КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 4 С. 855–870
DOI: 10.20537/2076-7633-2026-18-4-855-870

Ки&М

МОДЕЛИ В ФИЗИКЕ И ТЕХНОЛОГИИ

УДК: 004.93’1

Обнаружение малоразмерных объектов на аэрофотоснимках с использованием сверточных нейронных сетей

С. А. Баженов[a], Р. Р. Ходырев[b], Т. В. Кабанова[c], С. Э. Шипилов[d], Д. А. Вражнов[e], Ю. В. Кистенев[f]

Национальный исследовательский Томский государственный университет,
Россия, 634050, г. Томск, пр. Ленина, д. 36

E-mail: [a] bazhenov.stepan1@gmail.com, [b] hodarev.ruslan@mail.ru, [c] tvk@bk.ru, [d] s.shipilov@gmail.com, [e] vda@iao.ru, [f] yuk@iao.ru

Получено 12.03.2026, после доработки — 05.07.2026
Принято к публикации 05.07.2026.

В работе решается задача обнаружения малоразмерных объектов на аэрофотоснимках в видимом спектре. Высокая вариативность фона и слабая выраженность признаков делают детекцию малоразмерных объектов нетривиальной задачей. В таких условиях низкая эффективность классических методов компьютерного зрения на основе дескрипторов стимулирует переход к архитектурам на основе глубоких нейронных сетей, которые демонстрируют большую обобщающую способность и устойчивость к ложным срабатываниям. В качестве метода поиска объектов на изображении был выбран одноэтапный подход на базе нейросетевой предиктивной модели. Проведен анализ нейросетевых архитектур, относящихся к данному классу, приведены их достоинства и недостатки. В качестве базовой архитектуры детектора была взята YOLO версии 11 с добавлением блока многомасштабной агрегации признаков (используется для объединения информации, полученной из слоев искусственной нейронной сети с разными масштабами) и блока селективной интеграции с учетом размерности (блок автоматически определяет, по какому измерению (канал, высота, ширина) обрабатывать данные, и выборочно объединяет их). Данные модификации направлены как на повышение вычислительной эффективности и возможности развертывания нейросетевых моделей на борту беспилотных летательных аппаратов для анализа изображений в режиме реального времени, так и для улучшения точности детекции объектов малых размеров. Учитывая сложность аннотации изображений, для обучения и тестирования (в соотношении 90/10 % соответственно) использовалась открытая синтетическая база изображений, содержащая порядка 4000 изображений. Обучающая выборка была расширена путем различных случайных аугментаций. Для оценки качества полученных предиктивных моделей использовалось среднее значение точности по всем классам как с порогом 50 % пересечения с экспертной разметкой, так и с варьируемым порогом 50–95 %. Переобучение контролировалось при помощи анализа кривых потерь. Предложенные модификации архитектуры YOLO позволили уменьшить время обработки изображений в два раза при сохранении точности распознавания.

Ключевые слова: распознавание малоразмерных объектов, компьютерное зрение, сверточные нейронные сети, аэрофотоснимки

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (соглашение № 075-15-2024-557 от 25.04.2024 г.).

COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 4 P. 855–870
DOI: 10.20537/2076-7633-2026-18-4-855-870

КиМ

MODELS IN PHYSICS AND TECHNOLOGY

UDC: 004.93'1

Small object detection in aerial images using convolutional neural networks

S. A. Bazhenov[a], R. R. Khodyrev[b], T. V. Kabanova[c], S. E. Shipilov[d], D. A. Vrazhnov[e], Yu. V. Kistenev[f]

Tomsk State University,
36 Lenin ave., Tomsk, 634050, Russia

E-mail: [a] bazhenov.stepan1@gmail.com, [b] hodarev.ruslan@mail.ru, [c] tvk@bk.ru, [d] s.shipilov@gmail.com, [e] vda@iao.ru, [f] yuk@iao.ru

Received 12.03.2026, after completion — 05.07.2026
Accepted for publication 05.07.2026.

This paper addresses the problem of detecting small objects in visible-spectrum aerial imagery. High background variability and weak feature saliency make small object detection a non-trivial task. Under such conditions, classical computer vision algorithms based on hand-crafted descriptors exhibit low efficiency, prompting a shift towards deep neural network architectures, which demonstrate superior generalization capability and robustness to false positives. We selected a one-stage approach based on a neural network predictive model as the primary object detection method. Also, several neural network architectures belonging to this class were analyzed, outlining their advantages and disadvantages. As the baseline detector, we adopted YOLO version 11 and incorporated a multi-scale feature aggregation module (which combines information from neural network layers operating at different scales) and a dimension-aware selective integration module (which automatically determines the dimension (channel, height, or width) along which to process features and fuses them selectively). These modifications aim to both enhance computational efficiency, enabling deployment of neural network models onboard aerial vehicles for real-time image analysis, and improve small object detection accuracy. Given the complexity of image annotation, we used an open synthetic image database containing approximately 4 000 images for training and testing (with a 90/10 split, respectively). We extend the training set using various random augmentation techniques. To evaluate the performance of the resulting predictive models, we employed mean Average Precision across all classes, using both a fixed 50 % intersection-over-union threshold and a varying threshold from 50 % to 95 %. Overfitting was monitored by analyzing loss curves during training process. The proposed modifications to the YOLO architecture reduced image processing time by a factor of two while maintaining detection accuracy.

Keywords: small object detection, computer vision, convolutional neural networks, aerial imagery

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 4, pp. 855–870 (Russian).

This research was funded by the Ministry of Science and Higher Education of the Russian Federation grant number 075-15-2024-557 dated 04/25/2024.

1. Введение

Развитие беспилотных авиационных систем открыло новые возможности в области дистанционного зондирования Земли и мониторинга труднодоступных территорий. Благодаря высокой маневренности, относительно низкой стоимости летного часа и возможности оснащения различными типами сенсоров летательные аппараты (ЛА) стали незаменимым инструментом для решения широкого круга прикладных задач — от сельскохозяйственного мониторинга до инспекции промышленных объектов и поисков опасных объектов, в том числе в рамках гуманитарного разминирования.

Особый практический интерес представляет задача автоматического распознавания объектов на аэрофотоснимках, получаемых с ЛА. В отличие от спутниковых снимков такие изображения характеризуются высоким пространственным разрешением, возможностью получения изображений в различных ракурсах, повторными съемками одного объекта для получения изображений без артефактов движения. В таких условиях ключевыми проблемами детекции являются частичное перекрытие или заслонение объектов (окклюзия) и их схожесть с фоном. Немаловажным аспектом в задачах, связанных с необходимостью оперативного реагирования, является реализация алгоритмов в режиме реального времени, причем в рамках мобильных вычислительных платформ, допускающих установку на ЛА. Увеличение размерности детектирующих матриц существенно увеличивает вычислительную нагрузку. Таким образом, вычислительная эффективность является важным требованием к подобным системам.

К малоразмерным объектам традиционно относят как естественные образования, так и антропогенные предметы, размер которых на снимке сопоставим с шумовой компонентой. Их распознавание осложнено низкой различимостью текстурных и геометрических признаков, высокой вариативностью фоновой обстановки, а также влиянием условий освещения и маскировкой естественным ландшафтом. Классические алгоритмы компьютерного зрения, основанные на ручном подборе дескрипторов, на таких выборках демонстрируют низкую обобщающую способность и высокий уровень ложных срабатываний, что привело к активному привлечению глубоких нейронных сетей для решения данной проблемы.

