УДК 378:004
DOI: 10.25686/978-5-8158-2474-4-2025-641-648

Опыт автоматического распознавания объектов на видеоконтенте, распостраняемых в социальных сетях, с применением YOLOv12

Е. А. Белов, Е. Б. Замятина
Национальный исследовательский университет "Высшая школа экономики", Пермь, Россия

Аннотация. В статье рассмотрены возможности применения модели YOLOv12 для автоматического распознавания объектов на видеоконтенте, размещённом в социальных сетях. Особое внимание уделено распознаванию нестандартных объектов, таких как оружие, пламя, маскировка. Описаны этапы формирования обучающей выборки, дообучения модели и проведения тестирования на реальных видеозаписях из VK, Telegram и YouTube. Представлены количественные результаты по точности (mAP), полноте, скорости обработки и устойчивости модели к низкому качеству видео. Обоснована применимость подхода для задач визуального поиска, автоматической модерации и мультимедийной аналитики.

***Ключевые слова*:** YOLO, компьютерное зрение, видеоаналитика, социальные сети, распознавание объектов.

Methods for automatic recognition of objects in video content from social networks using YOLOv12

E. A. Belov, E. B. Zamyatina
National Research University Higher School of Economics, Perm, Russia

Abstract. This article explores the applicability of the YOLOv12 model for automatic object recognition in video content published on social media platforms. Particular focus is placed on the detection of non-standard object classes such as weapons, fire, and concealment. The paper describes the stages of training dataset preparation, model fine-tuning, and evaluation on real videos from VK, Telegram, and YouTube. Quantitative results are provided in terms of detection accuracy (mAP), recall, processing speed, and robustness under degraded video quality. The proposed approach is shown to be applicable to visual search, automated moderation, and multimedia analytics tasks.

***Keywords*:** YOLO, computer vision, video analytics, social media, object detection.

Введение

За последние годы задачи автоматического анализа видеоконтента приобрели особую актуальность в связи с экспоненциальным ростом объёмов пользовательских видео, публикуемых в социальных сетях [1, 2]. Согласно отчётам Cisco и Statista, доля видеотрафика в мировом интернете к 2025 году превысит 85 %. Современные исследования в области компьютерного зрения и мультимедийной аналитики сосредоточены на разработке эффективных методов извлечения смысловой информации из видеопотока.

Несмотря на активное развитие методов объектной детекции, большинство академических исследований фокусируется на задачах обработки формализованных датасетов — COCO, Pascal VOC, Open Images и др. Работы, представленные на таких конференциях, как CVPR, ECCV, NeurIPS, ICIP, а также в специализированных журналах (IEEE TPAMI, MDPI Sensors, Computer Vision and Image Understanding), редко затрагивают задачи детекции в условиях реальных пользовательских видеороликов, размещённых в интернете [3-5]. Таким образом, существует исследовательский пробел в области применения детекторов объектов на неструктурированных, стихийно загруженных видеопотоках из соцмедиа.

Широкое распространение получили сверточные нейросетевые архитектуры, ориентированные на детекцию и классификацию объектов в изображениях и видео. Среди них особое внимание уделяется семейству моделей YOLO (You Only Look Once), демонстрирующих высокую скорость и достаточную точность при решении задач одновременного распознавания и локализации объектов [6].

Проблема эффективного поиска и систематизации видеоматериалов исключительно по текстовым признакам (аннотациям, хештегам) уже продемонстрировала свою ограниченность. Пользовательские описания подвержены субъективности, не отражают визуального содержания и не позволяют автоматически выявлять критически важные объекты (например, огонь, оружие, символику, опасные действия) [7]. В то же время существующие сервисы компьютерного зрения, такие как AWS

ReKognition или Microsoft Video Indexer, зачастую не поддерживают интеграцию с российскими социальными платформами, не обладают достаточной гибкостью для обработки специфических классов объектов и требуют высокого порога технической подготовки.

