ЗАЩИТА ОТ СОСТЯЗАТЕЛЬНЫХ АТАК НЕЙРОСЕТЕЙ СИСТЕМЫ УПРАВЛЕНИЯ РОБОТОТЕХНИЧЕСКИМИ КОМПЛЕКСАМИ НА ОСНОВЕ МАРКОВСКИХ ПРОЦЕССОВ

Сацута А. И.[1], Липатников В. А.[2]

DOI:10.21681/3034-4050-2026-4-13-24

Ключевые слова: *безопасность искусственного интеллекта, защита моделей машинного обучения, состязательное обучение, компьютерное зрение, распознавание объектов, системы управления роботизированными комплексами.*

Аннотация

Цель работы: *состоит в анализе уязвимостей нейросетевых моделей систем управления робототехническими комплексами к состязательным воздействиям и разработке способа повышения их защиты на основе методов обучения с подкреплением.*

Метод исследования: *подход, основанный на формализации процесса генерации состязательных примеров как задачи Марковского процесса принятия решений.*

Результаты исследования: *разработан агент, адаптивно генерирующий состязательные примеры путем учета состояния среды и характеристик модели. Предложен способ обучения агента генерации состязательных примеров с последующим их использованием в состязательном обучении. Состояние агента формируется на основе карт признаков модели, градиентной информации и величины искажения входных данных, оцениваемой по норме l2. Действия агента соответствуют выбору параметров атаки типа PGD. Функция награды учитывает как снижение качества детекции, так и ограничение на уровень вносимых искажений.*

Научная новизна: *предложена новая функция награды, учитывающая одновременно качество детекции и величину искажения, что обеспечивает баланс между эффективностью атаки и её незаметностью.*

Введение

В условиях растущей нагрузки на наземные сети связи требуется внедрение новых решений для обеспечения стабильного и качественного обслуживания пользователей. Одним из перспективных подходов является применение робототехнических комплексов, которые могут оперативно развертываться в районах с высокой плотностью пользователей или при чрезвычайных ситуациях. Актуальной является задача повышения защиты нейронных сетей систем управления робототехническими комплексами от состязательных атак.

Современные системы управления робототехническими комплексами (СУ РТК), основанные на методах компьютерного зрения, играют ключевую роль в широком спектре прикладных задач, включая мониторинг окружающей среды, автономную навигацию, принятие решений в реальном времени и распознавание объектов, в том числе объектов, создающих непосредственную угрозу для таких РТК. Такие системы активно применяются в гражданских и военных областях, а также в задачах повышенной ответственности, включая транспортные системы, охранные комплексы и военные приложения. Центральным элементом СУ РТК являются нейросетевые модели, в частности, глубокие сверточные нейронные сети, обеспечивающие детекцию и классификацию объектов в стандартных условиях эксплуатации [1, 2, 3].

Данные нейросетевые модели, как элемент СУ РТК, обладают различными уязвимостями, реализуемые посредством воздействия на пакеты данных в сети передачи данных (СПД) или посредством воздействия на оптические элементы системы управления. Рассматриваемая поверхность атаки дополняется уязвимостью НС моделей к состязательным атакам (СА). Данный тип атак представляют из себя специально сконструированные возмущения входных данных. Такие возмущения способны приводить к искажениям предсказаний моделей. Данные возмущения называются состязательными, а итоговые примеры, полученные с их помощью, называются состязательными примерами. Состязательные атаки приводят к ошибкам классификации и распознавания объектов что критично для систем, функционирующих без участия оператора РТК в автономном режиме [1, 4, 5].

1 ***Сацута Анатолий Игоревич***, старший оператор научной роты военной академии связи им. Маршала Советского Союза С. М. Буденного, г. Санкт-Петербург, Россия. E-mail: toha.satzuta@yandex.ru

2 ***Липатников Валерий Алексеевич***, доктор технических наук, профессор, старший научный сотрудник научно-исследовательского центра Военной академии связи им. Маршала Советского Союза С. М. Буденного, Санкт-Петербург. E-mail: lipatnikovanl@mail.ru

Для СУ РТК проблема устойчивости НС к состязательным воздействиям [6], пример которых представлен на рисунке 1, имеет особое значение, так как ошибки в восприятии окружающей среды напрямую влияют на корректность анализа обстановки и управления. Кроме того, в реальных условиях эксплуатации входные данные подвержены не только преднамеренным атакам, но и различным видам естественного шума, который создаётся посредством изменения освещения, погодных условий и другими физическими эффектами [7].

Существующие методы защиты от СА, одним из которых является состязательное обучение [8], демонстрируют определенную эффективность. Однако такие методы обладают рядом существенных ограничений [9]. В большинстве случаев они используют фиксированные алгоритмы генерации атак, так, например, методы состязательного обучения могут использовать Projected Gradient Descent (PGD). Это приводит к ограниченному разнообразию генерируемых состязательных примеров и, как следствие, к недостаточной обобщающей способности моделей в условиях более сложных или адаптивных атак. Кроме того, статический характер параметров атак не позволяет учитывать текущее состояние модели, динамику обучения и особенности входных данных [1, 5].

В связи с этим актуальной является задача разработки адаптивных методов генерации состязательных возмущений. Одним из перспективных направлений решения данной задачи является использование методов обучения с подкреплением, на примере Марковского процесса принятия решений (MDP). Данный метод позволяет формализовать процесс генерации атак как последовательное принятие решений в условиях неопределенности [10].

В данной работе предлагается подход, основанный на формализации процесса генерации состязательных примеров в виде Марковского процесса принятия решений для последующего их использования в состязательном обучении НС модели распознавания объектов. В рамках предложенной модели агент, адаптивно генерирующий состязательные примеры, обучается выбирать параметры атаки уклонения (атаки). При выборе параметров атаки он опирается на текущее состояние среды, включающее характеристики входных данных и реакцию нейросетевой модели. В качестве базового алгоритма для генерации таких примеров был выбран алгоритм PGD, параметры которого будут оптимизироваться.

Такая постановка задачи позволяет реализовать адаптивную стратегию генерации атак, направленную на максимальное ухудшение качества предсказаний модели при сохранении ограничений на величину возмущений.

Пусть задана модель $f(x)$, где x – входное изображение. Требуется найти такое возмущение δ (формула 1) так, что одновременно выполняются условия ухудшения качества предсказания модели и ограничения на допустимую величину искажения входного сигнала.

Формально задача генерации необходимого состязательного возмущения может быть записана как задача условной оптимизации:

$$\delta^* = \arg\max_{\|\delta\| \le \epsilon} L(f(x+\delta), y), \quad (1)$$

где $L(\cdot)$ – функция потерь модели детекции, y – истинные аннотации объектов, а ограничение $\|\delta\| \le \epsilon$ обеспечивает малую визуальную заметность возмущений и сохранение семантической структуры изображения.