В рамках задачи обнаружения объектов на изображениях выделяют две подзадачи:

а) обнаружение на изображении подозрительных фрагментов (детекция кандидатов);

б) идентификация данных фрагментов путем присвоения им меток принадлежности заданному классу (распознавание).

В настоящее время используется одноэтапное (ОЭ) и многоэтапное (МЭ) обнаружение объектов, различающееся тем, что в ОЭ подзадачи (а) и (б) архитектурно объединены в рамках одной искусственной нейронной сети (ИНС), а в МЭ это независимые этапы [Juhartini et al., 2025]. Это позволяет снизить вычислительную нагрузку в рамках ОЭ-подхода, позволяя осуществлять обработку данных в реальном времени. Предиктивные модели на основе МЭ-подхода потенциально более точны, однако в настоящее время эта разница становится незначительной [Chen et al., 2024; Muzammul, Li, 2025].

В рамках подзадачи (а) применяется детекция объектов путем выделения прямоугольных фрагментов на изображениях (рамки, bounding boxes). Такой подход упрощает задачу путем стандартизации интерфейса распознающего модуля, так как на вход последнего подаются рамки заданных размеров и исключается из конвейера обработки изображений сложный этап сегментации [Aziz et al., 2020].

Также при решении подзадачи (а) в целях экономии ресурсов при обучении глубоких нейросетей используется подход «трансферное обучение» [Kumar et al., 2022]. Его суть заключается в переиспользовании предварительно обученных весов нейросети в новых задачах. Для этого

в архитектуре нейросетей выделяют две последовательные составляющие: «основа» (англ. backbone) и «голова» (англ. head) [Evci et al., 2022]. Слои, формирующие «основу», принимают на вход сырые данные (изображение) и последовательно, слой за слоем, преобразуют их в набор абстрактных, семантически значимых признаков. На выходе «основы» — многомерное представление входных данных, так называемые карты признаков. Слои «головы» осуществляют классификацию карт признаков путем отнесения последних к одному из известных классов. Основную вычислительную нагрузку при обучении глубоких нейросетей несет «основа», поэтому ее переиспользование позволяет существенно сокращать время. Таким образом, учитывая точность предиктивных моделей и вычислительную эффективность, наиболее перспективным является ОЭ-подход.

На сегодняшний день доминирующее положение среди архитектур для детекции и распознавания объектов на изображениях в рамках ОЭ-подхода занимает YouOnlyLookOnce (YOLO). Актуальная версия архитектуры — 11, поддерживающая модуль Convolutional block with Parallel Spatial Attention (C2PSA), повышающий точность для мелких и перекрывающихся объектов и блок C3k2 для эффективности [Khanam, Hussain, 2024]. Блок C2PSA работает по принципу двойного (параллельного) внимания, анализируя два аспекта признаков, извлеченных сверточными слоями: «канальное внимание», определяющее важные для детекции объекта признаки, и «пространственное внимание», области интереса на изображении. Блок Cross Stage Partial Network with 3 convolutions (C3k2–C3) представляет собой стандартный блок из трех сверток с механизмом частичного соединения между стадиями. Он используется для извлечения и объединения признаков. K2 означает, что внутри этого блока используется архитектурный паттерн «узкое место» (bottleneck) с двумя сверточными ядрами, которые работают параллельно. Это основное отличие от предшественников. Архитектура YOLO предлагает несколько предобученных размеров нейросети (n, s, m, l, x), позволяющих осуществить ее развертывание как на Raspberry Pi, так и на крупных серверах с графическими ускорителями.

Альтернативными YOLO архитектурами в рамках ОЭ-подхода являются Single Shot MultiBox Detector (SSD) [Liu et al., 2016], RetinaNet [Lin et al., 2017], CenterNet [Duan et al., 2019], EfficientDet [Tan et al., 2020]. В SSD используются многоуровневые карты признаков для детекции объектов разных размеров в одном проходе. Изначально именно такой подход обеспечивал высокую точность как детекции, так и распознавания мелких объектов. Принимая во внимание вычислительную сложность данной модели, авторами был применен подход «трансферное обучение» с использованием «основы» на базе архитектуры VGG-16 (Visual Geometry Group) [Benali Amjoud, Amrouch, 2020]. Сеть VGG-16 состоит из 16 слоев с обучаемыми весами (13 сверточных и 3 полносвязных). В ее основе лежит следующий принцип: стеки из 2–3 сверточных слоев с фильтрами малого размера (3×3), за которыми следует слой пулинга (объединения) для уменьшения размерности. Повсеместное использование сверток 3×3 (вместо больших, как в AlexNet) позволяет сократить число параметров при том же рецептивном поле и увеличить глубину нелинейностей.

RetinaNet — одноэтапный детектор, предложенный в 2017 году. Его ключевая инновация в использовании фокальной функции потерь, которая решает проблему дисбаланса между простыми (фоновыми) и сложными (объектными) примерами. Это позволило одноэтапным детекторам впервые достичь точности, сопоставимой с двухэтапными, сохранив при этом высокую скорость. Его архитектура основана на Feature Pyramid Network (FPN) для обнаружения объектов на различных масштабах.

CenterNet — одноэтапный «безъякорный» детектор, представленный в 2019 году. Он рассматривает объект как единую точку, являющуюся центром его ограничивающей рамки (якорем). Сеть не учитывает остальные свойства объекта, такие как его размер и ориентация. Этот подход

позволяет отказаться от трудоемкого процесса настройки якорей и операции подавления немаксимумов, упрощая архитектуру детектора.

EfficientDet — семейство высокоэффективных детекторов, предложенное в 2020 году. В нем использованы два ключевых нововведения: взвешенная двунаправленная пирамида признаков (позволяющая эффективно объединять многомасштабные признаки, что критически важно для детекции объектов разного размера) и метод комплексного масштабирования (compound scaling) всей сети.

Ниже приведены недостатки альтернативных методов по отношению к YOLOv11.

Метод SSD использует якоря, которые необходимо тщательно настраивать под конкретную задачу и массив данных. Неправильно подобранные якоря являются одной из основных причин низкой эффективности SSD при работе с мелкими объектами. Кроме того, SSD обладает более низкой точностью локализации объектов по сравнению с YOLOv11 [Wei et al., 2025].

Сети типа RetinaNet с фокальной потерей способны фокусироваться на сложных примерах, однако стандартная архитектура RetinaNet может терять информацию о мелких объектах из-за многократного уменьшения размерности (субдискретизации) при извлечении признаков. В результате модель нередко пропускает объекты малого размера. RetinaNet плохо определяет точные границы объектов, особенно при высоком пороге Intersection over Union (IoU). RetinaNet — одна из самых вычислительно «тяжелых» моделей. Базовая версия RetinaNet с архитектурой ResNet-50 имеет около 138 миллионов параметров. Это затрудняет развертывание данной модели на БПЛА [Aldubaikhi, Patel, 2025].

У сетей с архитектурой CenterNet возникают проблемы с перекрытием объектов. Подход с детекцией по одной центральной точке создает неоднозначность: если центры двух объектов близки, сеть может детектировать только один из них. При высокой вариативности фона (что характерно для аэрофотоснимков) модели CenterNet могут давать большее количество ложных срабатываний или пропускать цели. Использование архитектуры Hourglass-104 в качестве «основы» для достижения высокой точности приводит к тому, что CenterNet не всегда может обеспечить обработку в реальном времени, особенно на ограниченных вычислительных мощностях. Вычислительная сложность CenterNet растет пропорционально квадрату количества предсказаний, что делает его особенно ресурсоемким при использовании высокого разрешения входных изображений [Duan et al., 2023].