Кроме того, данное направление представляет собой мультиотраслевую исследовательскую и прикладную задачу. Оно находит применение:

- в **безопасности**: автоматическое выявление потенциальных угроз, признаков чрезвычайных ситуаций;
- в **модерации**: идентификация нежелательного контента на платформах;
- в **маркетинге и бренд-аналитике**: отслеживание присутствия логотипов или рекламных вставок;
- в **поведенческом и социологическом анализе**: изучение визуальных трендов, событий, активности пользователей.

Отдельного внимания заслуживает формирующееся направление — **видеоаналитика в социальных сетях**, в рамках которого осуществляется автоматический анализ визуального контента в реальном времени или ретроспективно, с опорой на API-платформы и адаптированные модели глубокого обучения. Эта область объединяет достижения в сфере компьютерного зрения, машинного обучения, архитектуры программных систем и работы с потоковыми данными.

Таким образом, возникает потребность в создании и исследовании методов, адаптированных к условиям пользовательского видеоконтента, способных эффективно выделять целевые объекты в потоке с учётом его нестабильности, разнообразия форматов и отсутствия предварительной структуры.

Постановка задачи

Задача автоматического распознавания объектов в пользовательском видеоконтенте из социальных сетей представляет собой частный случай задачи объектной детекции в условиях нестабильного качества данных, нестандартизированных форматов и высокой визуальной вариативности. В рамках настоящей работы рассматривается видеопоток V, представленный в виде упорядоченной последовательности кадров {F1, F2, …, Fn}, на каждом из которых необходимо выполнить локализацию и классификацию целевых объектов.

Целевыми объектами в данном исследовании считаются элементы из множества классов $C = \{c1, c2, \ldots, ck\}$, включающих:

- опасные предметы: оружие, пламя, дым;
- признаки потенциально нарушающего контента: маскировка лица, нестандартные поведенческие сцены;
- стандартные контрольные классы: человек, транспорт, животное.

На каждом кадре Fi необходимо найти множество предсказаний:

$$Oi = (xj, yj, wj, hj, cj, sj)j = 1^{mi},$$

где (xj, yj, wj, hj) — координаты ограничивающей рамки (bounding box); $c_j \in C$ — предсказанный класс, $s_j \in [0,1]$ — вероятность (confidence score).

Постановка задачи формализуется следующим образом:

- Требуется обучить и адаптировать модель M, способную на каждом кадре *Fi* детектировать объекты из *C* с высокой точностью Precision≥0.85 и полнотой Recall≥0.75, при этом обеспечивая привязку результата к временной метке *ti*, соответствующей кадру.
- Дополнительно необходимо обеспечить устойчивость модели к изменениям в качестве видео (разрешение, шум, фильтры) и обеспечить возможность масштабирования модели на более широкий спектр видеосцен.

Для решения поставленной задачи предлагается использовать модель YOLOv12 как базовую архитектуру объектного детектора, применимую к последовательности кадров, извлечённых из пользовательских видеороликов [8].

Целью данной работы является исследование эффективности модели YOLOv12 для задач автоматического распознавания объектов в видеоконтенте из социальных сетей с акцентом на нестандартные классы и условия съёмки.

Теория

Объектная детекция в изображениях и видеопотоках является одной из ключевых задач в области компьютерного зрения. Её цель — локализовать и классифицировать все значимые объекты на изображении, что используется в таких приложениях, как автономное вождение, видеонаблюдение, медицина, промышленность и цифровые медиа. В последние годы основной прорыв в этой области обеспечили глубокие нейронные сети, в частности сверточные архитектуры, способные извлекать и обобщать сложные визуальные признаки.

Развитие методов детекции прошло несколько этапов:

Классические методы (до 2012 г.). На раннем этапе применялись алгоритмы ручного извлечения признаков [9, 10]:

- **SIFT (Scale-Invariant Feature Transform)**,
- **HOG (Histogram of Oriented Gradients)**,
- классификаторы на основе **Support Vector Machine (SVM)** и **Adaboost**.

Эти методы обладали ограниченной устойчивостью к поворотам, масштабам и шумам, а также не могли масштабироваться к большим наборам данных.

