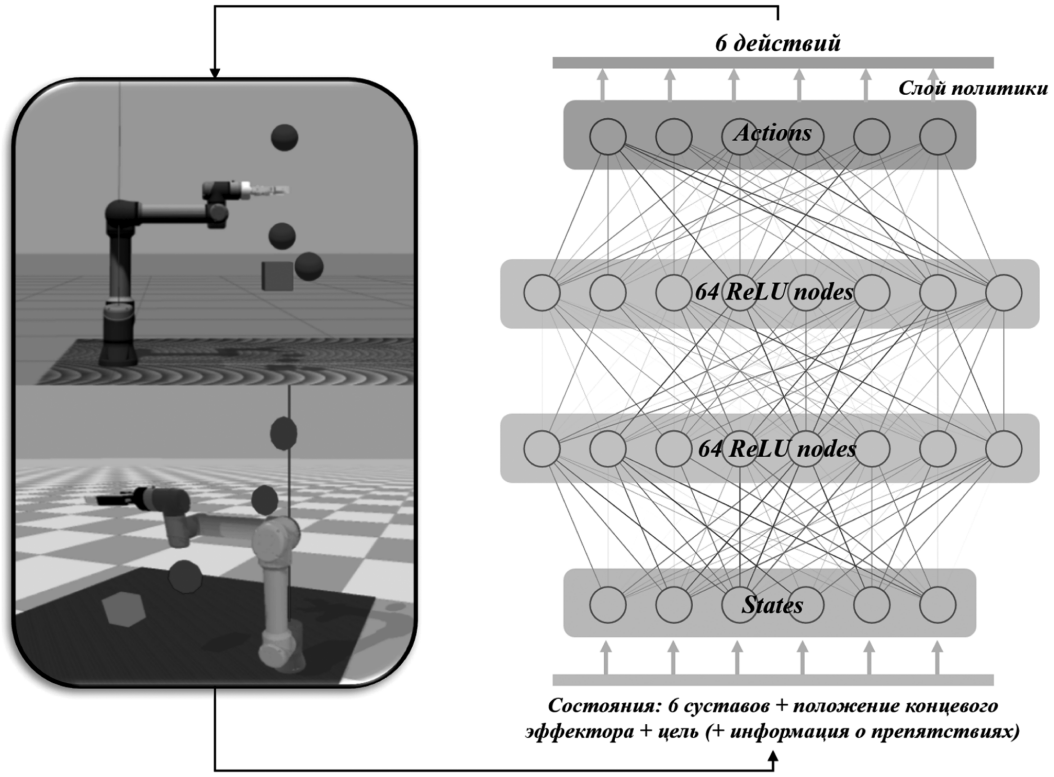

Рис. 4. Структура нейросети (поведенческое клонирование + PPO) для управления UR5