EfficientDet на этапе постобработки использует операцию подавления немаксимумов для фильтрации перекрывающихся рамок, что может стать серьезным узким местом, ограничивающим скорость обработки в реальном времени. Это критично для систем, работающих на БПЛА. Модель требует значительно больше усилий и вычислительных ресурсов для обучения и тонкой настройки гиперпараметров по сравнению с YOLOv11, у которой этот процесс хорошо оптимизирован. Несмотря на высокую эффективность по параметрам, по скорости работы EfficientDet на практике часто уступает YOLOv11 [Wang et al., 2021]. В работе [Mirzaei et al., 2023] показано, что в задачах с большим количеством мелких объектов на изображении (как на аэрофотоснимках) эффективность EfficientDet снижается.

Таким образом, проведенный анализ существующих решений позволяет сделать вывод о перспективности модификации архитектуры YOLOv11 в решения поставленной задачи.

Данная работа направлена на решение проблемы, заключающейся в необходимости разработки специализированных подходов к детекции малоразмерных объектов, которые, с одной стороны, обеспечивают высокую точность распознавания в гетерогенных условиях, а с другой — пригодны для развертывания на аппаратной платформе ЛА в режиме, приближенном к реальному времени.

В статье представлены результаты оптимизации модифицированной архитектуры нейросетевой модели для повышения эффективности обнаружения целевых объектов малого размера.

Основное внимание уделено достижению выигрыша в вычислительной сложности без потери качества классификации.

Новизна работы заключается в том, что впервые предложена гибридная архитектура на базе YOLOv11, объединяющая несколько специализированных модулей: MFAM (Multi-scale Feature Aggregation Module) [Zhang et al., 2024], который предназначен для многомасштабного сбора признаков с использованием параллельных глубинных сверток (ядра $3 \times 3 \ldots 9 \times 9$) и skip-соединений; DASI (Dimension-Aware Selective Integration Module) [Lu et al., 2025], который предназначен для адаптивного взвешенного слияния низкоуровневых (детальных) и высокоуровневых (контекстных) признаков с управлением через механизм gating. RAT (Region Attention Transformer) [Yang et al., 2024], ключевая идея которого заключается в использовании «регионального внимания» (Region Attention), которое заменяет стандартное «многоголовое самовнимание» (Multihead Self Attention), работающее по всему изображению или его фиксированным участкам. Ранее такие блоки применялись по отдельности или в других задачах; их совместное использование и адаптация внутри YOLOv11 для детекции малоразмерных объектов на аэрофотоснимках выполнены впервые.

В качестве объекта исследования были выбраны синтетические изображения мин («лепесток»), находящиеся в открытом доступе, которые могут быть зарегистрированы при помощи ЛА в рамках задачи гуманитарного разминирования [Vivoli et al., 2024]. Известные массивы данных для детекции на аэрофотоснимках, например VisDrone [Zhu et al., 2021], UAVDT [Du et al., 2018], DOTA [Xia et al., 2018], не содержат таргетные объекты, однако могут быть использованы для расширения выборки путем искусственного добавления объектов интереса.

2. Материалы и методы

Для обучения моделей использовался синтетический массив данных [Synthetic PFM-1 Dataset Computer Vision Model, 2026]. Изображения в массиве данных прошли следующую предобработку:

- изменение размера (Resize): масштабирование до разрешения 640×640 пикселей;
- конвертация в оттенки серого (Grayscale).

Кроме того, для каждого исходного изображения были использованы следующие аугментации:

- отражение (Flip) по горизонтали;
- поворот на 90° (90° Rotate) по часовой стрелке, против часовой стрелки, переворот на 180°;
- случайное изменение яркости в диапазоне от −25 % до +25 %;
- размытие при помощи функции Гаусса в радиусе до 4,5 пикселей;
- добавлено до 5 % случайных шумовых пикселей;
- mosaic (значение 1): 4 разных изображения склеиваются в одно мозаичное, на каждое из 4 изображений применяются индивидуальные аугментации (например, изменение цвета, поворот);
- fliplr (значение 0,5): с вероятностью 50 % зеркально отражает изображение по горизонтали;
- degrees (значение 5): случайно поворачивает изображение на угол в диапазоне от −5,0 до 5,0 градусов;

- translate (значение 0,02): случайно сдвигает изображение по горизонтали и вертикали на до 2 % от его размеров;
- scale (значение 0,05): случайным образом масштабирует изображение в диапазоне от 95 % до 105 % от исходного размера;
- close_mosaic (значение 10): отключает mosaic-аугментацию за 10 эпох до конца обучения;
- hsv_h (значение 0,01): случайно изменяет оттенок изображения на величину до ±1 % от полного диапазона;
- hsv_s (значение 0,5): случайно изменяет насыщенность цветов в диапазоне от 0,5 (50 %) до 1,5 (150 %) от исходной;
- hsv_v (значение 0,4): случайно изменяет яркость изображения в диапазоне от 0,4 (40 %) до 1,4 (140 %) от исходной;
- mixup (значение 0,15): с вероятностью 15 % берет два изображения и линейно их интерполирует (смешивает пиксели), рамки и метки классов также смешиваются соответственно;
- copy_paste (значение 0,2): с вероятностью 20 % копирует случайные объекты (вместе с сегментацией или рамкой) с одного изображения и вставляет их на другое изображение;
- erasing (значение 0,3): с вероятностью 30 % случайным образом выбирает прямоугольную область на изображении и заполняет ее случайными пикселями или нулями.

Разделение массива данных было следующим:

- обучающая выборка — 3148 изображений;
- валидационная выборка — 397 изображений;
- тестовая выборка — 393 изображений.

При обучении случайно сгенерированные подмножества обучающей и валидационной выборок использовались для итерационного обучения модели.

Обучение YOLO

Для ускорения обучения и процесса инференса были реализованы блоки MFAM, DASI и RAT. Блок-схемы MFAM, DASI и RAT приведены на рис. 1, 2 и 3 соответственно. DWConv обозначает раздельную глубинную свертку, а Conv 1×1 — точечную свертку.

MFAM решает проблему обнаружения мелких объектов на аэрофотоснимках с ЛА за счет эффективного сбора признаков разных масштабов. Ключевой механизм заключается в использовании параллельных глубинных сверток с разными размерами ядер (3×3, 5×5, 7×7, 9×9). На этом этапе каждый канал обрабатывается отдельно. Это позволяет одновременно захватывать детали и контекст на различных уровнях. Полученные карты признаков объединяются друг с другом. Также используются skip-соединения для расширения рецептивного поля сети без чрезмерного роста вычислительной сложности. DASI позволяет дополнительно усилить эффект слияния мультимасштабных признаков при балансировке глобального контекста и локальных деталей. Основная идея заключается в выполнении адаптивного взвешенного объединения низкоуровневых высокодетализированных признаков и высокоуровневых наполненных глобальным контекстом признаков. DASI использует механизм управления (gating), чтобы динамически определять, какой вклад должны вносить признаки разной размерности на разных этапах обработки.