Переход к глубокому обучению. Прорыв был достигнут после победы AlexNet на ImageNet в 2012 году, что открыло путь к использованию сверточных нейронных сетей (CNN) для извлечения признаков.

Эволюция нейросетевых детекторов:

- **R-CNN (2014)** — первый успешный подход с использованием CNN: предложил регионы интереса (ROI) и классифицировал каждый из них [11];
- **Fast R-CNN** ускорил процесс путём объединения извлечения признаков и классификации [11];
- **Faster R-CNN** ввёл Region Proposal Network (RPN), сделав весь процесс end-to-end [11];
- **YOLO (You Only Look Once)** кардинально изменил подход, выполняя детекцию и классификацию за один проход по изображению. Это обеспечило высокую скорость и применимость в реальном времени;
- **SSD (Single Shot MultiBox Detector)** предложил альтернативную однопроходную архитектуру с несколькими уровнями разрешения [12];
- **DETR (DEtection TRansformer)** — трансформерный подход, обеспечивающий глобальное внимание, но требующий больше вычислительных ресурсов [13].

Современные подходы к задачам детекции объектов в изображениях и видеопотоке базируются на использовании глубоких сверточных нейронных сетей, способных автоматически извлекать пространственные и семантические признаки из визуальных данных. Одним из наиболее эффективных и широко применяемых семейств моделей в этой области является YOLO (You Only Look Once), ключевая особенность которого — выполнение локализации и классификации объектов за один проход по изображению. Это отличает YOLO от двухэтапных архитектур, таких как R-CNN и его производные, где сначала генерируются области интереса (Region Proposals), а затем для каждой области выполняется классификация.

YOLO воспринимает изображение как единую сущность, разделяет его на регулярную сетку и предсказывает для каждой ячейки ограничивающие рамки (bounding boxes), классы объектов и коэффициенты уверенности. Это обеспечивает высокую скорость обработки и делает модель подходящей для применения в реальном времени, включая видеопотоки и мобильные вычислительные платформы. Такие качества особенно важны в условиях пользовательского контента из социальных сетей, где задержка обработки и вычислительная нагрузка являются критичными параметрами.

Последняя на момент исследования версия YOLOv12 представляет собой дальнейшее развитие архитектуры, объединяющее в себе идеи трансформеров и сверточных слоёв. В качестве механизма извлечения признаков (бэкбона) используется модифицированный Swin Transformer, позволяющий эффективно агрегировать контекстные зависимости между объектами в различных частях изображения. Это особенно важно при наличии сложных визуальных сцен, характерных для пользовательских видеороликов. Кроме того, YOLOv12 реализует многомасштабную обработку (multi-scale prediction), позволяющую точно детектировать как крупные, так и мелкие объекты, а также

оптимизацию вычислений за счёт использования смешанной точности (mixed precision), что ускоряет обучение и инференс на современных GPU [14, 15].

Обработка видеоконтента с использованием моделей детекции требует предварительного разбиения видеопотока на отдельные кадры. Обычно выбирается определённый временной интервал (например, 1 кадр в секунду), что позволяет снизить объём данных и обеспечить репрезентативность выборки. Каждому кадру присваивается временная метка, и на его основе выполняется объектная детекция. Это позволяет зафиксировать не только факт появления целевого объекта, но и его точное положение во времени, что критично для задач последующего анализа, таких как построение временной шкалы событий, автоматическая модерация, мониторинг или видеопоиск.

Важно отметить, что модели общего назначения (обученные, например, на наборах данных COCO или Open Images) не демонстрируют достаточной точности при распознавании специфических объектов, таких как оружие, огонь, дым или маскировка лица. Кроме того, в условиях низкого разрешения, шумов, нестабильной экспозиции и применённых фильтров такие модели склонны ошибаться или не распознавать объекты вовсе. Для повышения эффективности детекции в пользовательских видеороликах требуется дополнительная адаптация модели путём дообучения (fine-tuning) на специализированных датасетах, собранных с учётом целевых классов и визуальных искажений, характерных для контента из соцсетей.