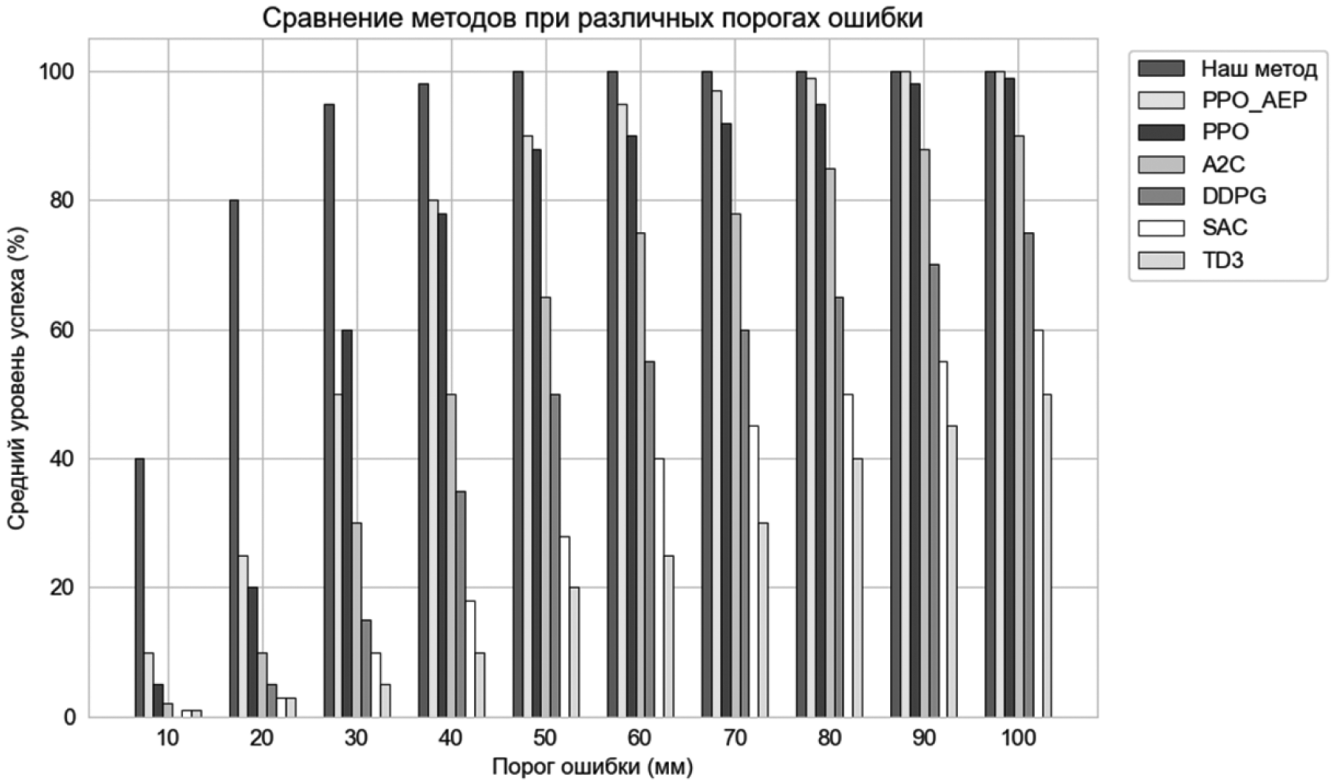

Рис. 5. Уровень успешности для разных методов при различных порогах ошибки позиционирования (в миллиметрах)