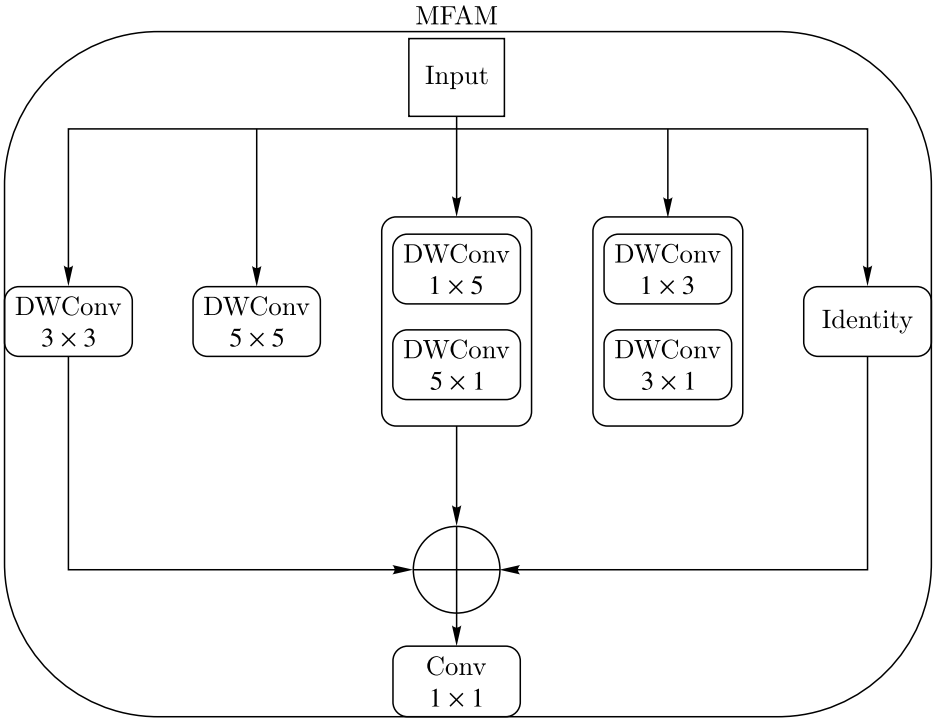

Рис. 1. Блок-схема MFAM (MFAM block flowchart) [Zhang et al., 2024]

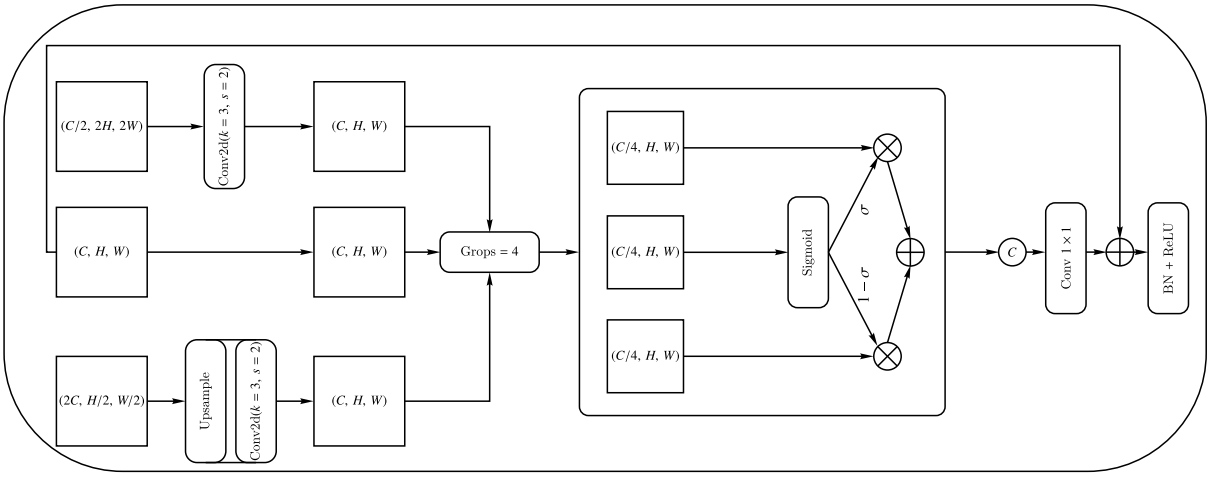

Рис. 2. Блок-схема DASI (DASI block flowchart) [Lu et al., 2025]

Это помогает сети более эффективно интегрировать информацию. RAT-блок позволяет трансформеру «видеть» и обрабатывать изображение не просто как набор пикселей или квадратиков, а как совокупность осмысленных объектов или частей, что значительно повышает качество восстановления.

Обучение и тестирование моделей производилось на системе с процессором AMD EPYC 7643 48-Core (базовая частота — 2,3 ГГц, максимальная в турбо-режиме — до 3,6 ГГц, 48 ядер, 96 потоков, архитектура Zen 3, объем кэша первого уровня — 4,6 Мб, второго — 24 Мб, третьего — 256 Мб) и графическим ускорителем NVIDIA RTX A5000 (архитектура NVIDIA Ampere, ядер CUDA 8192, RT (2-го пок.) — 64, тензорных (3-го пок.) — 256, объем памяти типа GDDR6 — 24 Гб). Следует отметить, что время инференса будет зависеть от конфигурации мобильного