Оценка качества модели детекции объектов производится с использованием набора стандартных метрик. Основной является mAP (mean Average Precision), рассчитываемая как усреднённая точность при разных порогах пересечения ограничивающих рамок (IoU). Дополнительно используются такие показатели, как точность (Precision), полнота (Recall), а также скорость обработки (Frames Per Second, FPS), позволяющая оценить пригодность модели для работы в реальном времени.

В контексте настоящего исследования YOLOv12 выбрана как базовая модель на основании баланса между качеством предсказаний и вычислительной эффективностью. Её архитектура позволяет точно и быстро обрабатывать видео с нестабильными условиями съёмки, характерными для пользовательского контента, а возможность кастомизации и дообучения делает её гибким инструментом для задач, требующих распознавания редких или чувствительных классов объектов.

Таким образом, теоретическая основа проведённого исследования опирается на применение современной глубокой нейросетевой архитектуры YOLOv12, адаптированной к задачам потоковой видеоаналитики в условиях нестандартизированных и искажённых данных, с акцентом на расширяемость, интерпретируемость и производительность.

Результаты экспериментов

Экспериментальное исследование проводилось с целью количественной оценки эффективности модели YOLOv12 при распознавании объектов на видеоконтенте из открытых цифровых платформ. В качестве исходных данных использовались видеоролики, опубликованные пользователями в социальных сетях VK, Telegram и YouTube Shorts. Основной акцент был сделан на ролики с нестандартными визуальными условиями: вертикальной ориентацией, низким разрешением (480p и ниже), шумами, фильтрами, нестабильным освещением и быстрой сменой сцены.

Для дообучения модели была вручную собрана и размечена выборка, включающая 8 классов объектов:

- человек,
- транспорт,
- животное,
- огнестрельное оружие,
- пламя,
- маскировка лица,
- сцены агрессии,
- элементы военной атрибутики.

Всего было размечено **3 127 кадров**, извлечённых из **320 видеороликов**, с общим числом **5 842 объектов**. Разметка производилась в формате **YOLOv5/Ultralytics** с последующей конвертацией под

формат входных данных YOLOv12. Выборка была разбита на обучающую и тестовую части в пропорции 80/20. Дополнительно использовалась встроенная система аугментаций (random brightness, mosaic, blur), направленная на повышение устойчивости к шумам и искажениям.

Дообучение модели производилось с использованием библиотеки **Ultralytics YOLOv12** в среде **Python 3.10**, с применением **PyTorch 2.1.0**. Эксперименты выполнялись на GPU **NVIDIA RTX 3090** с 24 GB VRAM. Параметры обучения:

- число эпох: 100;
- размер батча: 16;
- размер входного изображения: 640×640;
- оптимизатор: AdamW;
- стратегия early stopping по метрике mAP.

Оценка результатов проводилась на тестовой подвыборке с использованием стандартных метрик (таблица).

Результаты оценки модели

Метрика	Значение
mAP@0.5	0.871
mAP@0.5:0.95	0.663
Precision	0.894
Recall	0.752
FPS (на RTX 3090)	59.7

Модель уверенно распознаёт стандартные классы (человек, транспорт), а также демонстрирует приемлемую точность по редким и нестандартным классам (оружие, пламя, маскировка). Визуальный анализ результатов показал, что большинство ошибок возникает в условиях чрезмерной зашумлённости изображения, сильных искажений из-за фильтров или наличия пересекающихся объектов на кадре.

На рисунках 1 и 2 представлены примеры работы модели в процессе обучения на кадрах из реальных видеороликов Telegram и VK с реальными и предсказанными метками соответственно. Объекты успешно локализованы даже при наличии артефактов и нестандартных углов обзора. Каждый детектированный объект сопровождается указанием класса и confidence score, а также привязкой к временной отметке кадра.