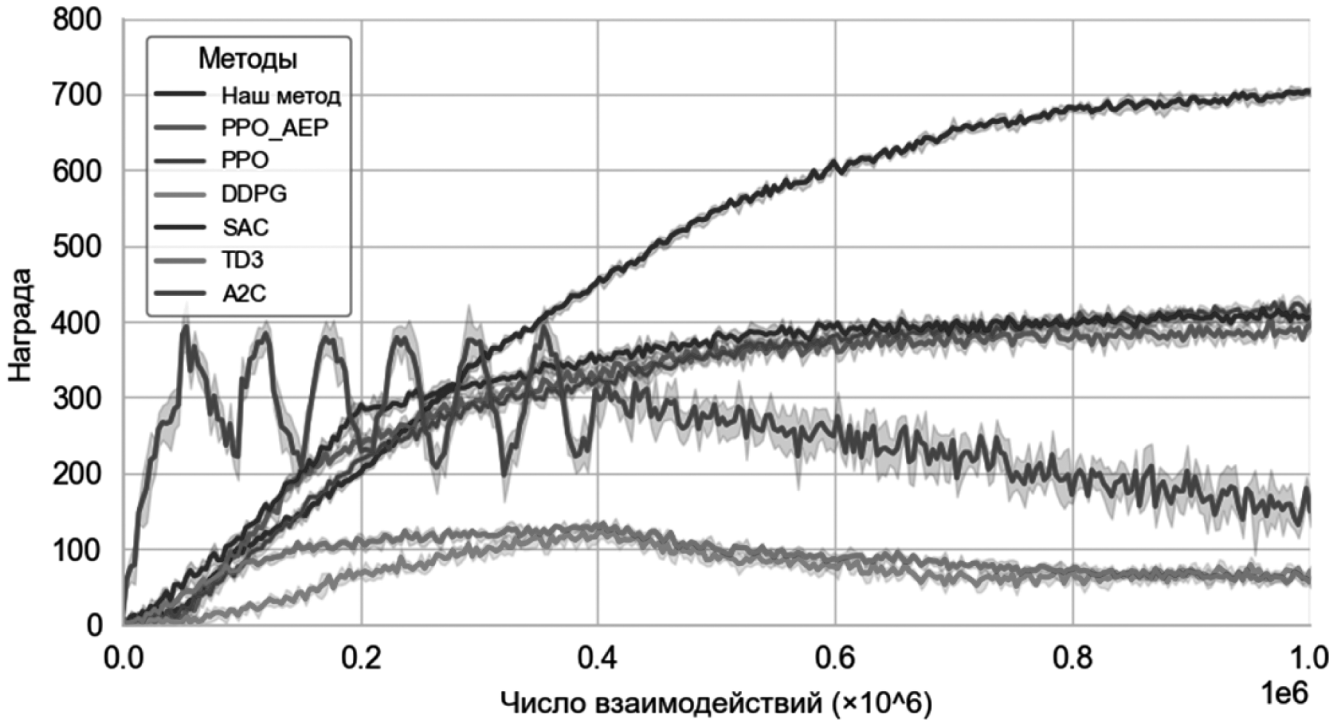

Рис. 6. Кривые обучения разных методов (по оси абсцисс — число шагов, по оси ординат — средняя награда)

Таблица 1.

Сравнение показателей различных методов в задаче управления манипулятором UR5

Метод	Успешность (%)	Вероятность столкновения (%)	Шаги до сходимости
Стандартный PPO	72,6	15,4	$1{,}1\times10^6$
Модифицированный PPO (AEP)	83,4	10,1	$0{,}9\times10^6$
A2C	70,3	14,6	$1{,}2\times10^6$
DDPG	68,5	16,2	$1{,}4\times10^6$
SAC	75,1	12,5	$1{,}1\times10^6$
TD3	78,4	11,9	$1{,}0\times10^6$
Демонстрации + обучение с подкреплением	90,1	5,7	$0{,}85\times10^6$

Заключение

В данной работе предложен трёхэтапный метод обучения манипулятора UR5 безопасному обходу препятствий. На первом этапе с помощью OMPL автоматически генерируются коллизионно-свободные «экспертные» траектории, устраняя необходимость ручного управления роботом и обеспечивая высокую надёжность исходных демонстраций. Затем путём поведенческого клонирования (ПК) формируется начальная политика, способная воспроизводить эти демонстрации и избегать столкновений. Наконец, с помощью алгоритма PPO выполняется дополнительная оптимизация полученной стратегии, что позволяет повысить точность и адаптивность управляющих действий.

Проведённые эксперименты подтверждают следующие преимущества разработанного подхода:

1. **Быстрая сходимость и высокая точность уже на ранних шагах обучения**, что существенно снижает совокупные затраты вычислительных ресурсов.
2. **Минимизация вероятности столкновений**, поскольку робот унаследует «экспертные» навыки избежания препятствий с самого начала обучения.
3. **Гибкость и масштабируемость**: метод легко адаптируется к новым конфигурациям препятствий и другим модификациям задач.

В перспективе планируется экспериментальная проверка предложенного метода на физическом манипуляторе UR5, а также дополнение сценария моделирования динамически перемещающимися препятствиями. Кроме того, особый интерес представляют формальные гарантии безопасности и интерпретируемость стратегий, что имеет критическое значение при внедрении подобных систем в промышленную среду.

ЛИТЕРАТУРА

1. Чжан Л., Цай К., Сун З. и др. (2025). Планирование движений для робототехники: обзор методов на основе выборок. Биомиметические интеллектуальные системы и робототехника, 100207.
2. Фэн Б., Цзян С., Ли Б. и др. (2024). Адаптивный метод multi-RRT для планирования движения робота. Экспертные системы с применениями, 252, 124281.
3. Ян С., Лю П., Пирс Н.Э. (2023). Сравнительный анализ алгоритмов планирования в библиотеке OMPL для роботизированных рук в среде с препятствиями. В Трудах конференции «Инновации в науке и технике Йоркшира (YISEC 2023)».
4. Дель Рио А., Хименес Д., Серрано Х. (2024). Сравнительный анализ алгоритмов A3C и PPO в обучении с подкреплением: обзор в универсальных средах. IEEE Access.
5. Мок Дж. В., Мукнахаллипатна С.С. (2023). Сравнение алгоритмов PPO, TD3 и SAC в обучении с подкреплением для генерации шагающих движений у четвероногих роботов. Журнал интеллектуальных систем обучения и приложений, 15(1), 36–56.

© Гао Тяньцы (gaotianci0088@gmail.com); Ян Бо (yangbo.123@hotmail.com); Жао Шэнжэнь (raoshengren@gmail.com)
Журнал «Современная наука: актуальные проблемы теории и практики»