Предлагаемый подход обеспечивает более гибкое и разнообразное пространство атак по сравнению с традиционными методами, а, следовательно, более разнообразные состязательные выборки для обучения. Разнообразные тренировочные выборки приводят к повышению обобщающей способности НС моделей при состязательном обучении. Это делает такие модели более устойчивыми.

Использование MDP позволяет учитывать долгосрочные эффекты последовательных действий агента, что является важным фактором при генерации сложных многошаговых атак [10, 11].

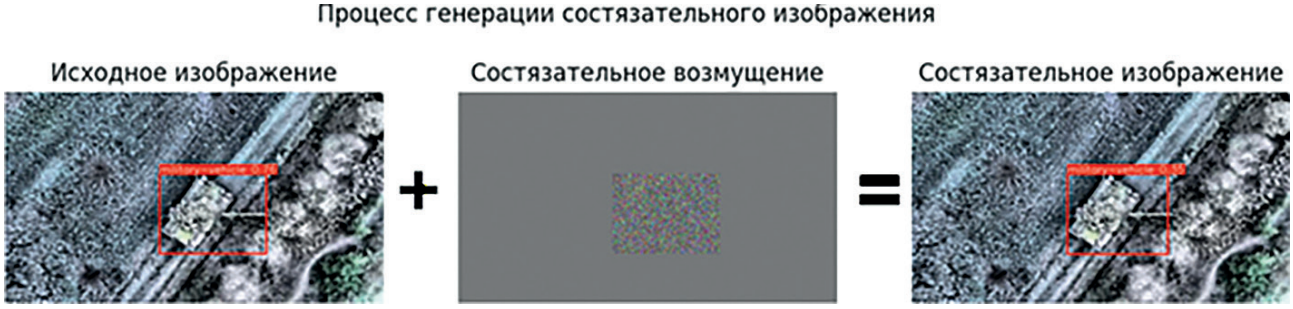


Рис. 1. Уязвимость нейросетевой модели распознавания объектов к состязательным атакам

Достижение поставленной цели по разработке способа и его реализации, в частности путём создания агента Марковского процесса принятия решений, требует выполнения следующих частных задач:

1. Формализация, обоснование и реализация пространства состояний и действий в составе среды агента Марковского процесса принятия решений.
2. Формализация, обоснование и реализация функции наград Марковского процесса принятия решений.
3. Определение подходящего алгоритма обучения агента, его обучение.
4. Выбор метрик для анализа эффективности разрабатываемого способа и агента.
5. Проведение эксперимента и анализ его результатов, сравнение предложенного способа с уже существующими альтернативами, обсуждение результатов.

Разработка агента Марковского процесса принятия решений генерации состязательных примеров

В рамках решаемой задачи генерация состязательных примеров для НС модели детекции объектов формализуется как марковский процесс принятия решений. Данная формализация позволяет рассматривать атаку не как фиксированную последовательность действий, а как последовательность действий, адаптирующихся к текущему состоянию модели и атакуемому изображению. Другими словами, такое представление обеспечивает возможность обучения агента, оптимизирующего стратегию формирования возмущений с учётом как эффективности подавления детекции, так и ограничений на величину искажения входных данных. Формально процесс задаётся кортежем: $\mathcal{MDP} = \{S, A, P, R, \gamma\}$, где S – пространство состояний, A – пространство действий, P – вероятностная функция переходов, R – функция награды, γ – коэффициент дисконтирования.

Состояние среды формируется как минимально достаточное представление информации о поведении модели детекции в текущий момент атаки и уровне внесённого возмущения. В отличие от подходов [12], основанных только на пиксельном представлении изображения, в предлагаемом пространстве состояний используется внутренние представление нейросети, что позволяет агенту учитывать семантические признаки объекта.

Состояние определяется следующим образом: $s_t = \{\|\delta_t\|_2, F_t, \nabla_x L(x_t)\} \in S$, где $\|\delta_t\|$ – l_2 норма текущего возмущения, отражающая уровень отклонения от исходного изображения и контролирующая ограничение заметности атаки, $F_t \in \mathbb{R}^{H \times W \times C}$ – карта признаков предпоследнего слоя детекции, что соответствует высокоуровневым признакам объектов, используемым детектором для принятия решений, $\nabla_x L(x_t)$ – градиент функции потерь по входному изображению, характеризующий чувствительность модели к локальным изменениям входа.

Предложенное пространство состояний позволяет агенту учитывать не только итоговый результат детекции, но и внутреннюю структуру принятия решений модели, включая области наибольшей уязвимости.

Пространство действий агента отражает процесс управления параметрами PGD-атаки. Такое пространство действий позволяет рассматривать генерацию состязательных примеров как процесс адаптивной оптимизации параметров: $a_t = \{\epsilon_t, \alpha_t\} \in A$, где ϵ_t является допустимым радиусом возмущения, ограничивающий максимальное отклонение от исходного изображения, а α_t – шаг градиентного обновления, определяющий скорость изменения изображения на каждой итерации.

Изменение изображения осуществляется с помощью проекционного градиентного спуска, вычисляемый по формуле 2.

$$x_{t+1} = \Pi_{\mathcal{P}_\epsilon(x)}(x_t + \alpha_t sign(\nabla_x L(x_t)), \quad (2)$$

где $\Pi_{\mathcal{P}_\epsilon(x)}$ обозначает проекцию на допустимое множество возмущений.

Таким образом, агент не генерирует возмущение напрямую, а управляет динамикой его формирования через параметры оптимизационного процесса.

Функция награды формализует противоречие между эффективностью атаки и её незаметностью и вычисляется по формуле 3. Такое определение награды приводит к максимизации эффекта подавления детекции при одновременном штрафе за рост искажения входного изображения. Ограниченность компонент на отрезке [0, 1] сохраняет вклад каждого компонента и предотвращает доминирование одного из критериев.

$$(r_t = (1 - \overline{C_t}) + (1 - \overline{IoU_t}) + (1 - \tanh(0{,}1 \times \|\delta_t\|_2)), \quad (3)$$

где $\overline{C_t}$ представляет собой среднюю уверенность модели в предсказаниях объектов, которые были найдены в той или иной степени, верно. В данной работе предсказание считается достаточно верным, если оно покрывает 25 % действительного объекта. Такое уточнение является важным, так как при атаке может возникать большое количество ложных детекций, в которых модель может быть сильно уверена. Большое количество ложных детекций приводит к компрометации среднего значения уверенности и не может быть использовано для обучения агента.

Показатель ($\overline{IoU_t}$) из (3) является средним значением долей правильно обнаруженных объектов, $\|\delta_t\|_2$ – является мерой величины возмущения и численно отображает, как сильно изменилось изображение. Данная мера является эквивалентом евклидовой нормы для матриц вида (Ш, В, Г). Использование гиперболического тангенса необходимо для сведения учитываемого возмущения к отрезку [0, 1], так как в отличии от других компонент функции наград оно неограниченно.