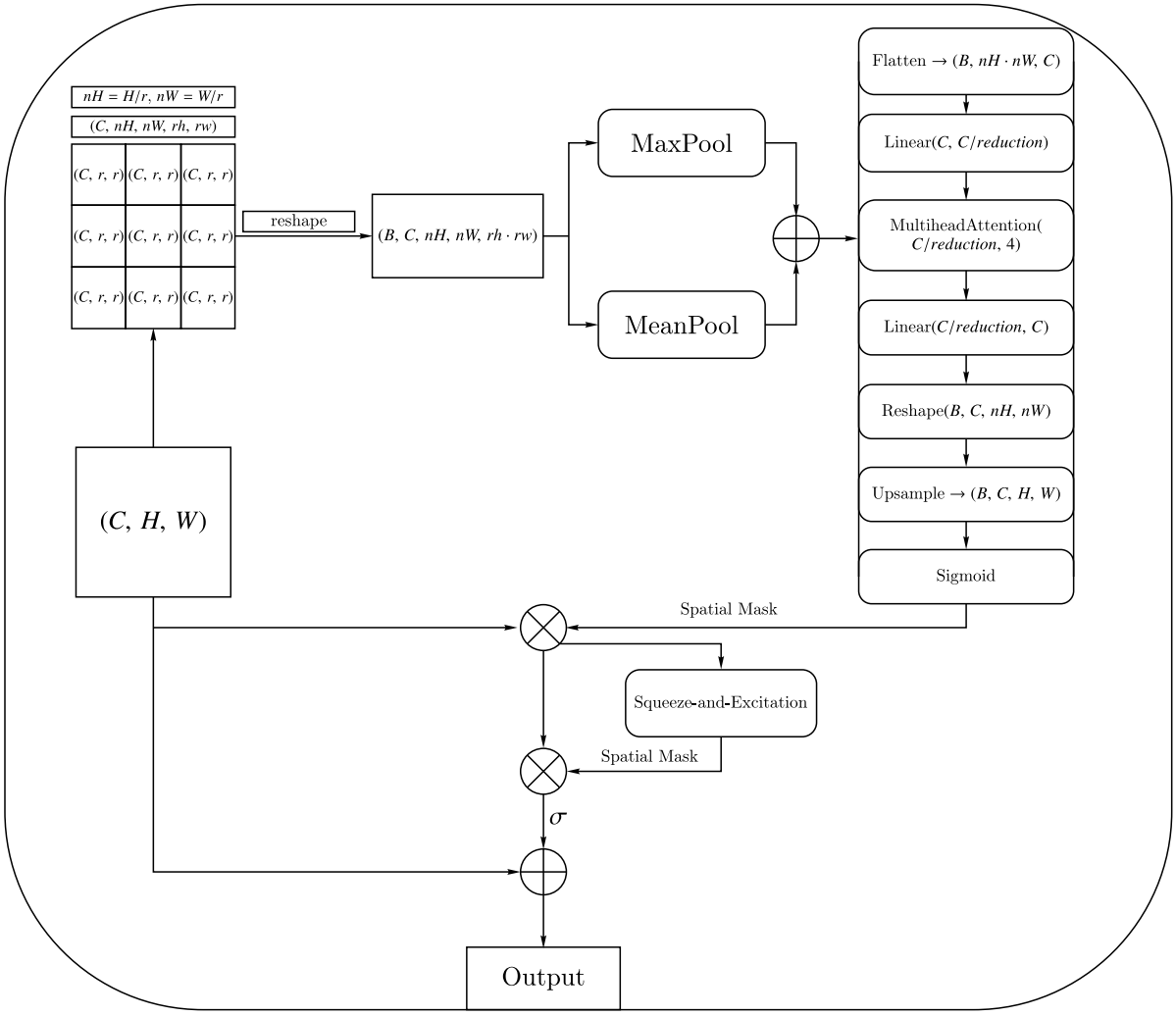

Рис. 3. Блок-схема RAT (RAT block flowchart) [Yang et al., 2024]

вычислителя, устанавливаемого на ЛА, однако величина ускорения будет пропорциональна полученным результатам.

В качестве метрик для оценки качества предиктивных моделей использовались прецезионность (precision, показывающая долю обнаруженных моделью объектов, являющихся действительно верными), полнота (recall, показывающая долю найденных реальных объектов на изображении), mAP (mean Average Precision, среднее значение точности по всем классам) и mAP50-95. Метрики mAP, mAP50-95 опираются на вычисление индекса Жаккара (также используется аббревиатура IoU — Intersection over union) или отношение меры пересечения двух множеств к мере их объединения. Чем больше множества пересекаются, тем ближе данный индекс к 1. Метрика mAP50-95 считается как среднее арифметическое между значениями mAP при разных порогах IoU, начиная с 0,5 и заканчивая 0,95, с шагом 0,05. Использование одновременно двух этих метрик позволяет исключить случайное угадывание положения рамки, содержащей искомый объект, поскольку mAP будет высоким, а mAP50-95 — низким.

Для качества обучения и валидации использовались одинаковые метрики box_loss, cls_loss, dfl_loss. Box_loss показывает точность нахождения координат рамок, содержащих искомый объект, cls_loss — точность предсказания класса объекта, dfl_loss используется для повышения точ-

ности положения рамки для небольших объектов и предсказывает распределение вероятностей для положения границ рамки.

Все программы реализованы на языке Python версии 3.12.11 с использованием стандартных библиотек, в скобках указана версия: torch (2.5.1), torchvision (0.20.1), torchaudio (2.5.1), ultralytics (8.3.152), numpy (2.3.5), opencv-python (4.13.0), cv2 (4.13.0), pandas (2.3.3), matplotlib (3.10.7), PIL (12.2.0), yaml (6.0.2), scipy (1.16.2), torch CUDA (12.4), cuDNN (90100). Ссылку на репозиторий проекта авторы могут предоставить по запросу на электронную почту.

3. Результаты и обсуждение

Было обучено 5 различных модификаций архитектуры YOLO. На рис. 4 показаны кривые и метрики обучения. Сильные осцилляции этих кривых свидетельствуют о нестабильности обучения и необходимости изменения архитектуры либо необходимости увеличения размеров обучающей выборки. Модели обучались со следующими гиперпараметрами.

Оптимизатор AdamW:

- lr0 = 0,001 — начальное значение learning rate;
- lrf = 0,05 — коэффициент конечного learning rate относительно начального;
- momentum = 0,8 — параметр momentum оптимизатора;
- weight_decay = 0,00005 — коэффициент l2-регуляризации;
- warmup_epochs = 20 — количество эпох warmup;
- warmup_momentum = 0,6 – значение momentum на этапе warmup;
- использовался линейный learning rate scheduler;
- lr изменялся от 0,001 до 0,00005.

Использовались функции потерь box loss, classification loss и distribution focal loss с параметрами:

- box = 7,5 — вес компоненты box loss;
- cls = 0,5 – вес компоненты classification loss;
- dfl = 1,5 — вес компоненты dfl.

В модели 2 было проведено упрощение головного блока, отвечающего за классификацию, путем удаления трех слоев Upsample-Concat-C3k2. Согласно рис. 5 графики функции потерь вышли на плато, что свидетельствует о завершении обучения модели.

В модели 3 также проведен эксперимент по увеличению размеров головного блока путем добавления трех блоков Upsample-Concat-C3k2 и слоев MFAM. Как показано на рис. 6, модель успешно обучилась, так как функция потерь вышла на плато.

В модели 4 «основа» была расширена за счет блоков DASI, при этом головной модуль уменьшен до одного блока Concat-C3k2-Upsample-C3k2. Кривые обучения при этом больше осциллируют, чем соответствующие кривые для модели 3, но меньше, чем для модели 1 (рис. 7). Заметим, что модели 3 и 4, где есть блоки MFAM, хоть и имеют мало параметров, но являются вычислительно тяжелыми.

