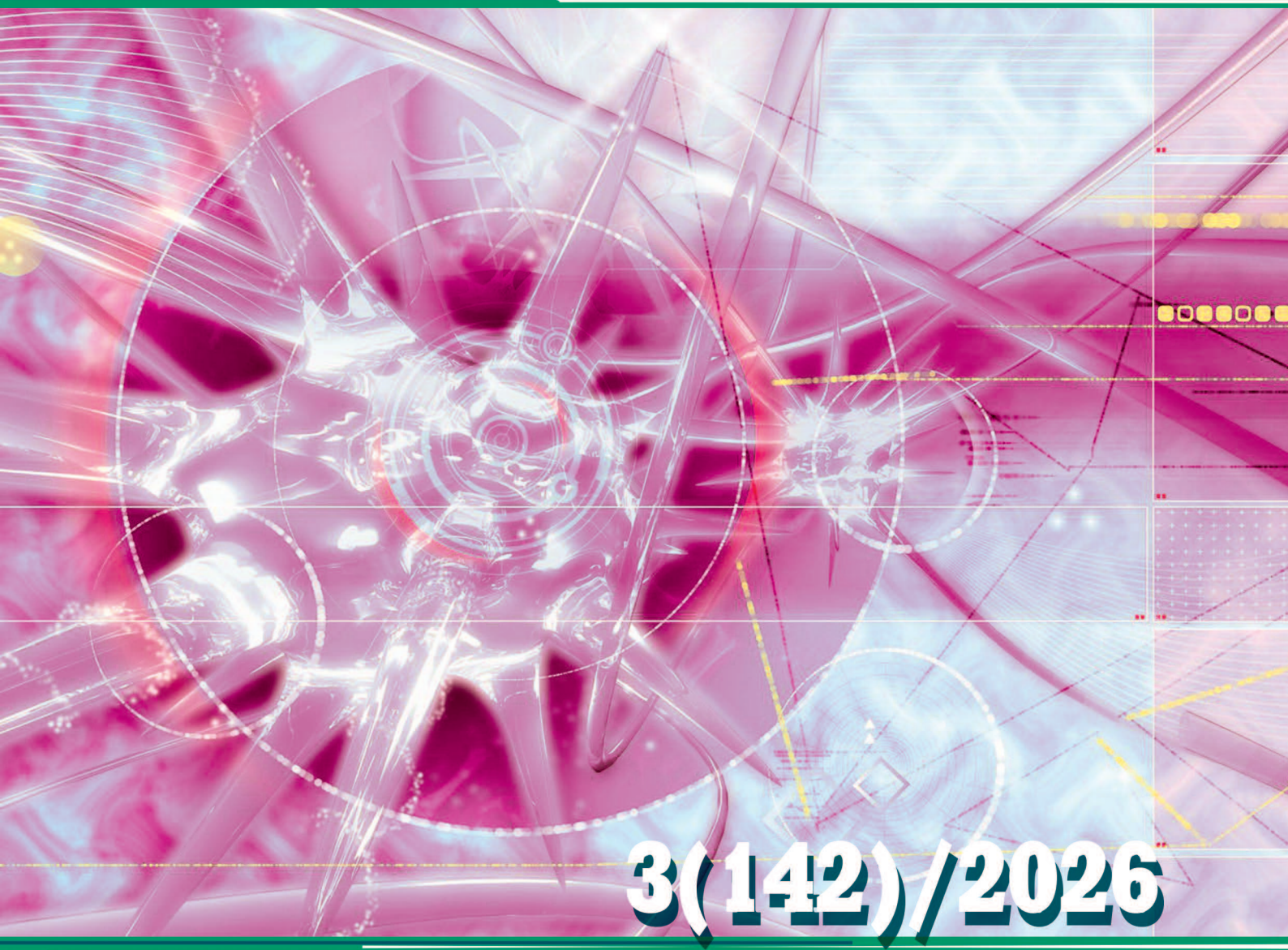
ISSN 1684-8853 (print); ISSN 2541-8610 (online)

ИНФОРМАЦИОННО-УПРАВЛЯЮЩИЕ СИСТЕМЫ

НАУЧНЫЙ ЖУРНАЛ



3(142)/2026

doi:10.31799/1684-8853-2026-3

ISSN 1684-8853 (print); ISSN 2541-8610 (online)

SCIENTIFIC JOURNAL

3(142)/2026

PEER REVIEWED JOURNAL

INFORMATSIONNO-UPRAVLIAIUSHCHIE SISTEMY (INFORMATION AND CONTROL SYSTEMS)

Contact information
The Editorial and Publishing Center, SUAI
67A, Bol'shaya Morskaia, 190000, Saint Petersburg, Russia
Website: http://i-us.ru/en, e-mail: ius.spb@gmail.com
Tel.: +7 - 812 494 70 02

The Journal was registered in the Ministry of Press, Broadcasting and Mass Media of the Russian Federation.
Registration Certificate JD № 77-12412 from April, 19, 2002.
Re-registration in the Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications (ROSKOMNADZOR) due to change of the founder A. Vostrikov,
PI № FS 77-82226 from November, 23, 2021

Submitted for publication 11.05.26. Passed for printing 26.06.26. Signed for printing: 30.06.26.
Format 60×84$_{1/8}$. Phototype CentSchbkCyrill BT. Digital printing.

Layout original is made at the Editorial and Publishing Center, SUAI.
67A, Bol'shaya Morskaia, 190000, Saint Petersburg, Russia
Printed from slides at the Editorial and Publishing Center, SUAI.
67A, Bol'shaya Morskaia, 190000, Saint Petersburg, Russia

The journal is indexed in Scopus.

Free distribution.

doi:10.31799/1684-8853-2026-3 ISSN 1684-8853 (print); ISSN 2541-8610 (online)
НАУЧНЫЙ ЖУРНАЛ

3(142)/2026

РЕЦЕНЗИРУЕМОЕ ИЗДАНИЕ

ИНФОРМАЦИОННО-УПРАВЛЯЮЩИЕ СИСТЕМЫ

Адрес редакции: 190000, г. Санкт-Петербург,
ул. Большая Морская, д. 67, лит. А, ГУАП, РИЦ
Тел.: (812) 494-70-02, эл. адрес: ius.spb@gmail.com,
сайт: http://i-us.ru

Журнал входит в БД Scopus и в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук.

Сдано в набор 11.05.26. Подписано в печать 26.06.26. Дата выхода в свет: 30.06.26.
Формат 60×841/8. Гарнитура CentSchbkCyrill BT. Печать цифровая.
Усл. печ. л. 9,2. Уч.-изд. л. 12,9. Тираж 1000 экз (1-й завод 50 экз.). Заказ № 215.
Оригинал-макет изготовлен в редакционно-издательском центре ГУАП.
190000, г. Санкт-Петербург, ул. Большая Морская, д. 67, лит. А.
Отпечатано в редакционно-издательском центре ГУАП.
190000, г. Санкт-Петербург, ул. Большая Морская, д. 67, лит. А.
Распространяется бесплатно.

Журнал зарегистрирован в Министерстве РФ по делам печати, телерадиовещания и средств массовых коммуникаций.
Свидетельство о регистрации ПИ № 77-12412 от 19 апреля 2002 г.
Перерегистрирован в Роскомнадзоре.
Свидетельство о регистрации ПИ № ФС77-82226 от 23 ноября 2021 г.