Дополнительной важной частью пространства действий, не относящейся к формализму Марковских процессов принятия решений, является локализованный характер воздействия: возмущение формируется не по всему изображению, а ограничивается областями, содержащими объекты интереса.

Такое ограничение обусловлено тем, что модели распознавания объектов принимают решение на основе признаков, сконцентрированных в пределах ограниченных пространственных регионов. Внесение возмущения в фоновые области, как правило, не приводит к существенной деградации предсказаний объектов, но увеличивает норму искажения, тем самым ухудшая вычисляемую награду. В связи с этим управление процессом генерации возмущений осуществляется с учётом маски, которая выделяет необходимые области изображения. Данная маска постоянна для каждого изображения (сцены) и не изменяется ни в процессе обучения агента, ни в процессе его работы.

Маска атаки M_t представляет собой бинарную или вещественную матрицу размерности входного изображения, формируемую на основе текущих предсказаний модели:

$$M_t(i,j) = \begin{cases} 1, (i,j) \in \text{рамка объектов} \\ 0, \text{иначе} \end{cases}.$$

Физический смысл введения маски заключается в концентрации возмущения в тех регионах, которые непосредственно влияют на формирование признакового описания объекта. Это позволяет более эффективно нарушать внутреннюю структуру представления. Кроме того, локализация воздействия позволяет снизить визуальную заметность атаки для человека.

Для обучения агента, управляющего процессом генерации состязательных примеров, был выбран алгоритм проксимальной оптимизации политики (Proximal Policy Optimization, PPO) [13]. Выбор метода обучения в данной задаче определяется предложенной средой: высокая размерность состояния, стохастичность переходов.

Предложенный агент формирует непрерывные параметры атаки (ϵ_t, α_t), что требует применения алгоритмов обучения с непрерывным пространством действий. Кроме того, функция награды является нестационарной, так как зависит от текущего состояния нейросетевой модели и динамики изменения предсказаний в процессе атаки. Это накладывает дополнительные требования на устойчивость алгоритма обучения.

Алгоритм PPO оптимизирует параметризованную стратегию $\pi_\theta(a_t|s_t)$ посредством максимизации ожидаемой награды. Важной особенностью PPO является использование ограниченного обновления политики, что достигается за счёт ограничения отношения вероятностей действий:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)], \quad (4)$$

где $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ – отношение вероятностей действий; $\hat{A}_t$ – оценка преимущества; ϵ – параметр ограничения.

Использование данного ограничения (4) позволяет избежать резких изменений политики и обеспечивает более стабильную сходимость по сравнению с классическими методами.

Анализ альтернативных подходов демонстрирует ограниченную применимость методов, опирающихся на функцию ценности (например, DQN), в силу их ориентации на дискретное пространство действий. Помимо этого, алгоритмы данного класса чувствительны к колебаниям функции вознаграждения и требуют применения вспомогательных механизмов стабилизации, таких как целевые сети и буфер повторного опыта, что усложняет процесс обучения.

Алгоритмы, работающие с непрерывным пространством действий, такие как DDPG и TD3, подходят для данной задачи, однако отличаются высокой чувствительностью к подбору гиперпараметров и тенденцией к завышению оценок функции ценности. В условиях нестационарной среды указанные особенности приводят к нестабильности обучения и деградации стратегии.

Таким образом, использование алгоритма PPO с параметрами, которые приведены в таблице 1, определяется его способностью поддерживать стабильную и эффективную оптимизацию

Таблица 1.

Параметры алгоритма PPO для обучения агента генерации

Параметр	Количество шагов обучения	Коэффициент дисконтирования	Количество эпох для обучения на подвыборке
Значение	2048	0,9	5

при непрерывном пространстве действий, изменчивой функции вознаграждения и высокой размерности пространства состояний. Данные характеристики полностью соответствуют специфике генерации состязательных воздействий.

Экспериментальная оценка предложенного метода генерации состязательных примеров и его влияние на результаты состязательного обучения

Экспериментальная часть работы направлена на оценку эффективности и сравнительный анализ предложенного подхода генерации состязательных воздействий на основе MDP. Сравнительный анализ проведён с классическим методом атаки. Данный анализ направлен на оценку разности влияния полученных примеров на устойчивость моделей детекции.

В качестве базовой архитектуры использовалась дообученная модель детекции объектов YOLOv8[3], реализованная в рамках фреймворка Ultralytics и адаптированная под задачу многоклассовой детекции военной техники. Обучение и предсказание модели выполнялись с использованием фреймворка PyTorch[4], который обеспечивает гибкую реализацию и удобную интеграцию процедур обучения и тестирования.

Обучение и тестирование агента и моделей проводились на специализированном наборе данных Military Vehicle Dataset[5]. Данный набор содержит изображения различных типов военной техники в условиях, приближенных к реальным сценам наблюдения с высоты полёта дрона, который представляет из себя элемент РТК. Выбор данного набора данных обусловлен необходимостью оценки устойчивости модели в прикладных задачах, где повышены требования к надежности детекции объектов.

Для реализации обучения агента в постановке MDP использовались инструменты библиотеки Stable-Baselines3[6]. Данная библиотека предоставляет стандартные реализации алгоритмов обучения с подкреплением и средства для построения пользовательских сред взаимодействия с моделью детекции.

3 Ultralytics. YOLOv8: State-of-the-art object detection and segmentation model. – URL: https://github.com/ultralytics/ultralytics (дата обращения: 20.04.2026).

4 Paszke A., Gross S., Massa F. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. – URL: https://pytorch.org (дата обращения: 20.04.2026).

5 Military Vehicle Dataset (Object Detection Dataset). Roboflow Universe. – URL: https://universe.roboflow.com/vgtu-n8zmy/military-vehicle (дата обращения: 20.04.2026).

6 Raffin A. et al. Stable-Baselines3: Reliable Reinforcement Learning Implementations. – URL: https://github.com/DLR-RM/stable-baselines3 (дата обращения: 20.04.2026).

Эксперименты для оценки эффективности и сравнительного анализа включали три основных сценария, отражающих различные режимы воздействия на модель.

Первый сценарий соответствует работе исходной модели без применения атак и используется как базовый уровень для последующего сравнения.

Второй сценарий предполагает применение фиксированной состязательной атаки (PGD), при которой возмущения формируются по детерминированному градиентному правилу без учета состояния модели в процессе атаки. Данный подход с заданными параметрами (табл. 2) используется как классический метод генерации состязательных примеров и служит ориентиром для сравнения. Ввиду особенности генерации состязательных примеров PGD алгоритмом, а именно сильное зашумление исходного изображения, были выбраны параметры, максимально ухудшающие производительность базовой YOLO-модели.