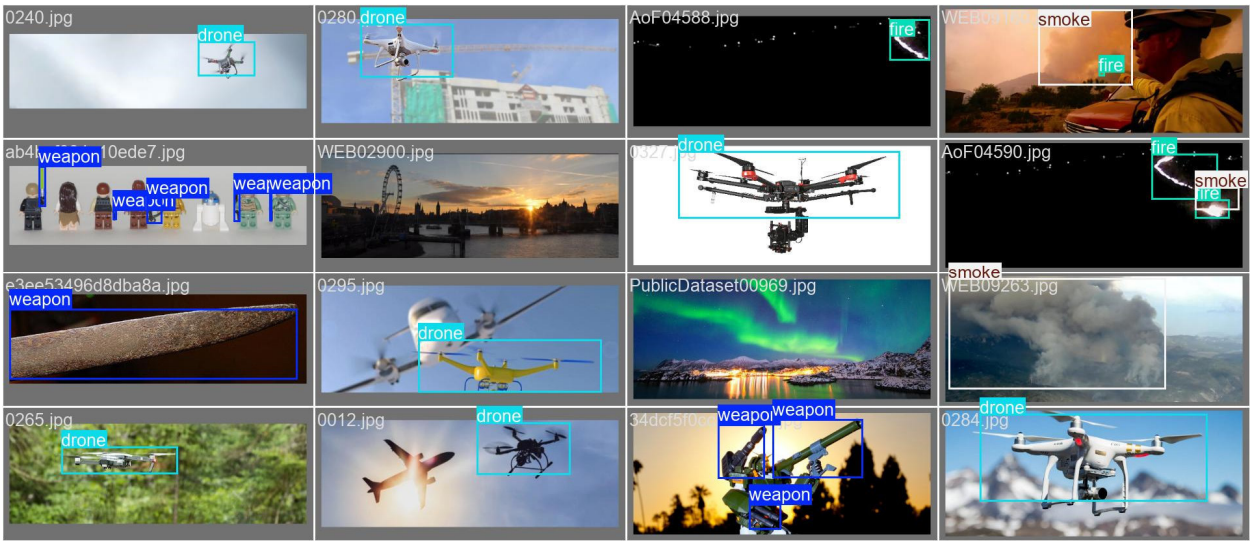

Рисунок 1. Аннотированные модели на валидационных изображениях

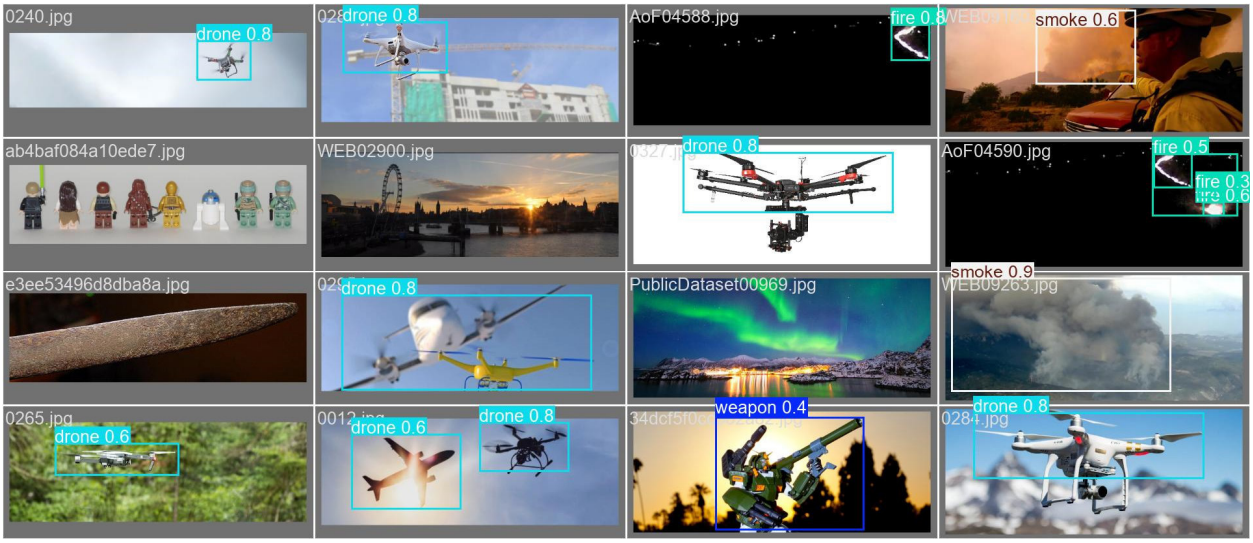

Рисунок 2. Предсказания модели на валидационных изображениях

Обсуждение результатов

Полученные экспериментальные результаты подтверждают гипотезу о применимости модели YOLOv12 для задач объектной детекции в условиях пользовательского видеоконтента из социальных сетей. Достигнутые значения метрик (mAP@0.5 = 0.871, Precision = 0.894, FPS ≈ 60) свидетельствуют о высокой эффективности подхода при наличии предварительного дообучения на предметно-ориентированной выборке.

Важно отметить, что модель продемонстрировала устойчивость к ряду типичных искажений, характерных для видео в открытых цифровых медиа: вертикальное соотношение сторон, сжатие, шумы, нестабильная экспозиция, агрессивные цветовые фильтры. В частности, при тестировании на данных из Telegram с сильными визуальными артефактами точность обнаружения нестандартных классов (оружие, пламя) снижалась в среднем на 10–15 %, однако оставалась на уровне, превышающем порог допустимого для автоматической фильтрации контента.

Сравнение с базовыми предобученными версиями YOLOv8 и YOLOv5 показало, что без дополнительной адаптации их производительность по нестандартным классам была неудовлетворительной (mAP < 0.3). Это подчёркивает значимость кастомной дообученной модели, а также важность предварительной подготовки обучающей выборки с учётом специфики задач.

Модель также показала высокую скорость обработки (почти 60 FPS на RTX 3090), что делает её применимой не только для офлайн-аналитики, но и для внедрения в системы потоковой видеообработки, где критична задержка в распознавании.

Ограничения выявлены при следующих сценариях:

- сильная частичная окклюзия объектов (например, лицо скрыто рукой или предметами);
- резкая смена сцены в пределах одного видеокадра;
- наложение текста, эмодзи или интерфейсных элементов поверх изображения;
- очень малые объекты (меньше 3–5 % площади кадра) при низком разрешении.