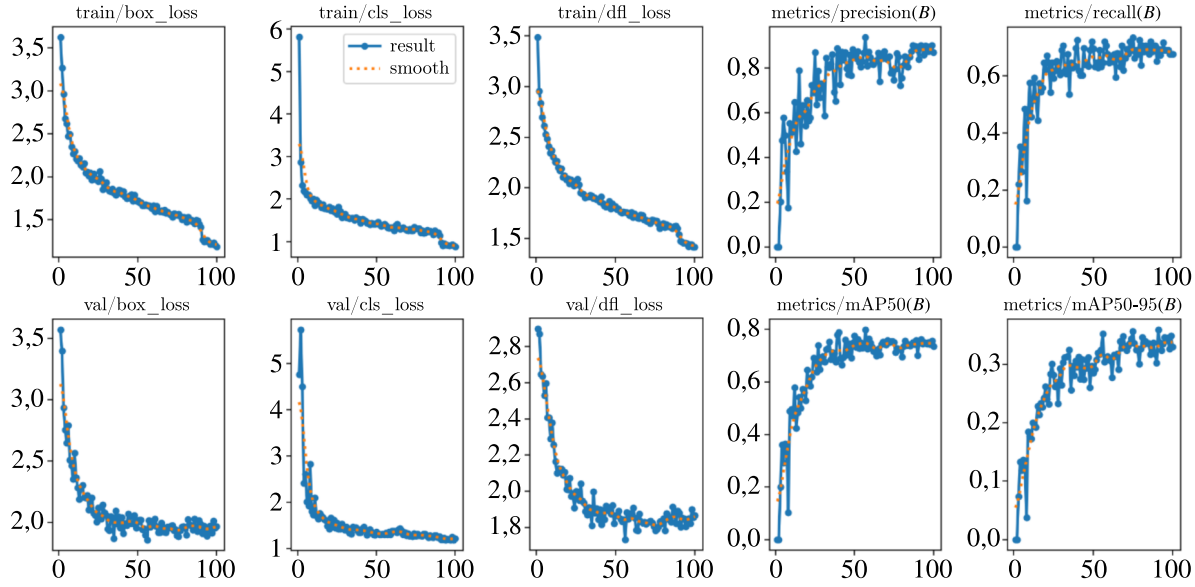

Рис. 4. Графики кривых потери и метрик точности для модели 1

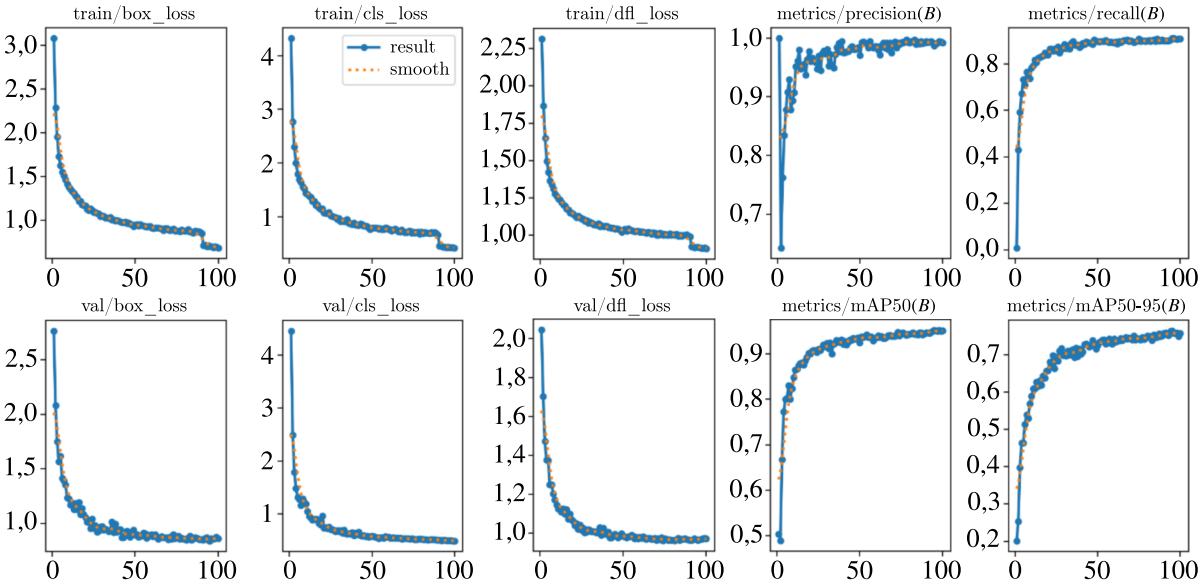

Рис. 5. Графики кривых потери и метрик точности для модели 2

В модели 5 были интегрированы блоки DASI и RAT. Это позволило улучшить стабильность обучения (на рис. 7 кривая потерь более гладкая).

Проведена серия экспериментов по оптимизации архитектур YOLO (5 модификаций) и представлены их оценки качества (таблица 1). Основными критериями оценки были стабильность процесса обучения (характер кривых потерь), точность и вычислительная эффективность.

Модель 1 (базовый вариант) выступила в качестве референсной модели. Наличие сильных осцилляций на кривых обучения указывает на нестабильность обучения. Это сигнализировало о необходимости либо изменения архитектуры, либо увеличения объема данных.

Модель 2 (упрощенная) была получена путем редуцирования части слоев в головном блоке. Этот шаг позволил добиться стабилизации — функция потерь вышла на плато. Это указывает