Третий сценарий основан на обучении и использовании предложенного адаптивного MDP-агента, который формирует состязательные воздействия с учетом текущего состояния модели.

Дополнительно рассматривались сценарии состязательного обучения, в рамках которого сгенерированные примеры включались в обучающую выборку. Под этим добавлением понимается расширение исходного набора данных за счет добавления модифицированных изображений. Далее модель дообучается на смеси исходных и атакованных данных, что позволяет повысить устойчивость к подобным возмущениям в дальнейшем.

Для корректного сопоставления результатов между различными методами генерации атак во всех сценариях состязательного обучения используются одинаковые исходные изображения. При этом доля состязательных примеров в обучающей выборке фиксирована и составляет 25 % от общего числа обучающих данных и также являются фиксированными. Такая доля состязательных примеров в обучающих выборках выбран как компромисс между сохранением качества обучения на чистых данных и достаточной интенсивностью воздействия для повышения устойчивости модели.

Экспериментальная постановка охватывает как анализ эффективности различных типов

Таблица 2.

Параметры алгоритма PGD

Параметр	Максимальная доля возмущения за шаг (ϵ)	Размер шага возмущения за шаг генерации (α)	Количество шагов генерации
Значение	0,5	0,1	10

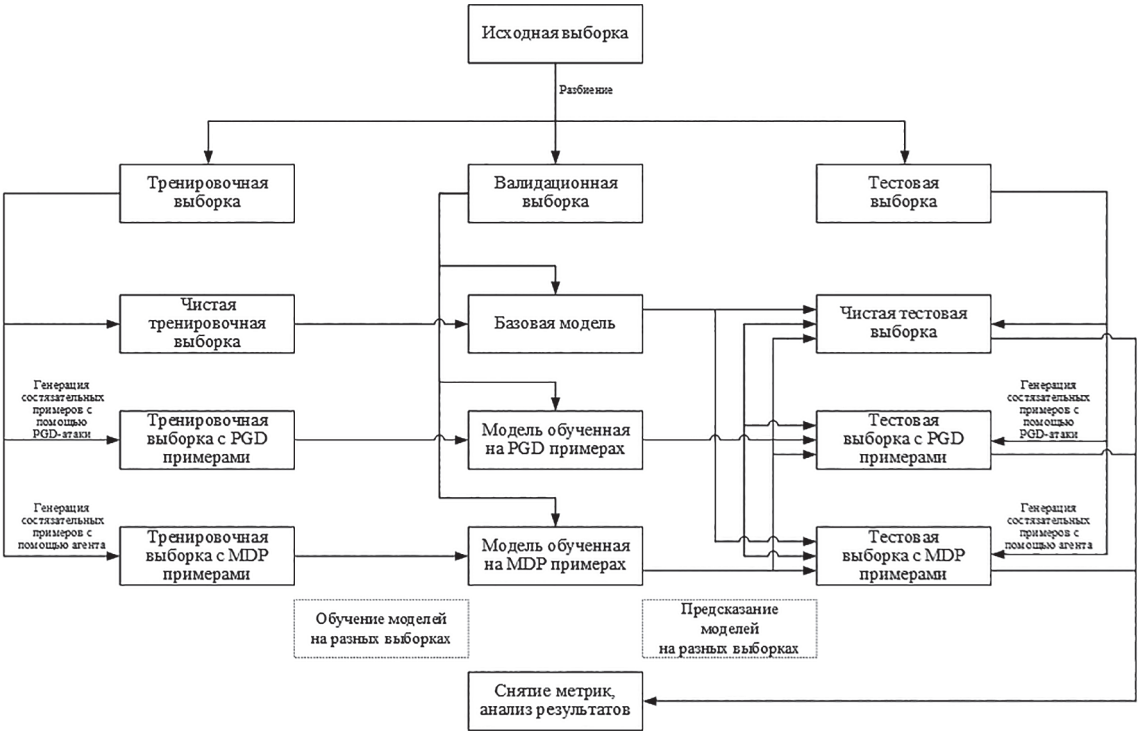


Рис. 2. План проведения эксперимента

атак, так и их влияние на процесс повышения устойчивости модели через состязательное обучение. Всего в ходе валидации результатов оговаривается 9 случаев (предсказание чистых и состязательных примеров двух видов для каждой полученной модели). Итоговый план проведения эксперимента представлен на рисунке 2.

Важным процессом в ходе генерации тренировочных и тестовых выборок является процесс генерации состязательных возмущений, а следовательно, состязательных изображений. Динамики генерации таких примеров относительно метрик-компонент функции наград агента $\overline{C_t}$, $\overline{IoU_t}$ и l_2 и функции наград непосредственно представлены на рисунке 3.

Полученные данные говорят, что с точки зрения влияния подходов на способность модели распознать изменённые объекты, алгоритм PGD обладает значительным преимуществом. В то же время MDP агент решает задачу минимизации не только рассматриваемых метрик, но и задачу минимизации степени возмущения в совокупности.

В соответствии с поставленной задачей анализ полученных результатов обучения моделей на состязательных примерах проводится с точки зрения устойчивости модели к состязательным воздействиям. Оценка эффективности предложенного подхода проводилась на основе совокупности численных метрик, позволяющих анализировать как качество детекции, так и устойчивость модели к состязательным воздействиям.

В качестве первичной оценки используются графики процесса обучения модели в трёх рассматриваемых конфигурациях. Во всех рассматриваемых случаях используются одинаковые параметры для обучения моделей. Данные кривые позволяют сравнить динамику сходимости модели, скорость достижения устойчивого качества и влияние различных типов состязательных данных на процесс оптимизации. Особое внимание уделяется разнице в стабильности обучения и финальном уровне метрик качества детекции.

На рисунке 4 представлены графики динамики функций потерь моделей, обученных на разных вариациях тренировочной выборки (чистой, с 25 % состязательных примеров, полученных с помощью PGD алгоритма и с 25 % состязательных примеров, полученных с помощью агента MDP). На данных графиках потеря точности локализации объекта оценивает точность определения рамки вокруг объекта моделью, Потеря ошибки классификации, оценивает правильность определения моделью класса объекта. Потеря точности границ, отображает точность определения границ объекта непосредственно. На валидации данные функции потерь означают то же самое, но считаются на проверочном наборе и помогают понять, как модель ведёт себя на данных, которых она не видела во время обучения.

В контексте проводимого анализа видно, что модели, обучаемые на состязательных примерах имеют повышенные значения функций потерь относительно обучения базовой модели YOLO на новых данных. Данный факт в совокупности с метриками fitness и fps, представленных на рисунке 5, говорит о том, что моделям тяжелее обучаться на состязательных примерах ввиду их неестественности, так как состязательные возмущения напрямую влияют на карту признаков моделей. В архитектуре YOLO комплексный показатель качества (fitness) служит оценкой работы модели распознования объектов. Он рассчитывается как взвешенная комбинация основных характеристик, включая точность, полноту и среднюю точность (mAP), и применяется для сравнения конфигураций моделей. Скорость обработки кадров (FPS) отражает вычислительную эффективность алгоритма и определяет его пригодность для систем реального времени.