УДК 004.08
doi:10.31799/1684-8853-2026-3-2-13
EDN: WKWZGY

Научные статьи
Articles

Метод сжатия глубоких нейронных сетей, основанный на геометрически контролируемом прореживании

Т. М. Татарникова[a], *доктор техн. наук, профессор, orcid.org/0000-0002-6419-0072, tm-tatarn@yandex.ru*
А. С. Раскопина[a], *аспирант, ассистент, orcid.org/ 0009-0002-0276-607X*
[a]*Санкт-Петербургский государственный университет аэрокосмического приборостроения, Б. Морская ул., 67, Санкт-Петербург, 190000, РФ*

Введение: *высокие вычислительные ресурсы, энергозатраты и время для решения задач с применением технологий глубокого обучения обусловили поиск решений сжатия моделей нейронных сетей без существенной потери качества результата.* ***Цель:*** *разработать метод сжатия глубоких нейронных сетей – сокращения вычислительной сложности и числа параметров сверточных нейронных сетей без существенной потери точности решения задачи классификации.* ***Результаты:*** *разработан новый метод геометрически контролируемого прореживания модели нейронной сети, основанный на жадном отборе кандидатов структурного прореживания с контролем изменения геометрии представлений. Предложена метрика контролирования сохранения геометрии представлений в виде матрицы межклассовых сходств, вычисляемой по центроидам классов в пространстве признаков. Введен параметр допустимого бюджета деформации геометрии представлений и предложен подход к его выбору на основе оценки шумового порога геометрической метрики. Результаты эксперимента показали, что предложенный метод сжатия моделей нейронных сетей обеспечивает сохранение точности классификации после дообучения, сопоставимой с базовой моделью без прореживания при сокращении вычислительной сложности на ~8 % и числа параметров на ~12 % на примере архитектуры ResNet-50 и набора данных CIFAR-100. Дополнительно показана переносимость разработанного метода на архитектуры глубоких нейронных сетей ResNet-18 и MobileNetV2.* ***Практическая значимость:*** *разработанный метод может найти применение при решении задачи классификации на мобильных и встраиваемых устройствах в реальном времени.*

Ключевые слова *– глубокое обучение, сжатие нейронной сети, прореживание, геометрия представлений, вычислительная сложность, задержка инференса, качество классификации.*

Для цитирования: Татарникова Т. М., Раскопина А. С. Метод сжатия глубоких нейронных сетей, основанный на геометрически контролируемом прореживани. *Информационно-управляющие системы*, 2026, № 3, с. 2–13. doi:10.31799/1684-8853-2026-3-2-13, EDN: WKWZGY
For citation: Tatarnikova T. M., Raskopina A. S. A deep neural network compression method based on geometrically controlled thinning. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 2–13 (In Russian). doi:10.31799/1684-8853-2026-3-2-13, EDN: WKWZGY

Введение

Сжатие нейронных сетей является основным решением при практическом применении глубокого обучения в вычислительно ограниченных условиях. Например, современные сверточные нейронные сети обеспечивают высокое качество распознавания изображений, однако их вычислительная сложность и объем параметров приводят к росту задержки инференса, потребления памяти и энергозатрат. Это существенно ограничивает использование моделей глубокого обучения в задачах реального времени, а также на мобильных и встраиваемых устройствах [1–4].

Одним из наиболее распространенных подходов к сжатию нейронных сетей является прореживание — удаление части параметров или структурных элементов модели нейронной сети с целью уменьшить ее вычислительную сложность. В зависимости от вида удаляемых элементов различают неструктурированное прореживание, предполагающее обнуление отдельных весов связей между нейронами; структурированное прореживание, при котором удаляются каналы сверточной нейронной сети, фильтры или блоки; полуструктурированное прореживание, которое выполняется в соответствии с регулярным шаблоном разреженности. Эти традиционные подходы хоть и позволяют уменьшить вычислительную сложность, приводящую к желаемому ускорению вывода решения, но сопровождаются существенным ухудшением качества решения задачи классификации [5–7].

Главная проблема традиционных методов прореживания — это выбор элементов сети, удаление которых минимально влияет на качество результата. Традиционные методы опираются на локальные критерии важности параметров, такие как нормы весов, или на приближенные оценки вклада каналов в итоговую ошибку. Однако качество классификации определяется не только значениями весов, но и формированием внутреннего пространства признаков, в котором классы становятся разделимыми. Поэтому

перспективно построение методов прореживания, контролирующих изменение структуры сжимаемой модели нейронной сети [8–10].

В данной работе предлагается метод геометрически контролируемого прореживания, основанный на жадной стратегии выбора степени структурного прореживания блоков сети при ограничении на изменение геометрии представлений. Геометрия представлений определяется через матрицу попарных косинусных сходств между центроидами классов в пространстве признаков. Поскольку оценка изменения геометрии вычисляется на основе выборочных пакетов данных и подвержена стохастической вариативности, вводится понятие порогового уровня шума — величины естественных флуктуаций метрики ΔG при неизменной архитектуре. Данный порог используется для формирования устойчивого бюджета допустимого изменения геометрии.

Постановка задачи и метрики оценивания

Рассматривается задача многоклассовой классификации изображений. Пусть задана обучающая выборка

$$D_{Learn} = (x_i, y_i)_{(i=1)}^{N},$$

где $x_i \in R^{(H\times W\times C)}$ — входное изображение, H, W, C — высота, ширина и число каналов (фильтров) изображения; $y_i \in \{1, \ldots, K\}$ — метка класса, K — число классов; N — количество обучающих примеров.

Пусть нейронная сеть с параметрами θ задает отображение

$$\mathbf{f}_\theta : R^{(H\times W\times C)} \to R^K,$$

где $\mathbf{f}_\theta(x)$ — вектор логитов, т. е. ненормированных оценок принадлежности к каждому из K классов.

Предсказанный класс определяется как

$$\hat{y}(x) = \arg_{k\in\{1,\ldots,K\}} \max\left[\mathbf{f}_\theta(x)\right].$$

Цель прореживания — сжатие модели $\mathbf{f}_{\theta'}$, где θ' получается из θ путем удаления части параметров или структурных элементов нейронной сети с выполнением следующих требований:

— сохранение качества классификации на уровне исходной модели нейронной сети;

— уменьшение числа параметров модели нейронной сети и ее вычислительной сложности;

— улучшение времени получения результата (задержки инференса).

В работе используются стандартные метрики Top-1 и Top-5 точности. Top-1 точность определяется как доля примеров, для которых предсказанный класс совпадает с истинным. Top-5 точность определяется как доля примеров, для которых истинный класс входит в множество пяти наиболее вероятных предсказаний [10, 11].

Для оценки степени сжатия модели нейронной сети используются следующие показатели:

1) число параметров модели, P;

2) объем модели в мегабайтах, V — суммарный размер параметров с учетом формата хранения;

3) вычислительная сложность, FLOPs — число операций умножения-сложения при одном прямом проходе модели [12];

4) задержка инференса в миллисекундах, T — продолжительность времени между подачей входных данных в модель и получением результата. На практике задержка инференса зависит не только от числа FLOPs, но и от особенностей аппаратной платформы, реализации операторов и структуры вычислительного графа. Поэтому задержка инференса рассматривается как эмпирическая метрика, отражающая реальное ускорение в условиях выбранного вычислительного окружения.

Для анализа реализуемых методов, кроме общепринятых метрик, введем понятие геометрии представлений для контроля устойчивости внутреннего пространства признаков.