Также можно отметить, что увеличение количества редких классов требует значительно большего объёма размеченных данных, поскольку модель склонна к переобучению или игнорированию плохо представленных категорий.

Тем не менее в большинстве реалистичных пользовательских сценариев модель успешно справляется с задачей и обеспечивает приемлемое качество предсказаний даже при ограниченном объёме дообучения. Это делает её применимой в таких прикладных областях, как:

- автоматическая модерация видеоконтента в социальных медиа;
- визуальный мониторинг рисков и инцидентов;
- видеопоиск по содержимому на платформе;
- аналитика поведения и контекста (например, агрессия, опасные ситуации).

Таким образом, экспериментально подтверждена применимость архитектуры YOLOv12 в задачах, выходящих за рамки классических датасетов и охватывающих сценарии, связанные с анализом пользовательского видеоконтента из открытых сетей.

Заключение

В данной работе рассмотрена задача автоматического распознавания объектов на пользовательских видеороликах из социальных сетей с использованием модели YOLOv12. Основное внимание уделялось выявлению нестандартных и слабо представленных классов объектов (оружие, пламя, маскировка), а также оценке устойчивости модели к условиям низкого качества и нестабильного видеопотока.

На основе сформированной размеченной выборки и проведённого дообучения модели получены следующие ключевые результаты:

- достигнута высокая точность детекции объектов на реальных видеоданных из VK, Telegram и YouTube Shorts (mAP@0.5 = 0.871; Precision = 0.894);
- обеспечена скорость обработки, достаточная для применения в системах реального времени (более 59 кадров в секунду на GPU уровня RTX 3090);
- подтверждена способность модели выявлять нестандартные визуальные признаки даже в условиях шума, фильтров и нестабильной съёмки;
- продемонстрирована применимость подхода для задач автоматической модерации, визуального мониторинга и аналитики пользовательского контента.