На рисунках 6–8 представлены результаты оценки НС моделей на различных вариантах тестовой выборки. В задачах обнаружения объектов точность характеризует долю верно идентифицированных объектов среди всех предложенных моделью, полнота – долю обнаруженных объектов относительно общего числа реальных, а F1-мера представляет их гармоническое среднее. Показатели mAP50 и mAP50-95 обозначают среднюю точность при пороге пересечения ограничивающих рамок (IoU) 0,5 и усреднённую точность в диапазоне IoU от 0,5 до 0,95 соответственно. Среднее значение IoU отражает степень совпадения предсказанных и эталонных рамок, а средняя уверенность вычисляется как математическое ожидание вероятности,

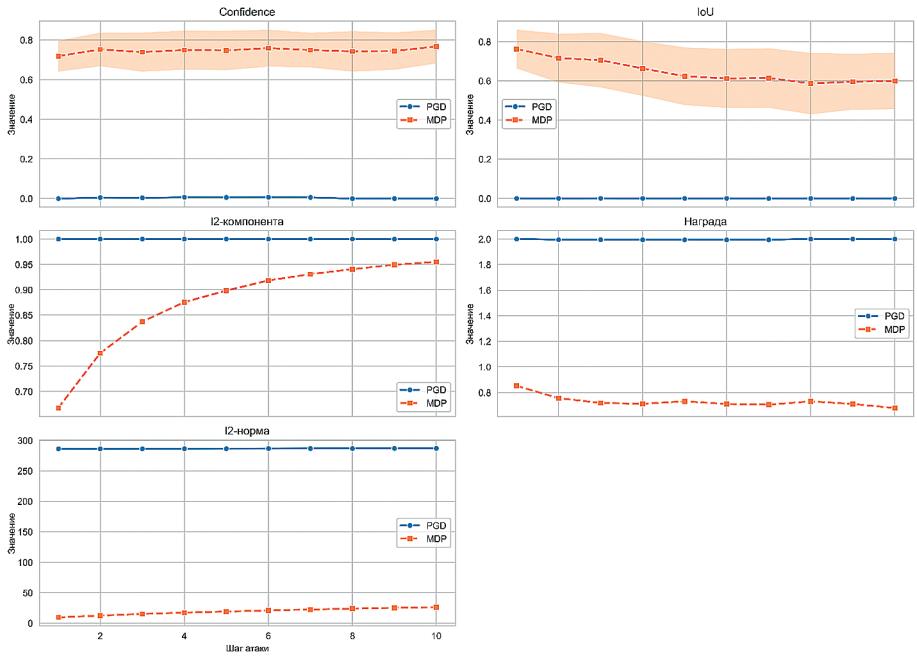


Рис. 3. Динамики генерации состязательных примеров подходами PGD и MDP

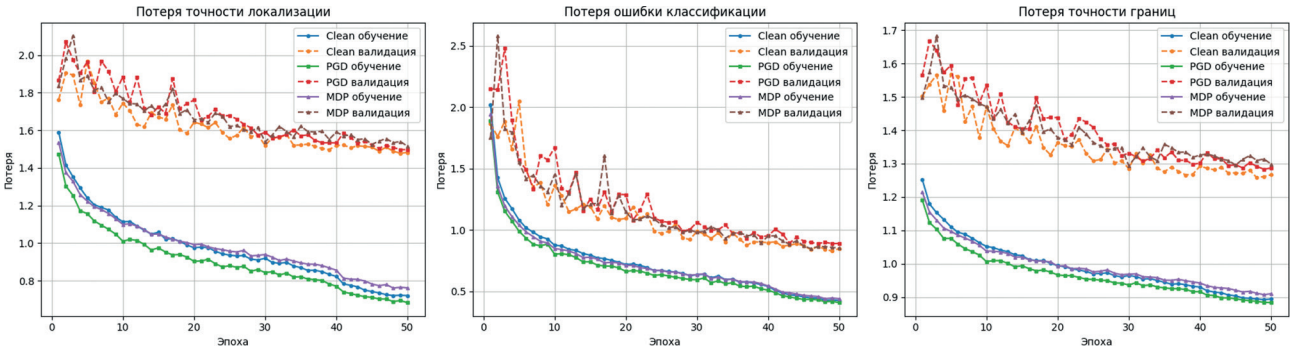


Рис. 4. Динамики функций потерь при обучении YOLO на чистой и состязательных выборках тренировочного набора