Пусть $\boldsymbol{\varphi}_\theta(x) \in R^d$ — вектор признаков, извлекаемый из нейронной сети на предпоследнем уровне. Выбор предпоследнего слоя позволяет анализировать геометрию признаков независимо от параметров классификатора и обеспечивает сопоставимость межклассовой структуры до и после структурного упрощения сети. Тогда для каждого класса $k \in \{1, \ldots, K\}$ определим множество объектов данного класса

$$D_k = \left\{(x, y) \in D_{Learn} : y = k\right\}.$$

Тогда множество признаков класса k в пространстве признаков определяется как

$$H_k = \left\{\boldsymbol{\varphi}_\theta(x_i) \mid y_i = k\right\}.$$

Центроид класса определяется как [12]

$$\mu_k = \frac{1}{|H_k|} \sum_{\boldsymbol{\varphi}_\theta(x_i)\in H_k} \boldsymbol{\varphi}_\theta(x_i).$$

Далее строится матрица попарных косинусных сходств между центроидами классов

$$\mathbf{S} = \left[s_{ij}\right]_{i,j=1}^{K}, \quad s_{ij} = \frac{\left\langle \mu_i, \mu_j \right\rangle}{\|\mu_i\|_2 \cdot \|\mu_j\|_2 + \delta'}, \qquad (*)$$

где $\delta > 0$ — малая константа, предотвращающая деление на ноль.

Матрица $\mathbf{S}$ отражает относительное расположение классов в пространстве представлений:

близкие классы имеют высокое косинусное сходство, плохо разделимые классы — низкое. Матрица $\mathbf{S}$ характеризует угловую структуру пространства признаков и описывает взаимное направление центроидов классов.

Пусть $\mathbf{S}_{Ref}$ — матрица сходств, вычисленная в пространстве признаков $\varphi_\theta(x)$ для базовой модели $\mathbf{f}_\theta$, а $\mathbf{S}$ — матрица сходств, вычисленная в пространстве признаков $\varphi_{\theta'}(x)$ для модифицированной модели $\mathbf{f}_{\theta'}$ (после прореживания). Тогда изменение геометрии представлений определяется величиной

$$\Delta G(\theta',\ \theta) = \frac{\left\|\mathbf{S} - \mathbf{S}_{Ref}\right\|_F}{\left\|\mathbf{S}_{Ref}\right\|_F + \delta},$$

где $||\cdot||_F$ — норма Фробениуса.

В работе используется косинусная мера сходства между центроидами классов. Для повышения корректности сравнения признаки перед вычислением сходств L2-нормируются. Изменение геометрии ΔG измеряется как относительная фробениусова норма разности матриц сходств.

Выбранная метрика обладает масштабной инвариантностью и частичной инвариантностью к ортогональным преобразованиям, однако не инвариантна к произвольным аффинным преобразованиям, что учитывается при интерпретации результатов.

Величина ΔG используется для оценки изменения межклассовой структуры признакового пространства после прореживания: малые значения соответствуют сохранению относительного расположения классов и рассматриваются как индикатор устойчивости представлений. При этом ΔG не гарантирует сохранение точности и выступает как критерий, ограничивающий чрезмерную деформацию.

Метрика основана на косинусных сходствах центроидов классов и отражает межклассовую структуру через первые моменты распределений. Она не учитывает внутриклассовую дисперсию, однако позволяет контролировать изменения линейной разделимости за счет сохранения взаимного расположения классов.

Традиционные методы прореживания нейронных сетей

В рамках работы анализируются три наиболее распространенные группы методов: неструктурированное, структурированное и полуструктурированное прореживание [13—15]. Следует отметить, что в современной литературе используются и другие классификации методов прореживания, основанные на критерии отбора параметров (magnitude-based, gradient-based, Hessian-based), стратегии оптимизации (one-shot, iterative) или режиме обучения (post-training, training-time pruning). В настоящей работе используется структурная типология, поскольку она непосредственно связана с изменением архитектуры сети и вычислительной сложности, что соответствует целям исследования.

Неструктурированное прореживание заключается в обнулении отдельных весов нейронной сети без изменения размерностей слоев [16]. Пусть $\mathbf{W}$ — тензор весов некоторого слоя. Тогда прореживание задается бинарной маской $\mathbf{M}$ той же размерности:

$$\mathbf{W}' = \mathbf{W} \times \mathbf{M}.$$

При заданной доле разреженности удаляются веса с наименьшими значениями модуля веса:

$$M_{ij} = \begin{cases} 0, \text{ если } \left|W_{ij}\right| \le \tau; \\ 1, \text{ если } \left|W_{ij}\right| > \tau, \end{cases}$$

где τ — пороговое значение.

Неструктурированное прореживание позволяет достигать высокой разреженности при небольшой потере точности после дообучения. Однако его практическое ускорение на стандартных платформах (GPU/CPU) ограничено из-за нерегулярной структуры и необходимости специализированных операций с разреженными матрицами [17, 18].

Структурированное прореживание предполагает удаление целых структурных единиц модели: каналов свертки, фильтров, нейронов, блоков или слоев. В отличие от неструктурированного подхода, структурное прореживание изменяет размерности весовых тензоров и активаций, что приводит к непосредственному уменьшению вычислительной сложности и ускорению инференса [19].

Для сверточного слоя с весами $\mathbf{W} \in \mathbf{R}^{C_{Out} \times C_{In} \times k \times k}$ структурированное прореживание по выходным каналам соответствует удалению некоторого подмножества индексов $P \in \{1, \ldots, C_{Out}\}$ [20]. Тогда новый слой имеет размерность

$$\mathbf{W}' \in \mathbf{R}^{(C_{Out} - |P|) \times C_{In} \times k \times k},$$

где C_{Out} — число выходных каналов (фильтров); C_{In} — число входных каналов слоя; P — число удаленных каналов (каналы с минимальной нормой фильтра удаляются в первую очередь).

Основным преимуществом структурированного прореживания является прямое сокращение вычислительной сложности модели нейронной сети и ускорение ее работы на стандартных устройствах. Недостатком является сильная деградация качества, поскольку удаление целых каналов уменьшает выразительную способность модели [21, 22].

Полуструктурированное прореживание вводит регулярную схему разреженности, которая поддерживается современными аппаратными платформами. Наиболее распространенным вариантом является шаблон (n:m), при котором в каждой группе из m весов сохраняются только n ненулевых [23, 24].

Пусть $\mathbf{W} \in \mathbf{R}^m$ — группа весов. Тогда (n:m)-прореживание задает маску $\mathbf{M} \in \{0, 1\}^m$, удовлетворяющую условию $||\mathbf{M}||_0 = n$, а веса после прореживания определяются как $\mathbf{W}' = \mathbf{W} \times \mathbf{M}$. Обычно сохраняют n весов с максимальным модулем.

Полуструктурированное прореживание сочетает лучшую аппаратную эффективность по сравнению с неструктурированным и большую гибкость по сравнению со структурированным подходом. При этом фактический выигрыш в задержке инференса зависит от поддержки (n:m)-разреженности на конкретной платформе.

После прореживания качество модели может снижаться из-за изменения отображения и распределения представлений, поэтому применяется дообучение для восстановления точности.

Несмотря на широкое распространение, традиционные методы прореживания имеют ряд ограничений.

1. Критерии важности часто основаны на локальных характеристиках (нормы весов, статистики активаций, приближенные оценки градиентов), что не гарантирует сохранение разделимости классов.