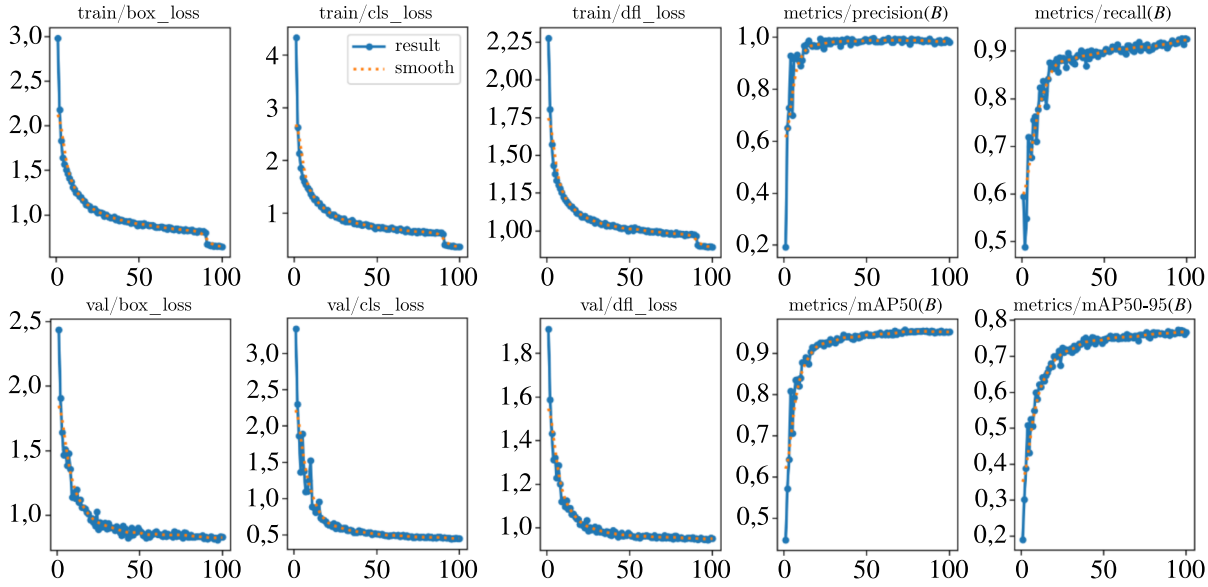

Рис. 6. Графики кривых потери и метрик точности для модели 3

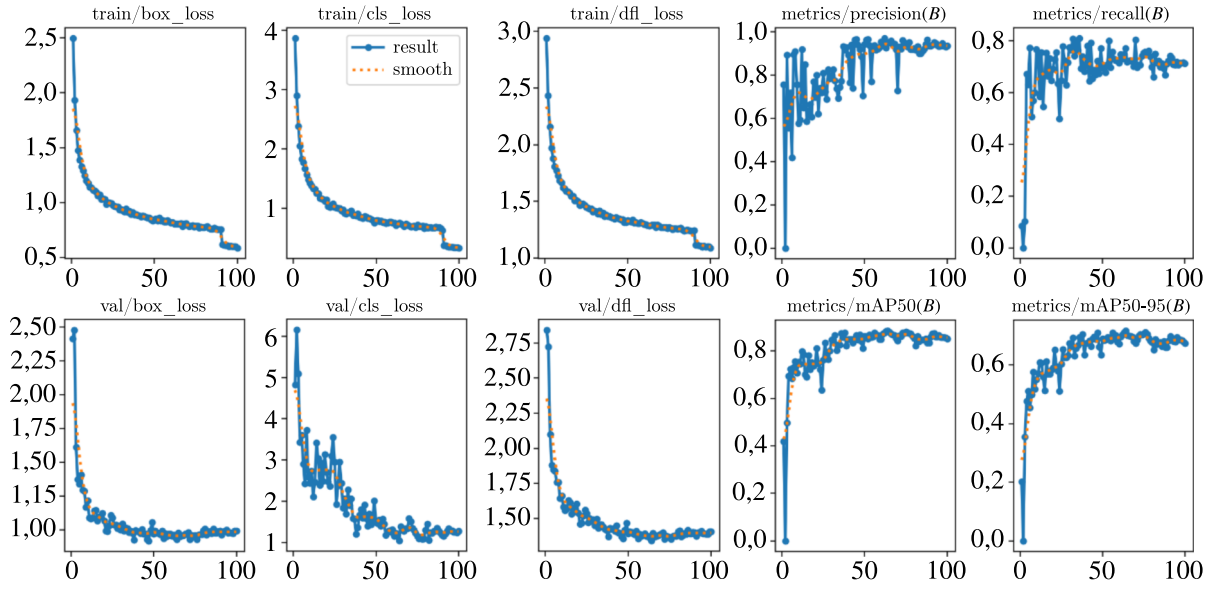

Рис. 7. Графики кривых потери и метрик точности для модели 4

на завершение обучения и потенциальное снижение риска переобучения, но может говорить об упрощении модели и, возможно, ограничении ее способности к обучению сложным признакам.

Для модели 3 (усложненная с блоком MFAM) был реализован противоположный подход — расширение головного блока и добавление специализированных блоков MFAM. Обучение также успешно завершилось (выход на плато). Это демонстрирует, что более сложная архитектура может быть стабильно обучена, но данная архитектура обладает более высокой вычислительной сложностью по сравнению с моделями 1 и 2 при малом числе параметров.

В модели 4 (гибридная с DASI) были внесены изменения как в backbone (расширение блоками DASI), так и головы (ее упрощение). Результат: осцилляции меньше, чем у модели 1, но больше, чем у модели 3. Это показывает, что изменения в backbone не компенсировали упроще-

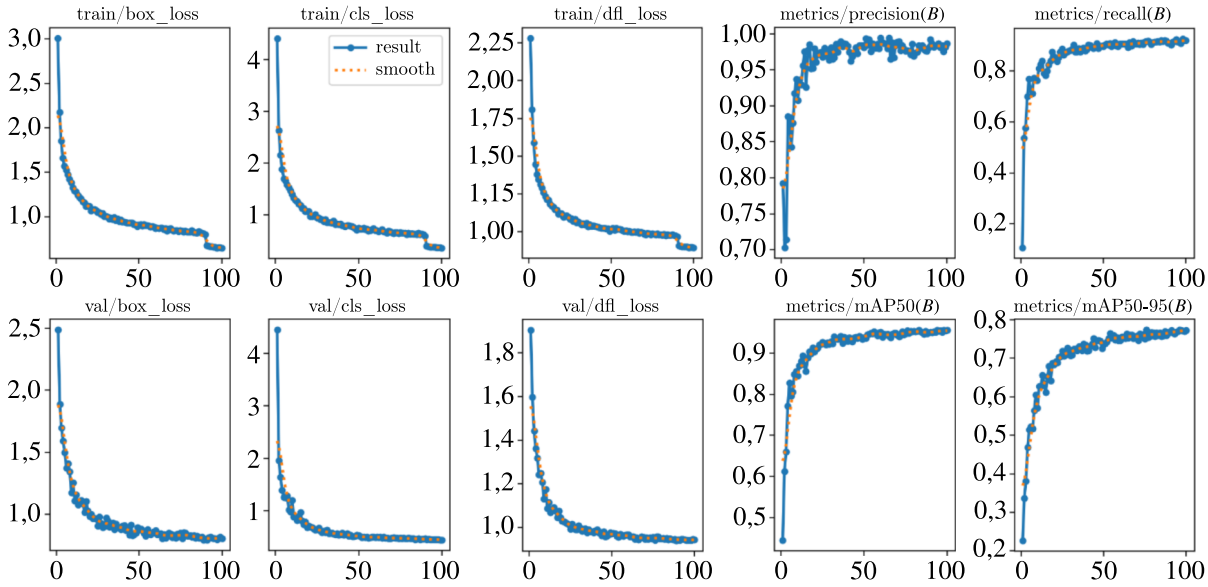

Рис. 8. Графики кривых потери и метрик точности для модели 5

Таблица 1. Сводная таблица сравнений различных протестированных предиктивных моделей

Модель	Изменения архитектуры	Стабильность обучения	Ключевой вывод
Модель 1	Базовая	Низкая (сильные осцилляции)	Нестабильность требует изменений
Модель 2	Упрощение головы	Высокая (выход на плато)	Упрощение стабилизирует обучение
Модель 3	Расширение головы + MFAM	Высокая (выход на плато)	Сложные блоки обучаемы, но тяжелы для вычислений
Модель 4	Расширенный backbone (DASI) + + упрощенная голова	Средняя (осцилляции умеренные)	Изменения в backbone не дали максимальной стабильности
Модель 5	DASI + RAT	Наилучшая (гладкая кривая)	Сбалансированная модификация дает лучшую стабильность

ние головной части для достижения максимальной стабильности. Также подтверждается высокая вычислительная стоимость моделей с MFAM.

В модели 5 (оптимизированная гибридная) были интегрированы блоки DASI и RAT, что привело к наилучшей стабильности обучения среди всех вариантов (самая гладкая кривая потерь). Этот результат демонстрирует успешную балансировку архитектуры.

Таким образом, эксперименты показали, что излишняя сложность (модель 1) и излишнее упрощение (модель 2) могут быть субоптимальны. Наилучшие результаты по стабильности обучения (таблица 2) были достигнуты за счет целенаправленного усложнения и модификации специализированных блоков (DASI, MFAM) в сбалансированной архитектуре (модели 3 и 5).

Время инференса для предложенных моделей 1, 2, 5 не превысило 7,3 мс, для YOLO — 14,7 мс, что говорит об ускорении работы не менее чем в 2 раза при незначительной потере точности на используемой выборке и вышеуказанном вычислительном устройстве.

Таблица 2. Оценки качества различных модификаций модели YOLO v11

Модель	mAP50	mAP50-95	Количество параметров	Вычислительная сложность GFLOPs
Модель 1	0,954	0,777	1 138 139	6,4
Модель 2	0,952	0,768	598 538	4,3
Модель 3	0,718	0,537	1 467 388	19,7
Модель 4	0,886	0,716	1 134 683	20,1
Модель 5	0,95	0,734	2 644 034	7,3
YOLO v11 без модификаций	0,976	0,841	20 030 803	67,6

4. Заключение

В данной работе приведены оценки пяти модификаций архитектуры YOLO v11 для задачи обнаружения малоразмерных объектов на аэрофотоснимках в видимом спектре на основе синтетической базы изображений с различными аугментациями. Было показано, что использование блока DASI в совокупности с RAT (в модели 5) позволило достичь наилучшей стабильности обучения (самая гладкая кривая потерь) по сравнению с исходной YOLOv11 и другими вариантами. Экспериментально показано, что предложенная архитектура (модель 5) обеспечивает двукратное сокращение времени инференса (с 14,7 мс до 7,3 мс) при незначительной потере точности (mAP50 снизилась с 0,976 до 0,95, mAP50-95 — с 0,841 до 0,734) на синтетическом массиве данных с муляжами PFM-1. Следует отметить, что данные метрики соответствуют следующим статистическим характеристикам для моделей (YOLOv11 базовой)/(модель 5): количество истинно положительных детекций — 378/374, ложноположительных — 10/5, ложноотрицательных — 15/19. Таким образом, было потеряно около 1 % детекций при существенном ускорении инференса.