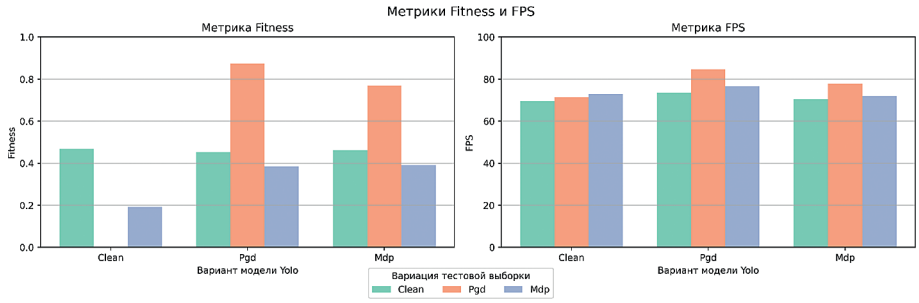


Рис. 5. Метрики качества обучения моделей распознавания

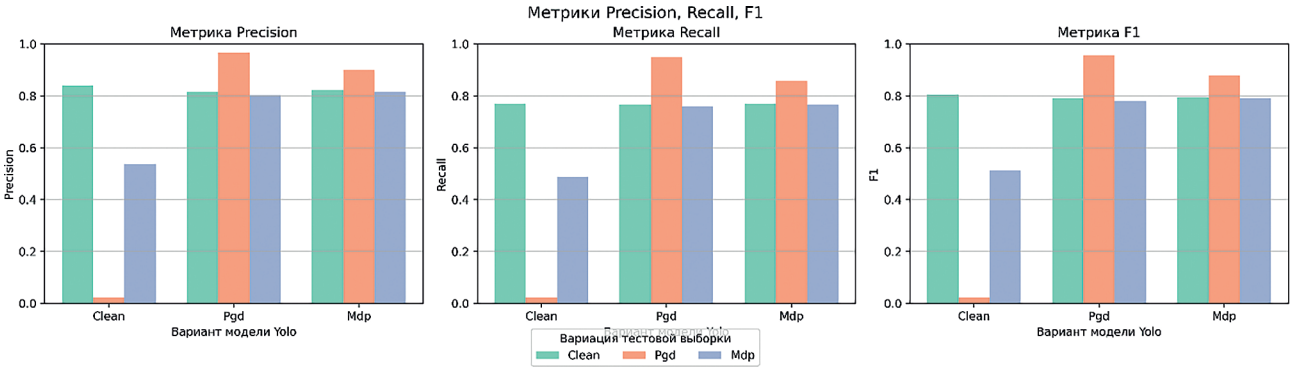


Рис. 6. Метрики precision, recall, F1 моделей распознавания

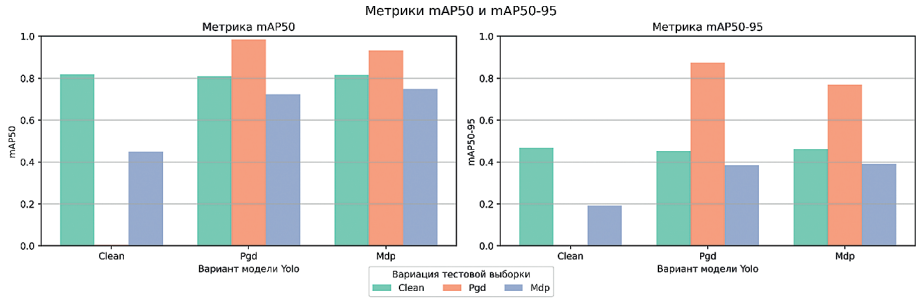


Рис. 7. Метрики mAP50 и mAP50-95 моделей распознавания

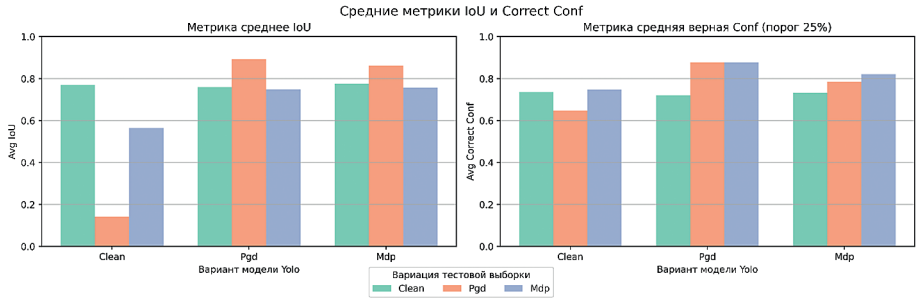


Рис. 8. Метрики IoU и верная Conf моделей распознавания

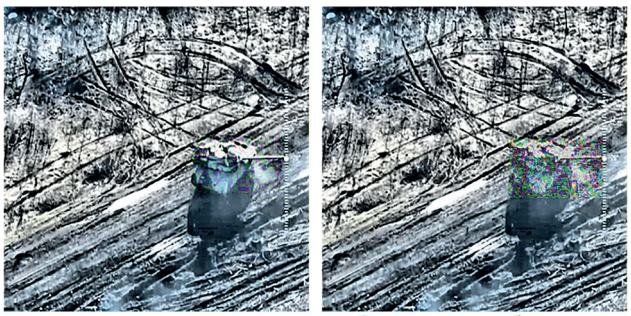
присвоенной моделью верным детекциям при пороге отсечения 0,25.

Базовая модель показывает высокие результаты на исходной выборке, однако её показатели существенно снижаются при обработке данных, подвергнутых атакам PGD и MDP, что свидетельствует о низкой устойчивости к состязательным воздействиям. Модификация, дообученная на примерах, сгенерированных методом PGD (25 % выборки), значительно повышает устойчивость к данному типу атак, демонстрируя наилучшие значения всех основных

метрик. Одновременно с этим модель проявляет признаки избыточной адаптации к специфике PGD: наблюдается завышение уверенности распознавании объектов, при этом на чистой выборке прироста качества не фиксируется. Вариант, использующий генерацию данных по методу MDP, отличается более сбалансированными характеристиками: показатели на атакованных выборках уступают версии, обученной на PGD, но превосходят базовую, а качество работы на исходных данных сохраняется на сопоставимом уровне. Таким образом, дообучение на примерах, полученных с помощью алгоритма PGD, обеспечивает максимальную специализированную устойчивость, тогда как применение MDP способствует лучшей способности к обобщению при умеренном снижении абсолютных значений метрик.

В проведённом эксперименте сравнивались классическая PGD-атака и подход, в котором параметры атаки (ϵ и α (2)) управляются MDP-агентом. Полученные результаты показывают, что, несмотря на способ внесения состязательных возмущений, данные методы в совокупности являются разными и решают разные по смыслу задачи и, соответственно, формируют различные типы состязательных примеров.

PGD, при подборе параметров на максимальное ухудшение качества детекции объектов, генерирует более сильные возмущения, что подтверждается большей l_2-нормой. Такие примеры эффективно снижают качество модели и позволяют обучать более устойчивые модели в условиях наиболее худших сценариев атак. Это достигается ценой сильного искажения изображений: шум визуально заметен (рис. 9), что указывает на выход за пределы естественного распределения данных.



Состязательное изображение, созданное с помощью MDP агента

Состязательное изображение, созданное с помощью алгоритма PGD

Рис. 9. Примеры состязательных изображений, сгенерированных различными способами

В отличие от этого, агент MDP является регулятором параметров атаки. За счёт встроенного штрафа на величину возмущения он находит компромисс между эффективностью атаки и сохранением структуры изображения. В результате генерируются состязательные примеры с существенно меньшей нормой и практически незаметными для человека искажениями (рис. 9). Такие примеры ближе к естественным данным и лучше сохраняют семантику сцены.

Данные различия отражаются на результатах обучения. Модели, обученные на MDP-примерах, демонстрируют сопоставимые, а в ряде случаев более высокие метрики на чистых и MDP-выборках, что говорит о меньшей деградации базового качества. Кроме того, такие модели обладают более стабильным поведением. Они лучше обобщают данные, близкие к исходному распределению. Это делает MDP-подход приоритетным в прикладных сценариях, где важен баланс между устойчивостью и качеством.

Так как MDP является более щадящим подходом, то его использование позволяет применять его как инструмент для регуляризации.

Таким образом, PGD остаётся более сильной атакой в терминах максимального ухудшения качества, тогда как MDP-подход обеспечивает более реалистичные, малозаметные и устойчивые возмущения, выступая эффективным механизмом адаптивной настройки атаки и улучшения обобщающей способности модели.