2. При высокой степени прореживания наблюдается существенное снижение качества, не всегда компенсируемое дообучением.

3. Неструктурированная разреженность не обеспечивает уменьшение задержки инференса на стандартном оборудовании.

4. Отмечается неочевидный выбор частей модели нейронной сети для удаления при структурированном прореживании.

Описание предлагаемого метода сжатия глубоких нейронных сетей

Разработанный метод сжатия глубоких нейронных сетей основан на эвристическом предположении о том, что существенная деформация межклассовой структуры представлений при структурном упрощении сети повышает риск ухудшения классиционной способности модели. Соответственно, ограничение изменения геометрии рассматривается как практический критерий контроля риска деградации качества, а не как строгое теоретическое условие сохранения точности.

Таким образом, задача прореживания глубокой нейронной сети формулируется как поиск ее структурно упрощенной модели при ограничении на допустимое изменение геометрии представлений.

Пусть базовая модель имеет параметры θ, а прореженная модель — θ'. В соответствии с формулой (*) введем матрицу сходств центроидов классов

$$\mathbf{S}_{Ref} = \mathbf{S}(\theta), \ \mathbf{S} = \mathbf{S}(\theta').$$

Изменение геометрии представлений определяется величиной

$$\Delta G(\theta', \theta) = \frac{\|S(\theta') - S(\theta)\|_F}{\|\mathbf{S}(\theta)\|_F + \delta}.$$

В предлагаемом методе сжатия модели нейронной сети вводится ограничение вида

$$\Delta G(\theta', \theta) \le \varepsilon,$$

где ε — допустимый бюджет изменения геометрии представлений.

Практическая реализация ограничения на ΔG требует учета того факта, что вычисление геометрии представлений выполняется по ограниченному числу наборов данных и поэтому имеет стохастическую составляющую. Даже при фиксированных параметрах θ значения матрицы $\mathbf{S}$ и, следовательно, ΔG могут отличаться при вычислении на разных выборках данных. Для учета данного эффекта вводится величина порога шума

$$\Delta G_{Noise} = \Delta G(\theta, B_1, B_2),$$

где B_1, B_2 – две выборки наборов данных, на которых независимо вычисляются матрицы $\mathbf{S}_1$ и $\mathbf{S}_2$ для одной и той же модели.

Тогда

$$\Delta G_{Noise} = \frac{\|\mathbf{S}_1 - \mathbf{S}_2\|_F}{\|\mathbf{S}_1\|_F + \delta}.$$

Далее итоговый бюджет задается как сумма шумового порога и дополнительного допуска

$$\varepsilon = \Delta G_{Noise} + \varepsilon_{\lim},$$

где $\varepsilon_{\lim} \ge 0$ — настраиваемый параметр метода.

Данный подход обеспечивает устойчивость метода к стохастическим колебаниям метрики и позволяет интерпретировать $\varepsilon_{\lim}$ как управляемый запас изменения геометрии.

Адаптация метода к модели сверточной нейронной сети

Рассмотрим работу предложенного метода геометрически контролируемого прореживания модели нейронной сети на архитектуре ResNet с прореживанием по ширине блоков архитектуры — сокращением числа внутренних каналов в отдельных блоках ResNet. Блок содержит несколько сверточных слоев.

Для выбора каналов, подлежащих удалению, используем критерий, основанный на статистике активаций. Рассмотрим первый блок (обозначим его BL1) с тензором активаций $\mathbf{a} \in \mathbf{R}^{H\times W\times C}$. Для каждого канала $c \in \{1, \ldots, C\}$ вычисляется дисперсия по пространственным координатам и по наборам данных, собранным на калибровочном наборе:

$$\vartheta_c = Var(\mathbf{a}_c).$$

Интуитивно каналы с более высокой дисперсией в среднем несут больше информации и сильнее участвуют в формировании признаков, поэтому в первую очередь сохраняются каналы с максимальными значениями ϑ_c.

Для заданной доли удаления $r \in (0, 1)$ выбирается число сохраняемых каналов

$$C_{Saved} = \max\left(C_{\min}\left\lfloor C(1-r)\right\rfloor\right),$$

где $C_{\min}$ — минимально допустимое число каналов в блоке; $\lfloor\cdot\rfloor$ — операция округления до ближайшего целого.

Далее выбирается множество индексов сохраняемых каналов

$$\zeta = TopK\left(\{\vartheta_c\}_{c=1}^{C}, C_{Saved}\right).$$

Метод геометрически контролируемого прореживания использует жадную стратегию. Пусть задан дискретный набор кандидатов долей прореживания

$$\Re = \{r_1, r_2, \ldots, r_m\}, \quad 0 < r_1 < \ldots < r_m < 1.$$

На практике кандидаты упорядочиваются по убыванию, т. е. сначала проверяются более агрессивные значения.

Для каждого блока нейронной сети выполняется перебор $r \in \Re$ и выбирается максимально возможное r, при котором выполняется ограничение $\Delta G \le \varepsilon$. Если ни один кандидат не удовлетворяет ограничению, блок не прореживается. Таким образом, метод автоматически регулирует степень прореживания в зависимости от чувствительности блоков к нарушению геометрии представлений. Ограничение $\Delta G \le \varepsilon$ накладывается на этапе структурного упрощения и используется для отбора допустимых вариантов прореживания. Следует отметить, что после дообучения значение $\Delta G \le \varepsilon$ может изменяться и не обязано совпадать с заданным порогом. В предлагаемом подходе геометрический контроль рассматривается как механизм ограничения начальной деформации векторного пространства перед оптимизацией, а не как жесткое инвариантное условие финального решения.

Фактически предложенный метод геометрически контролируемого прореживания вводит дополнительный критерий качества: удаление параметров допускается лишь при сохранении структуры взаимного расположения классов в признаковом пространстве. Это снижает риск нарушения разделимости классов при структурном упрощении модели.

При этом учитывается стохастическая природа оценки геометрии: значение ΔG может варьироваться в зависимости от выборки. В связи с этим ограничение задается относительно порога шума ΔG_{Noise}, а параметр $\varepsilon_{\lim}$ интерпретируется как допустимый запас деформации признакового пространства.

Методы прореживания, учитывающие структуру признакового пространства, обычно основаны на анализе активаций или сходства признаков [25–27], а также на различных критериях важности, включая значимость фильтров и структурные свойства сети [13, 15]. Такие подходы, как правило, требуют хранения промежуточных представлений или вычисления дополнительных метрик, что увеличивает вычислительную сложность.

В отличие от этого предлагаемый метод использует агрегированное описание геометрии классов через центроиды и косинусное сходство, обеспечивая более эффективную вычислительно альтернативу. При этом задача прореживания формулируется как задача контролируемого изменения структуры модели при ограничении на геометрию представлений, а не как задача максимального сокращения числа параметров.

Отметим, что ряд современных методов требует индивидуальной настройки протоколов обучения, поэтому применение единой схемы дообучения без адаптации может приводить к некорректной оценке. В данной работе для обеспечения сопоставимости условий рассматриваются базовые классы методов, тогда как расширенное сравнение с индивидуальной настройкой оставлено на дальнейшие исследования.