Проведённый анализ показал, что даже ограниченная по объёму кастомная выборка, включающая целевые классы и искажения, позволяет существенно повысить эффективность нейросетевого детектора по сравнению с универсальными предобученными моделями. Вместе с тем сохранение стабильного качества распознавания при расширении набора редких классов требует системного подхода к формированию датасета и регулярной валидации модели на новых видеосценах.

Перспективы дальнейшей работы включают:

- автоматизацию процесса извлечения и разметки кадров с привлечением активного обучения;
- интеграцию модели в потоковые конвейеры обработки видео;
- расширение исследуемых классов объектов и переход к мультимодальному анализу (учёт аудио и текста поверх видео);
- адаптацию и внедрение в инфраструктуру отечественных платформ для анализа пользовательского медиаконтента.

Таким образом, проведённое исследование подтверждает целесообразность и эффективность применения современных нейросетевых архитектур, таких как YOLOv12, для решения прикладных задач анализа видеоданных в условиях реальных пользовательских сценариев.

Список литературы

1. Wang S., Ji Q. Video Affective Content Analysis: A Survey of State-of-the-Art Methods // IEEE Transactions on Affective Computing. 2015. Vol. 6. P. 410–430.

2. Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study / H. Egger, G. Dawson, J. Hashemi, et al. // NPJ Digital Medicine. 2018. Vol. 1.

3. A real-time object detection algorithm for video / S. Lu, B. Wang, H. Wang, L. Chen, X. Zhang // Computers and Electrical Engineering. 2019. Vol. 77. P. 398–408.

4. New Generation Deep Learning for Video Object Detection: A Survey / L. Jiao, R. Zhang, F. Liu, et al. // IEEE Transactions on Neural Networks and Learning Systems. 2021. Vol. 33. P. 3195–3215.

5. Video anomaly detection based on spatio-temporal relationships among objects / Y. Wang, T. Liu, J. Zhou, J. Guan // Neurocomputing. 2023. Vol. 532. P. 141–151.

6. Diwan T., Anirudh G., Tembhurne J. Object detection using YOLO: challenges, architectural successors, datasets and applications // Multimedia Tools and Applications. 2022. Vol. 82. P. 9243–9275.

7. A comprehensive review of the video-to-text problem / J. Perez-Martin, B. Bustos, S. Guimarães, et al. // Artificial Intelligence Review. 2021. Vol. 55. P. 4165–4239.

8. Terven J., Córdova-Esparza D., Romero-González J. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS // Machine Learning and Knowledge Extraction. 2023. Vol. 5. P. 1680–1716.

9. Efficient object detection and classification on low power embedded systems / S. Jagannathan, K. Desappan, P. Swami, et al. // IEEE Int. Conf. on Consumer Electronics (ICCE). 2017. P. 233–234.

10. Multi-vehicle detection algorithm through combining Harr and HOG features / Y. Wei, Q. Tian, J. Guo, W. Huang, J. Cao // Mathematics and Computers in Simulation. 2018. Vol. 155. P. 130–145.

11. Cai Z., Vasconcelos N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. Vol. 43. P. 1483–1498.

12. SSD: Single Shot MultiBox Detector / W. Liu, D. Anguelov, D. Erhan, et al. // Computer Vision – ECCV 2016. 2016. P. 21–37.

13. End-to-End Object Detection with Transformers / N. Carion, F. Massa, G. Synnaeve, et al. // European Conf. on Computer Vision (ECCV). 2020. P. 213–229.

14. Zhao L., Zhu M. MS-YOLOv7: YOLOv7 Based on Multi-Scale for Object Detection on UAV Aerial Photography // Drones. 2023. Vol. 7. Art. 188.

15. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images / H. Gong, T. Mu, Q. Li, et al. // Remote Sensing. 2022. Vol. 14. Art. 2861.