Заключение

В настоящем исследовании рассмотрена задача повышения устойчивости НС, входящих в состав ИИ СУ РТК, к состязательным атакам. В качестве основного инструмента генерации возмущений использован формализм марковского процесса принятия решений. Данный подход заменяет статические схемы формирования атак на адаптивную стратегию, учитывающую текущее состояние модели распознавания объектов и специфику входных данных. Обученный агент формирует состязательные примеры с ограниченным уровнем искажений. Он сохраняет баланс между снижением точности детекции и сохранением визуальной естественности изображений. По сравнению с классическими итеративными методами (например, PGD), генерируемые возмущения отличаются большей вариативностью и структурной правдоподобностью, что повышает их ценность как тренировочного материала для состязательной настройки моделей.

Экспериментальная проверка подтвердила, что включение сгенерированных примеров в процесс дообучения НС повышает их устойчивость без существенного падения производительности на исходных данных. Подобный

результат важен для робототехнических систем, функционирующих в условиях неполной информации и внешних помех, где необходимы как высокая точность распознавания, так и отказоустойчивость. Метод усиливает безопасность РТК за счёт создания моделей восприятия, устойчивых не только к целевым атакам, но и к естественным шумам окружающей среды. Это напрямую снижает вероятность ошибочных управляющих решений, повышает общую надёжность системы и расширяет допустимую область её применения в задачах повышенной ответственности.

Полученные результаты подтверждают целесообразность применения методов обучения с подкреплением для формирования адаптивных состязательных воздействий. Данное направление представляется перспективным инструментом обеспечения киберфизической безопасности [14–16] интеллектуальных СУ РТК.

В заключение следует отметить, что представленный подход соотносится с результатами работы[7] по решению задачи защиты ИИ СУ РТК от атак отравления данных. В указанном исследовании механизм MDP применялся для контроля входных данных и принятия решений об их использовании в процессе обучения, что позволяло предотвращать деградацию модели за счёт фильтрации потенциально скомпрометированных выборок. В отличие от этого, в данной работе MDP используется на этапе генерации состязательных воздействий, где агент управляет параметрами атаки с целью формирования релевантных примеров для последующего состязательного обучения.

Несмотря на различие в уровне применения (контроль входных данных и активное формирование атакующих воздействий), оба подхода опираются на единую концепцию принятия решений в условиях неопределённости и динамики состояния модели. Их совместное использование позволяет реализовать комплексную стратегию повышения устойчивости ИИ СУ РТК. Данная стратегия объединяет механизмы предотвращения внедрения вредоносных данных и повышения защищённости моделей к атакам уклонения, что в совокупности сужает поверхность для атаки РТК и обеспечивает более высокий уровень устойчивости их интеллектуальных подсистем.

7 Программа генератора агентов для обнаружения атак отравления наборов данных: свидетельство о гос. регистрации программы для ЭВМ RU 2026613545 / правообладатель А. А. Шевченко. – № 2026612373; зарегистрировано 02.02.2026; опубл. 06.02.2026. – 1 электрон. опт. диск (Python; 77521 байт).

Литература

1. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey // IEEE Access. – 2021. – Vol. 9. – P. 12315–12336. – DOI: 10.1109/ACCESS.2021.3052020.
2. Wang Z., Wang X., Liu J., et al. Adversarial attacks on deep learning-based object detection systems: A survey // Neurocomputing. – 2022. – Vol. 493. – P. 1–19. – DOI: 10.1016/j.neucom.2022.03.045.
3. Qiu S., Liu Q., Zhou S., Wu C. Review of adversarial attacks and defense strategies in deep learning-based computer vision systems // Applied Sciences. – 2022. – Vol. 12 (15). – Art. 7583. – DOI: 10.3390/app12157583.
4. Xu H., Ma Y., Liu H., Deb D., Liu H., Tang J., Jain A. D. Adversarial attacks and defenses in images, graphs and text: A review // International Journal of Automation and Computing. – 2021. – Vol. 18 (2). – P. 151–178. – DOI: 10.1007/s11633-021-1274-3.
5. Li Y., Li X., Wang X., et al. Survey of adversarial machine learning in deep neural networks: attacks, defenses and robustness evaluation // ACM Computing Surveys. – 2023. – Vol. 55 (10). – P. 1–38. – DOI: 10.1145/3562695.
6. Pavlitskaya S. et al. Adversarial Vulnerability of Temporal Feature Networks for Object Detection // Sensors. – 2023. – Vol. 23 (23). – DOI: 10.3390/s23239579.
7. Zhou Z., Chen X., Li E., et al. Robustness of deep learning models under real-world corruption: A survey // Pattern Recognition. – 2023. – Vol. 138. – Art. 109386. – DOI: 10.1016/j.patcog.2023.109386.
8. Nguyen K. N. T., Zhang W., Lu K., Wu Y.-H., Zheng X., Tan H. L., Zhen L. A Survey and Evaluation of Adversarial Attacks in Object Detection // IEEE Transactions on Neural Networks and Learning Systems. – 2025. – DOI: 10.1109/TNNLS.2025.3561225.
9. Thunuguntla A., Tadepalli P., Raffa G. Defenses Against Adversarial Attacks on Object Detection: Methods and Future Directions // Information. – 2025. – Vol. 16 (11). – DOI: 10.3390/info16111003.
10. Huang Y., Zhang Y., Wu X., et al. Reinforcement learning for adversarial example generation and defense: A comprehensive survey // IEEE Transactions on Neural Networks and Learning Systems. – 2024. – DOI: 10.1109/TNNLS.2024.3351123.
11. Chen J., Wu Y., Liang J., et al. Robust adversarial reinforcement learning: A survey and taxonomy // Information Sciences. – 2022. – Vol. 608. – P. 1–24. – DOI: 10.1016/j.ins.2022.06.033.
12. Domico K., Sheatsley R., Pauley E., Hanna J., McDaniel P. Adversarial Agents: Black-Box Evasion Attacks with Reinforcement Learning // ResearchGate: научная социальная сеть. – 2025. – DOI: 10.48550/arXiv.2503.01734.
13. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal Policy Optimization Algorithms // arXiv.org: электронный архив. – 2021. – arXiv:1707.06347. – Режим доступа: https://arxiv.org/abs/1707.06347 (дата обращения: 20.04.2026).

14. Липатников В. А., Тихонов В. А. Распознавание вторжений нарушителя при управлении кибербезопасностью инфраструктуры интегрированной организации на основе нейро-нечетких сетей и когнитивного моделирования // В сборнике: Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО, 2019). сборник научных статей VIII Международной научно-технической и научно-методической конференции: в 4 т. – 2019. – С. 659–664.
15. Липатников В. А., Парфиров В. А. Метод поддержки принятия решения при управлении сетью связи на основе технологии нейронных сетей // В сборнике: Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО, 2024). Материалы XIII Международной научно-технической и научно-методической конференции. Санкт-Петербург, 2024. – С. 467–472.
16. Синдеев М. А., Васильев Н. А., Смирнов В. А., Шевченко А. А., Косолапов В. С., Липатников В. А., Парфиров В. А. Программный комплекс для прогнозирования и поддержки принятия решений администратором сети по парированию многоэтапной атаки. Свидетельство о регистрации программы для ЭВМ RU 2022616751, 15.04.2022. Заявка № 2022615448 от 29.03.2022.