Постановка эксперимента

Эксперименты проводились на наборе данных CIFAR-100, содержащем 60 000 цветных изображений размером 32 × 32, распределенных по 100 классам (50 000 – обучающая выборка, 10 000 – тестовая). Этот набор данных широко применяется для оценки методов сжатия и ускорения сверточных нейронных сетей. Принятая постановка эксперимента ограничена задачей классификации изображений малого разрешения. Несмотря на формальную инвариантность используемой геометрической метрики к масштабу признаков,

переносимость результатов на задачи с высоким разрешением не гарантируется, поскольку такие задачи сопровождаются изменением архитектурных решений, числа классов и сложности распределения данных.

В качестве базовой архитектуры определена ResNet-50. Поскольку исходная версия сети ориентирована на изображения более высокого разрешения, применялась стандартная адаптация под CIFAR-100: свертка 7×7 заменена на 3×3 с шагом 1, операция maxpool удалена, а выходной слой модифицирован для 100 классов. Это позволило сохранить пространственное разрешение признаков на ранних этапах обработки и должным образом адаптировать архитектуру к данным малого размера.

Для корректного сравнения всех методов используется единый протокол обучения при равном вычислительном бюджете. Протокол включает следующие этапы:

1) обучение базовой модели (baseline-40) в течение 40 эпох на обучающей выборке CIFAR-100 и сохранение ее в качестве контрольной точки;

2) для оценки эффекта дополнительного обучения продолжается оптимизация базовой модели еще на 20 эпохах без применения прореживания.

Для всех методов прореживания реализована единая последовательность: baseline-40 → прореживание → дообучение на 20 эпохах. Таким образом, итоговые модели сравниваются при одинаковом числе эпох оптимизации (60 эпох), что исключает влияние «дополнительного времени обучения» как источника роста точности.

В рамках эксперимента рассматриваются следующие методы: базовая модель без прореживания, неструктурированное прореживание, структурированное прореживание, полуструктурированное прореживание – (n:m)-прореживание, геометрически контролируемое прореживание.

В экспериментах использовались следующие параметры:

— доля разреженности $r = 0{,}7$;

— на каждом шаге обучения модель обрабатывает 128 примеров данных одновременно;

— для полуструктурированного прореживания применялся шаблон n:m в виде 2:4;

— прореживанию подвергались блоки 3 и 4 архитектуры ResNet.

Параметры предложенного алгоритма, включая набор коэффициентов прореживания и минимальное число каналов, выбирались с целью обеспечить устойчивость процедуры и предотвратить деградацию представлений. Ограничение на минимальное число каналов позволяет избежать вырождения признакового пространства. Прореживание применялось к более глубоким слоям сети, формирующим высокоуровневые признаки и содержащим основную долю параметров, тогда как ранние слои сохранялись без изменений для поддержания стабильности базовых представлений.

Результаты контрольного обучения сравниваемых методов и их анализ

Результаты контрольного обучения без прореживания (baseline-60) и трех традиционных методов прореживания модели нейронной сети: структурированного, неструктурированного и полуструктурированного — представлены в табл. 1. Для повышения достоверности результатов эксперименты на ResNet-50 выполнялись с тремя фиксированными значениями начальных условий генераторов случайных чисел, поэтому в табл. 1 приведены средние значения метрик (mean) и стандартные отклонения (std). Это позволяет оценить устойчивость методов к стохастическим факторам обучения и дообучения.

■ ***Таблица 1.*** Значения метрик качества традиционных методов прореживания

■ ***Table 1.*** Values of quality metrics for traditional thinning methods

Метод	Метрика					
	Top-1, % (mean±std)	Top-5, % (mean±std)	P, млн	V, МБ	FLOPs, млрд	T, мс (mean±std)
Baseline-60 (без прореживания)	73,34±0,13	92,73±0,08	23,71	90,43	1,311	7,73±0,55
Структурированное прореживание по глубине	60,84±0,41	86,40±0,36	15,26	58,21	0,807	7,12±3,06
Неструктурированное прореживание	38,65±0,32	69,88±0,17	23,71	90,43	1,311	9,86±3,33
Полуструктурированное (n:m)-прореживание	58,68±1,03	84,04±0,93	23,71	90,43	1,311	6,91±1,31

Анализ результатов из табл. 1 показывает, что число параметров, размер модели нейронной сети и FLOPs не меняются при неструктурированном и (n:m)-прореживании — методы не изменяют размерности тензоров, а меняется структура разреженности. Структурированное прореживание уменьшает число параметров и FLOPs за счет удаления каналов.

Базовая модель (baseline-60), обученная на 40 эпохах, показала Top-1 = 58,33 %, тогда как после продолжения обучения до 60 эпох точность возрасла до 73,34 %. Это подтверждает, что корректное сравнение методов прореживания требует фиксации вычислительного бюджета обучения, поскольку дополнительное дообучение существенно улучшает качество.

Неструктурированное прореживание в проведенных экспериментах демонстрирует заметное снижение точности при заданной доле разреженности 0,7 (Top-1 ≈ 38,65 %). Следует отметить, что эффективность неструктурированного прореживания существенно зависит от выбора схемы дообучения и гиперпараметров, а также может требовать специализированных процедур восстановления качества.

В рамках данной работы использовалась единая схема дообучения для всех методов, что позволило обеспечить сопоставимость условий, однако не предполагает отдельного тюнинга под неструктурированное прореживание. Кроме того, неструктурированное прореживание не приводит к снижению FLOPs и размера модели без использования специализированных аппаратных или программных оптимизаций, что ограничивает его практическую применимость на стандартных устройствах.

Структурированный метод сокращает число параметров примерно на 36 % (с 23,7 до 15,3 млн) и уменьшает FLOPs с 1,31 до 0,81 млрд, однако итоговая точность остается на уровне около 60,84 %, что существенно ниже baseline-60.

Метод (n:m)-прореживания дает точность 59,19 %, что сопоставимо со структурированным подходом, но без сокращения числа параметров и FLOPs.

Далее продемонстрируем результаты предложенного метода геометрически контролируемого прореживания для модели ResNet-50 на датасете CIFAR-100. В эксперименте варьировался параметр ε_{lim}, определяющий дополнительный допустимый бюджет деформации геометрии представлений. Для каждого значения $\varepsilon_{\text{lim}} \in \{0{,}06; 0{,}1; 0{,}14; 0{,}18\}$ выполнялось три независимых запуска с различными фиксированными значениями начальных условий генераторов случайных чисел.

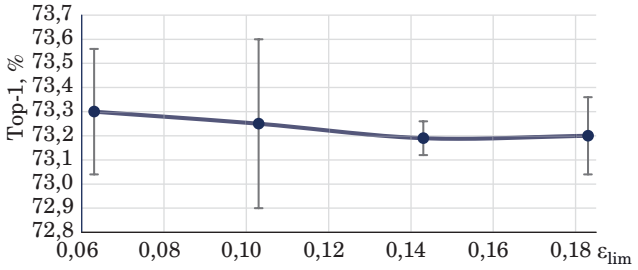
Анализ метрик предложенного метода геометрически контролируемого прореживания по табл. 2 показывает, что увеличение ε_{lim} приводит к более выраженному структурному упрощению модели: число параметров снижается с 22,88 млн при $\varepsilon_{\text{lim}} = 0{,}06$ до 20,80 млн при $\varepsilon_{\text{lim}} = 0{,}18$, а вычислительная сложность уменьшается с 1,256 млрд FLOPs до 1,212 млрд FLOPs. Таким образом, метод ориентирован не на агрессивное сжатие, а на контролируемое структурное упрощение с минимальной деформацией геометрии представлений. Параметр ε_{lim} обеспечивает интерпретируемое управление компромиссом между степенью сжатия и ограничением на деформацию геометрии представлений. При этом итоговая точность Top-1 во всех рассмотренных режимах остается близкой к уровню базовой модели после дообучения, демонстрируя значения порядка 73,2–73,3 % (рис. 1). Это указывает на то, что предложенный метод позволяет выполнять сокращение вычислительных затрат без существенного ухудшения качества классификации, что особенно важно для практического применения на ресурсно-ограниченных устройствах.