При этом количество параметров модели снижено более чем на порядок. Таким образом, впервые показана возможность кардинального ускорения без развертывания облегченных версий (n/s/m/l/x), а именно за счет вставки MFAM и DASI + RAT в стандартную YOLOv11. Впервые выполнена систематическая оценка влияния MFAM и DASI + RAT на стабильность обучения, точность детекции и скорость инференса для малоразмерных объектов. Было проведено сравнение 5 моделей (базовая, упрощенная голова, голова + MFAM, backbone + DASI + упрощенная голова, DASI + RAT) и показано, что излишняя сложность или излишнее упрощение субоптимальны, а наилучший баланс достигается при использовании DASI + RAT. Продемонстрирована применимость такого подхода для задач гуманитарного разминирования на открытом синтетическом массиве данных изображений мин типа «лепесток» с расширенным набором аугментаций (mosaic, mixup, copy-paste, erasing и др.). Ранее аналогичные работы [Vivoli et al., 2024] использовали другие архитектуры и не фокусировались на вычислительной эффективности для развертывания на мобильных платформах.

Несмотря на высокие показатели точности обнаружения объектов, данное исследование носит фундаментальный характер и не учитывает заслонения, вариации положения в пространстве, отличные от горизонтальных, а также помехи, возникающие в реальных условиях. Полученные предиктивные модели позволили уменьшить время обработки изображений минимум в 2 раза при сохранении точности распознавания.

Список литературы (References)

Aldubaikhi A., Patel S. Advancements in small-object detection (2023–2025): Approaches, datasets, benchmarks, applications, and practical guidance // Applied Sciences. — 2025. — Vol. 15, No. 22. — P. 11882.

Aziz L., Salam M. S. B. H., Sheikh U. U., Ayub S. Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: a comprehensive review // IEEE Access. — 2020. — Vol. 8. — P. 170461–170495. — DOI: 10.1109/ACCESS.2020.3021508

Benali Amjoud A., Amrouch M. Convolutional neural networks backbones for object detection // International Conference on Image and Signal Processing. — Cham: Springer, 2020. — P. 282–289. — DOI: 10.1007/978-3-030-51935-3_30

Chen W., Luo J., Zhang F., Tian Z. A review of object detection: Datasets, performance evaluation, architecture, applications and current trends // Multimedia Tools and Applications. — 2024. — Vol. 83, No. 24. — P. 65603–65661. — DOI: 10.1007/s11042-023-17949-4

Du D., Qi Y., Yu H., Yang Y., Duan K., Li G., Zhang W., Huang Q., Tian Q. The unmanned aerial vehicle benchmark: Object detection and tracking // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 370–386.

Duan K., Bai S., Xie L., Qi H., Huang Q., Tian Q. Centernet: Keypoint triplets for object detection // Proceedings of the IEEE/CVF international conference on computer vision. — 2019. — P. 6569–6578.

Duan K., Bai S., Xie L., Qi H., Huang Q., Tian Q. CenterNet++ for object detection // IEEE transactions on pattern analysis and machine intelligence. — 2023. — Vol. 46, No. 5. — P. 3509–3521.

Evci U., Dumoulin V., Larochelle H., Mozer M. C. Head2toe: utilizing intermediate representations for better transfer learning // Proceedings of the International Conference on Machine Learning (ICML). — PMLR, 2022. — P. 6009–6033. — DOI: 10.48550/arXiv.2201.03529

Juhartini, Arwidiyarti D., Desmiwati. Single shot multibox detector (SSD) in object detection: a review // IJACI: International Journal of Advanced Computing and Informatics. — 2025. — Vol. 1, No. 2. — P. 118–127. — DOI: https://doi.org/10.71129/ijaci.v1i2.pp118-127

Khanam R., Hussain M. YOLOv11: an overview of the key architectural enhancements // arXiv preprint. — 2024. — DOI: 10.48550/arXiv.2410.17725

Kumar J. S., Anuar S., Hassan N. H. Transfer learning based performance comparison of the pre-trained deep neural networks // International Journal of Advanced Computer Science and Applications. — 2022. — Vol. 13, No. 1. — P. 1–10. — DOI: 10.14569/IJACSA.2022.0130193

Lin T.-Y., Goyal P., Girshick R., He K., Dollár P. Focal loss for dense object detection // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 2980–2988.

Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C. Y., Berg A. C. SSD: single shot multibox detector // European Conference on Computer Vision (ECCV 2016). — Cham: Springer International Publishing, 2016. — P. 21–37.

Lu L., He D., Liu C., Deng Z. MASF-YOLO: an improved YOLOv11 network for small object detection on drone view // arXiv preprint. — 2025. — https://arxiv.org/abs/2504.18136

Mirzaei B., Nezamabadi-pour H., Raoof A., Derakhshani R. Small object detection and tracking: a comprehensive review // Sensors. — 2023. — Vol. 23, No. 15. — P. 6887.

Muzammul M., Li X. Comprehensive review of deep learning-based tiny object detection: challenges, strategies, and future directions // Knowledge and Information Systems. — 2025. — P. 1–89. – DOI: 10.1007/s10115-025-02375-9

Synthetic PFM-1 dataset computer vision model. — [Electronic resource]. — https://universe.roboflow.com/synthetic-data/synthetic-pfm-1-dataset (accessed: 19.02.2026).

Tan M., Pang R., Le Q. V. Efficientdet: Scalable and efficient object detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — P. 10781–10790.

Vivoli E., Bertini M., Capineri L. Deep learning-based real-time detection of surface landmines using optical imaging // Remote sensing. — 2024. — Vol. 16, No. 4. — P. 677. — DOI: 10.3390/rs16040677

Wang C. Y., Bochkovskiy A., Liao H. Y. M. Scaled-yolov4: Scaling cross stage partial network // Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. — 2021. — P. 13029–13038.

Wei W., Huang Y., Zheng J., Rao Y., Wei Y., Tan X., OuYang H. YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images // Journal of Radiation Research and Applied Sciences. — 2025. — Vol. 18, No. 1. — P. 101309.

Xia G.-S., Bai X., Ding J., Zhu Z., Belongie S., Luo J., Datcu M., Pelillo M., Zhang L. DOTA: A large-scale dataset for object detection in aerial images // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 3974–3983.

Yang Z., Chen H., Qian Z., Zhou Y., Zhang H., Zhao D., Wei B., Xu Y. Region attention transformer for medical image restoration // International Conference on Medical Image Computing and Computer-Assisted Intervention. — Cham: Springer Nature Switzerland, 2024. — P. 603–613.

Zhang Y., Zhang K., Wang J., Wu Y., Wang W. Multi-level feature aggregation and recursive alignment network for real-time semantic segmentation // arXiv preprint. — 2024. — https://arxiv.org/html/2402.02286v1

Zhu P., Wen L., Du D., Bian X., Fan H., Hu Q., Ling H. Detection and tracking meet drones challenge // IEEE transactions on pattern analysis and machine intelligence. — 2021. — Vol. 44, No. 11. — P. 7380–7399.