PROTECTION AGAINST ADVERSARIAL ATTACKS OF NEURAL NETWORKS CONTROL SYSTEMS FOR ROBOTIC COMPLEXES BASED ON MARKOV PROCESSES

Satsuta A. I.[8], Lipatnikov V. A.[9]

Keywords: *artificial intelligence security, protection of machine learning models, adversarial learning, computer vision, object recognition, control systems for robotic complexes.*

Abstract

Objective: *the purpose of the work is to analyze the vulnerabilities of neural network models of robotic control systems to adversarial influences and to develop a way to increase their protection based on reinforcement learning methods.*

Research method: *an approach based on the formalization of the process of generating adversarial examples as a task of the Markov decision-making process.*

Results of the study: *an agent has been developed that adaptively generates adversarial examples by taking into account the state of the environment and the characteristics of the model. It is proposed to train the agent to generate adversarial examples with subsequent use in adversarial training. The agent's state is formed on the basis of model feature maps, gradient information, and the value of input data distortion, estimated according to the l2 norm. The agent's actions correspond to the choice of PGD attack parameters.*

Scientific novelty: *a new reward function is proposed, which takes into account both the quality of detection and the amount of distortion, which provides a balance between the effectiveness of the attack and its stealth.*

References

1. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey // IEEE Access. – 2021. – Vol. 9. – P. 12315–12336. – DOI: 10.1109/ACCESS.2021.3052020.
2. Wang Z., Wang X., Liu J., et al. Adversarial attacks on deep learning-based object detection systems: A survey // Neurocomputing. – 2022. – Vol. 493. – P. 1–19. – DOI: 10.1016/j.neucom.2022.03.045.
3. Qiu S., Liu Q., Zhou S., Wu C. Review of adversarial attacks and defense strategies in deep learning-based computer vision systems // Applied Sciences. – 2022. – Vol. 12 (15). – Art. 7583. – DOI: 10.3390/app12157583.
4. Xu H., Ma Y., Liu H., Deb D., Liu H., Tang J., Jain A. D. Adversarial attacks and defenses in images, graphs and text: A review // International Journal of Automation and Computing. – 2021. – Vol. 18 (2). – P. 151–178. – DOI: 10.1007/s11633-021-1274-3.
5. Li Y., Li X., Wang X., et al. Survey of adversarial machine learning in deep neural networks: attacks, defenses and robustness evaluation // ACM Computing Surveys. – 2023. – Vol. 55 (10). – P. 1–38. – DOI: 10.1145/3562695.
6. Pavlitskaya S. et al. Adversarial Vulnerability of Temporal Feature Networks for Object Detection // Sensors. – 2023. – Vol. 23 (23). – DOI: 10.3390/s23239579.
7. Zhou Z., Chen X., Li E., et al. Robustness of deep learning models under real-world corruption: A survey // Pattern Recognition. – 2023. – Vol. 138. – Art. 109386. – DOI: 10.1016/j.patcog.2023.109386.
8. Nguyen K. N. T., Zhang W., Lu K., Wu Y.-H., Zheng X., Tan H. L., Zhen L. A Survey and Evaluation of Adversarial Attacks in Object Detection // IEEE Transactions on Neural Networks and Learning Systems. – 2025. – DOI: 10.1109/TNNLS.2025.3561225.

8 ***Anatoly I. Satsuta***, Senior Operator of the Scientific Company of the Military Academy of Communications named after Marshal of the Soviet Union S. M. Budyonny, St. Petersburg, Russia. E-mail: toha.satzuta@yandex.ru

9 ***Valery A. Lipatnikov***, Dr.Sc. of Technical Sciences, Professor, Senior Researcher of the Research Center of the Military Academy of Communications named after Marshal of the Soviet Union S. M. Budyonny, St. Petersburg. E-mail: lipatnikovanl@mail.ru

9. Thunuguntla A., Tadepalli P., Raffa G. Defenses Against Adversarial Attacks on Object Detection: Methods and Future Directions // Information. – 2025. – Vol. 16 (11). – DOI: 10.3390/info16111003.
10. Huang Y., Zhang Y., Wu X., et al. Reinforcement learning for adversarial example generation and defense: A comprehensive survey // IEEE Transactions on Neural Networks and Learning Systems. – 2024. – DOI: 10.1109/TNNLS.2024.3351123.
11. Chen J., Wu Y., Liang J., et al. Robust adversarial reinforcement learning: A survey and taxonomy // Information Sciences. – 2022. – Vol. 608. – P. 1–24. – DOI: 10.1016/j.ins.2022.06.033.
12. Domico K., Sheatsley R., Pauley E., Hanna J., McDaniel P. Adversarial Agents: Black-Box Evasion Attacks with Reinforcement Learning // ResearchGate: nauchnaya social'naya set'. – 2025. – DOI: 10.48550/arXiv.2503.01734.
13. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal Policy Optimization Algorithms // arXiv.org: ehlektronnyj arhiv. – 2021. – arXiv:1707.06347. – Rezhim dostupa: https://arxiv.org/abs/1707.06347 (data obrashcheniya: 20.04.2026).
14. Lipatnikov V. A., Tihonov V. A. Raspoznavanie vtorzhenij narushitelya pri upravlenii kiberbezopasnost'yu infrastruktury integrirovannoj organizacii na osnove nejro-nechetkih setej i kognitivnogo modelirovaniya. V sbornike: Aktual'nye problemy infotelekommunikacij v nauke i obrazovanii (APINO, 2019). sbornik nauchnyh statej VIII Mezhdunarodnoj nauchno-tehnicheskoj i nauchno-metodicheskoj konferencii: v 4 t. – 2019. – S. 659–664.
15. Lipatnikov V. A., Parfirov V. A. Metod podderzhki prinyatiya resheniya pri upravlenii set'yu svyazi na osnove tehnologii nejronnyh setej. V sbornike: Aktual'nye problemy infotelekommunikacij v nauke i obrazovanii (APINO, 2024). Materialy XIII Mezhdunarodnoj nauchno-tehnicheskoj i nauchno-metodicheskoj konferencii. Sankt-Peterburg, 2024. – S. 467-472.
16. Sindeev M. A., Vasil'ev N. A., Smirnov V. A., Shevchenko A. A., Kosolapov V. S., Lipatnikov V. A., Parfirov V. A. Programmnyj kompleks dlya prognozirovaniya i podderzhki prinyatiya reshenij administratorom seti po parirovaniyu mnogoehtapnoj ataki. Svidetel'stvo o registracii programmy dlya EHVM RU 2022616751, 15.04.2022. Zayavka № 2022615448 ot 29.03.2022.