Отдельного внимания заслуживает поведение метрики геометрии представлений. Значение ΔG до дообучения закономерно возрастает при увеличении ε_{lim}, что отражает факт более сильного вмешательства в архитектуру сети (рис. 2). Однако после дообучения ΔG стабилизируется около 0,14–0,15, что позволяет предположить наличие эффекта частичного восстановления структуры представлений в процессе дообучения.

■ ***Таблица 2.*** Значения метрик метода геометрически контролируемого прореживания

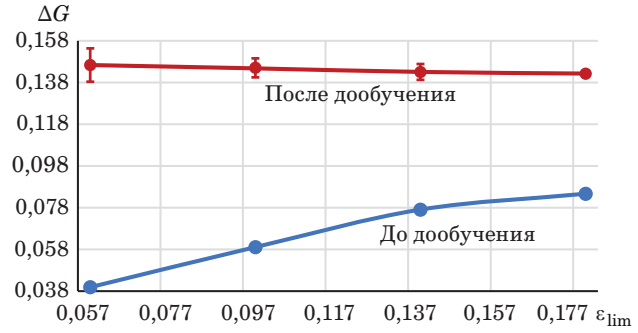
■ ***Table 2.*** Values of the metrics of the geometrically controlled thinning method

ε_{lim}	Метрика						
	Top-1, % (mean±std)	Top-5, % (mean±std)	P, млн	V, МБ	FLOPs, млрд	ΔG до дообучения (mean±std)	ΔG после дообучения (mean±std)
0,06	**73,30±0,26**	92,80±0,07	22,88	87,29	1,256	**0,0399±0,0008**	0,1464±0,0080
0,10	73,24±0,36	92,65±0,06	22,11	84,35	1,233	0,0591±0,0014	0,1448±0,0024
0,14	73,18±0,09	**92,74±0,24**	21,19	80,84	1,218	0,0771±0,0006	**0,1431±0,0038**
0,18	73,21±0,15	92,80±0,10	**20,80**	**79,35**	**1,212**	0,0847±0,0009	0,1423±0,0014



■ ***Рис. 1.*** Точность Top-1 в зависимости от ε_{lim} геометрически контролируемого метода прореживания

■ ***Fig. 1.*** Top-1 accuracy as a function ε_{lim} of the geometrically controlled pruning method



■ ***Рис. 2.*** Значение ΔG до и после дообучения

■ ***Fig. 2.*** The value of ΔG before and after fine-tuning

В ходе экспериментов наблюдается устойчивый рост значения ΔG после дообучения по сравнению со значением на этапе прореживания. Это связано с тем, что дообучение оптимизирует функцию потерь и не содержит явного ограничения на геометрию представлений, вследствие чего происходит дополнительная перестройка признакового пространства.

Таким образом, геометрическое ограничение не является инвариантным свойством финальной модели, а выступает как механизм ограничения области поиска на этапе структурного упрощения.

Несмотря на наблюдаемый рост значения ΔG после дообучения, практическая значимость геометрического ограничения подтверждается сравнением при равном бюджете параметров. При сопоставимом числе параметров (~21 млн) структурированное прореживание без геометрического контроля обеспечивает Top-1 = 59,93 %, тогда как с предложенным методом достигается 73,21 %.

Все методы сравниваются при едином протоколе дообучения и фиксированных гиперпараметрах, что позволяет изолировать влияние критерия прореживания. В этой постановке наблюдаемое преимущество не может быть объяснено различиями в настройке или вычислительном бюджете.

Полученный результат показывает, что геометрическое ограничение играет ключевую роль на этапе прореживания, ограничивая разрушение структуры представлений и формируя более устойчивую инициализацию для последующей оптимизации, несмотря на дальнейшую перестройку признакового пространства.

Проведен дополнительный эксперимент на модели ResNet-18 для CIFAR-100. В отличие от ResNet-50 с Bottleneck-блоками, ResNet-18 использует BasicBlock, что позволяет оценить применимость предлагаемого геометрического критерия для различных типов сверточных архитектур.

В качестве базовой модели использовалась нейронная сеть, обученная в течение 40 эпох. Далее выполнялось дообучение на 20 эпохах без прореживания, а также применялся геометрически контролируемый метод прореживания с параметром $\varepsilon_{lim} = 0{,}18$ и последующим дообучением в течение 20 эпох. Итоговые результаты приведены в табл. 3.

Для ResNet-18 при $\varepsilon_{lim} = 0{,}18$ наблюдается снижение ΔG с 0,0913 до 0,0305, что указывает на частичное восстановление структуры пространства признаков в процессе оптимизации.

Таким образом, эксперимент демонстрирует, что геометрически контролируемое прореживание обеспечивает мягкое уменьшение числа параметров и вычислительной сложности при сохранении точности классификации относительно контрольного варианта дообучения без прореживания.

Для оценки переносимости предложенного метода на архитектуры иного типа дополнительно проведены эксперименты на MobileNetV2

■ ***Таблица 3.*** Результаты работы геометрически контролируемого прореживания на модели ResNet-18

■ ***Table 3.*** Results of geometrically controlled pruning on the ResNet-18 model

Метод	Метрика				
	Top-1, %	Top-5, %	*P*, млн	*V*, МБ	FLOPs, млрд
Baseline ResNet-18 (40 эпох)	71,15	91,83	11, 22	42,80	0,56
Baseline ResNet-18 + дообучение (20 эпох)	71,63	91,78	11,22	42,80	0,56
Геометрически контролируемое прореживание при $\varepsilon_{lim} = 0{,}18$ + дообучение (20 эпох)	**71,72**	**91,81**	**10, 19**	**38,86**	**0,53**

■ ***Таблица 4.*** Результаты работы геометрически контролируемого прореживания на модели MobileNetV2
■ ***Table 4.*** Results of geometrically controlled pruning on the MobileNetV2 model

Метод		Метрика			
		P, млн	FLOPs, млн	*V*, МБ	Top-1, %
MobileNetV2 + дообучение (20 эпох)		2,35±0,00	26,17±0,00	8,97±0,00	59,78±0,15
Геометрически контролируемое прореживание	при ε_{lim} = 0,10	2,25±0,01	24,85±0,11	8,58±0,02	59,81±0,07
	при ε_{lim} = 0,18	2,25±0,01	24,54±0,13	8,57±0,02	59,42±0,17
	при ε_{lim} = 0,30	2,22±0,03	24,30±0,05	8,33±0,15	59,50±0,23
	при ε_{lim} = 0,70	2,17±0,02	23,93±0,00	8,28±0,00	59,55±0,35

(CIFAR-100). Процедура GC-GPEC применена к InvertedResidual-блокам с ограничением на изменение межклассовой геометрии и анализом различных значений параметра ε_{lim}, задающего допустимое изменение геометрии представлений. Результаты (табл. 4) показывают, что увеличение ε_{lim} приводит к последовательному снижению числа параметров (с ~2,35M до ~2,17M) и FLOPs (с ~26,17M до ~23,93M) при минимальной деградации точности (в диапазоне 59,4—59,8 % Top-1 по сравнению с (59,78±0,15) % для baseline).

Особое внимание уделено практической применимости метода. Для этого проведено измерение задержки инференса на реальном встраиваемом устройстве Raspberry Pi 4B (CPU, batch size = 1) с использованием OpenCV DNN backend. Для каждой модели выполнялось 10 независимых запусков. Каждый запуск включал этап прогрева (warm-up, 30 итераций) и 100 измерений инференса. Результаты показали, что предложенный метод обеспечивает стабильное снижение задержки инференса: с (7,80±0,15) мс для baseline и до (7,30±0,10) мс для прореженных моделей (ε_{lim} = 0,18), что соответствует ускорению порядка 4—5 %.

Следует отметить, что наблюдаемое ускорение инференса соответствует умеренному уровню структурного прореживания. В отличие от агрессивных методов, ориентированных на максимальное снижение вычислительной сложности, предложенный подход направлен на сохранение геометрии представлений и стабильности модели.

Дополнительно был проведен анализ зависимости итоговой точности от значения ΔG до дообучения на основе уже полученных экспериментальных данных.

В экспериментах на архитектурах ResNet-50 и MobileNetV2 не выявлено случаев, при которых малое значение ΔG приводило бы к деградации точности после дообучения. При этом зависимость между ΔG и точностью не является строго функциональной: близкие значения ΔG могут соответствовать незначительно различающимся результатам, а увеличение ΔG не всегда приводит к ухудшению качества. Это подтверждает, что метрика ΔG выступает как критерий, ограничивающий чрезмерную деформацию пространства представлений, но не являющийся точным предиктором итоговой точности.

Дополнительно проведен анализ чувствительности метода к параметру C_{min} на архитектуре MobileNetV2. Установлено, что в умеренных режимах прореживания (ε_{lim} = 0,1 и 0,3) изменение C_{min} в диапазоне от двух до 16 не влияет на итоговые характеристики модели, что указывает на доминирующую роль геометрического ограничения ΔG. В то же время в более агрессивном режиме (ε_{lim} = 0,7) влияние параметра становится заметным: уменьшение C_{min} приводит к более сильному сжатию модели и росту ΔG до дообучения, что подтверждает его роль как дополнительного структурного ограничения.

Анализ чувствительности к выбору набора кандидатов прореживания показал, что более плотная дискретизация (0,05; 0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6) позволяет достигать большей степени сжатия при сопоставимом значении ΔG и практически неизменной точности. Это связано с более точным приближением к границе допустимой деформации геометрии представлений, при этом метод в целом демонстрирует устойчивость к выбору набора кандидатов.

Заключение

Основным результатом работы является новый метод сжатия модели глубокой нейронной сети, основанный на геометрически контролируемом прореживании ее структуры.

Эксперимент на сверточной нейронной сети с архитектурой ResNet-50 показал, что геометрически контролируемый метод прореживания модели нейронной сети обеспечивает устойчивое снижение вычислительной сложности модели при сохранении точности классификации: метод демонстрирует сопоставимую точность Top-1 и Top-5 по сравнению с базовой моделью после дообучения, одновременно снижая FLOPs и число параметров.

Дополнительно выполнена проверка переносимости геометрически контролируемого метода прореживания модели нейронной сети на архитектуру ResNet-18 и MobileNetV2. Результаты эксперимента показали, что предложенный метод сохраняет эффективность и для модели с другой структурой блоков, обеспечивая уменьшение числа параметров и FLOPs без снижения точности классификации в сравнении с обучением без прореживания при одинаковом бюджете дообучения.

Полученные результаты подтверждают перспективность использования геометрических ограничений как критерия качества при прореживании, а также открывают возможности для дальнейших исследований. К таким направлениям относятся оптимальное распределение геометрического бюджета по слоям с учетом внутриклассовой дисперсии представлений, расширение подхода на другие архитектуры, а также перенос на задачи классификации с более высокоразрешенными данными.

Литература

1. **Dantas P. V., Sabino da Silva W., Cordeiro L. C., Carvalho C. B.** A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 2024, vol. 54, no. 22, pp. 11804–11844. doi:10.1007/s10489-024-05747-w
2. **Кузнецов А. В.** Цифровая история и искусственный интеллект: перспективы и риски применения больших языковых моделей. *Новые информационные технологии в образовании и науке*, 2022, № 5, с. 53–57. doi:10.17853/2587-6910-2022-05-53-57, EDN: VFYSAN
3. **Liu D., Zhu Y., Liu Z., Liu Y., Han C., Tian J., Li R., Yi W.** A survey of model compression techniques: Past, present, and future. *Front Robot AI*, 2025. doi:10.3389/frobt.2025.1518965
4. **Богачев И. В., Булканов Д. Е.** Обзор современных нейросетевых методов сжатия для задачи обработки измерительных данных. *Вестник Тихоокеанского государственного университета*, 2024, № 2 (73), с. 83–92. doi:10.38161/1996-3440-2024-2-83-92, EDN: EBDQZC
5. **He Y., Xiao L.** Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, iss. 5, pp. 2900–2919. doi:10.1109/TPAMI.2023.3334614
6. **Чернышов Н. Д., Буряк Д. Ю.** Исследование влияния метода сравнения каналов на эффективность алгоритмов поканального прореживания сверточных нейронных сетей. *Системный анализ в науке и образовании*, 2025, № 1, с. 16–22. EDN: JFTUOQ. https://sanse.ru/index.php/sanse/article/view/648 (дата обращения: 20.03.2026).
7. **Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu**. *Full parameter fine-tuning for large language models with limited resources*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024, vol. 1: Long Papers. Bangkok, Association for Computational Linguistics, 2024, pp. 8187–8198.
8. **Zhang Q., Zhang R., Sun J., Liu Y.** How sparse can we prune a deep network: A fundamental limit perspective. *Advances in Neural Information Processing Systems: Proceedings of the NeurIPS Conference*, 2024, vol. 37, pp. 91337–91372. doi:10.52202/079017-2898
9. **Zhu K., Hu F., Ding Y., Zhou W., Wang R.** A comprehensive review of network pruning based on pruning granularity and pruning time perspectives. *Neurocomputing*, 2025, vol. 626, Article 129382. doi:10.1016/j.neucom.2025.129382
10. **Татарникова Т. М., Мокрецов Н. С.** Оптимизация моделей дистилляции знаний для языковых моделей. *Научно-технический вестник информационных технологий, механики и оптики*, 2025, т. 25, № 4, с. 737–743. doi:10.17586/2226-1494-2025-25-4-737-743, EDN: PSPNOU
11. **Кузьмин В. Н., Менисов А. Б., Сабиров Т. Р.** Метод оптимизации нейронных сетей на основе структурной дистилляции с применением генетического алгоритма. *Научно-технический вестник информационных технологий, механики и оптики*, 2024, т. 24, № 5, с. 770–778. doi:10.17586/2226-1494-2024-24-5-770-778, EDN: SKBJQT
12. **Park C., Park M., Moon H., Yoon M.K., Go S., Kim S., Ro W. W.** DEPrune: Depth-wise separable convolution pruning for maximizing GPU parallelism. *Advances in Neural Information Processing Systems: Proceedings of NeurIPS Conference*, 2024, pp. 106906–106923. doi:10.52202/079017-3394
13. **Sun X., Shi H.** Towards better structured pruning saliency by reorganizing convolution. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 2193–2203. doi:10.1109/WACV57701.2024.00220
14. **Wright D., Igel C., Selvan R.** BMRS: Bayesian model reduction for structured pruning. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024, pp. 64119-64144. doi:10.52202/079017-2045
15. **Dong Z., Duan Y., Zhou Y., Duan S., Hu X.** Weight-adaptive channel pruning for CNNs based on closeness-centrality modeling. *Applied Intelligence*, 2024, vol. 54, pp. 201–215. doi:10.1007/s10489-023-05164-5
16. **Khan N. A., Rafat A. M. S.** Pruning convolution neural networks using filter clustering based on normalized cross-correlation similarity. *Journal of Information and Telecommunication*, 2024, vol. 9(2), pp. 190–208. doi:10.1080/24751839.2024.2415008

17. **Duan Z., Lu M., Ma J., Huang Y., Ma Z., Zhu F.** Quantization-aware ResNet VAE for lossy image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, iss. 1, pp. 436–450. doi:10.1109/TPAMI.2023.3322904
18. **Татарникова Т. М., Раскопина А. С.** Анализ эффективности постобучающего квантования для оптимизации нейронных сетей. *Наукоемкие технологии в космических исследованиях Земли*, 2025, т. 17, № 2, с. 4–10. doi:10.36724/2409-5419-2025-17-2-4-10, EDN: LUUUAK
19. **Yu H., Zhang W., Ji M., Zhen C.** Automatic channel pruning method by introducing additional loss for deep neural networks. *Neural Processing Letters*, 2022, vol. 55, pp. 1071–1085. doi:10.1007/s11063-022-10926-2
20. **Liu Y., Wu D., Zhou W., Fan K., Zhou Z.** An effective automatic channel pruning for neural networks. *Neurocomputing*, 2023, vol. 526, pp. 131–142. doi:10.1016/j.neucom.2023.01.014
21. **Hu Y., Zhu J., Chen J.** Continuous pruning function for efficient 2:4 sparse pre-training. *Conference: Advances in Neural Information Processing Systems*, 2024, pp. 33756–33778. doi:10.52202/079017-1063
22. **Shi Y., Tang A., Niu L., Zhou R.** Sparse optimization guided pruning for neural networks. *Neurocomputing*, 2024, vol. 574, Article 127280. doi:10.1016/j.neucom.2024.127280
23. **Krizhevsky A., Hinton G.** *Learning Multiple Layers of Features from Tiny Images*: Technical report. University of Toronto, 2009. https://www.cs.toronto.edu/~kriz/cifar.html (дата обращения: 14.01.2026).
24. **Huang H., Pao H.-K.** Interpretable deep model pruning. *Neurocomputing*, 2025, vol. 647, Article 130485. doi:10.1016/j.neucom.2025.130485
25. **Ganguli T., Chong E.** Activation-based pruning of neural networks. *Algorithms*, 2024, vol. 17, iss. 1, pp. 48. doi:10.3390/a17010048
26. **Cui J., Wang Z., Yang Z., Guan X.** A pruning method based on feature map similarity score. *Big Data and Cognitive Computing*, 2023, vol. 7, iss. 1, pp. 159. doi:10.3390/bdcc7040159
27. **Zhao C., Zhang Y., Ni B.** Exploiting channel similarity for network pruning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, vol. 33, iss. 9, pp. 5049–5061. doi:10.1109/TCSVT.2023.3248659

UDC 004.08
doi:10.31799/1684-8853-2026-3-2-13
EDN: WKWZGY

A deep neural network compression method based on geometrically controlled thinning

T. M. Tatarnikova[a], Dr. Sc., Tech., Professor, orcid.org/0000-0002-6419-0072, tm-tatarn@yandex.ru
A. S. Raskopina[a], Post-Graduate Student, Assistant Professor, orcid.org/ 0009-0002-0276-607X
[a]Saint-Petersburg State University of Aerospace Instrumentation, 67, B. Morskaya St., 190000, Saint-Petersburg, Russian Federation

Introduction: High computational resources, energy costs, and time required to solve problems using deep learning technologies have necessitated the search for solutions to compressing neural network models without significant loss of result quality. **Purpose:** To develop a method for compressing deep neural networks, reducing the computational complexity and the number of parameters of convolutional neural networks without significantly losing the accuracy of the classification problem. **Results:** A new method for geometrically controlled thinning of neural network models is developed, based on the greedy selection of structural thinning candidates with control over changes in representation geometry. We propose a metric for controlling the preservation of representation geometry in the form of an interclass similarity matrix calculated from the class centroids in the feature space. We introduce a parameter for the admissible budget of representation geometry deformation, and propose an approach to its selection based on an estimate of the noise threshold of the geometric metric. The experimental results have shown that the proposed method for compressing neural network models ensures that classification accuracy after additional training is maintained, comparable to the base model without thinning, while reducing computational complexity by ~8% and the number of parameters by ~12% using the ResNet-50 architecture and the CIFAR-100 dataset as an example. Additionally, the portability of the developed method to the architectures of deep neural networks ResNet-18 and MobileNetV2 is demonstrated. **Practical relevance:** The developed method can find application in solving classification problems on mobile and embedded devices in real time.

Keywords – deep learning, neural network compression, thinning, representation geometry, computational complexity, inference delay, classification quality.

For citation: Tatarnikova T. M., Raskopina A. S. A deep neural network compression method based on geometrically controlled thinning. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 2–13 (In Russian). doi:10.31799/1684-8853-2026-3-2-13, EDN: WKWZGY

References

1. Dantas P. V., Sabino da Silva W., Cordeiro L. C., Carvalho C. B. A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 2024, vol. 54, no. 22, pp. 11804–11844. doi:10.1007/s10489-024-05747-w
2. Kuznetsov A. V. Digital history and artificial intelligence: perspectives and risks of pretrained language models. *New Information Technologies in Education and Science*, 2022, no. 5, pp. 53–57 (In Russian). doi:10.17853/2587-6910-2022-05-53-57, EDN: VFYSAN
3. Liu D., Zhu Y., Liu Z., Liu Y., Han C., Tian J., Li R., Yi W. A survey of model compression techniques: Past, present, and future. *Front Robot AI*, 2025. doi:10.3389/frobt.2025.1518965
4. Bogachev I. V., Bulkanov D. E. Review of modern neural network compression methods for the task of processing measurement data. *Bulletin of Pacific National University*, 2024, vol. 2, no. 73, pp. 83–92 (In Russian). doi:10.38161/1996-3440-2024-2-83-92, EDN: EBDQZC
5. He Y., Xiao L. Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 2024, vol. 46, iss. 5, pp. 2900–2919. doi:10.1109/TPAMI.2023.3334614
6. Chernyshov N. D., Buryak D. Yu. Research into impact of channel comparison method on efficiency of algorithms for channel-wise pruning of convolutional neural networks. *System Analysis in Science and Education*, 2025, vol. 1, pp. 16–22. EDN: JFTUOQ. Available at: https://sanse.ru/index.php/sanse/article/view/648 (accessed 20 March 2026) (In Russian).
7. Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu. *Full parameter fine-tuning for large language models with limited resources*. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, vol. 1: Long Papers. Bangkok, Association for Computational Linguistics, 2024, pp. 8187–8198.
8. Zhang Q., Zhang R., Sun J., Liu Y. How sparse can we prune a deep network: A fundamental limit perspective. *Advances in Neural Information Processing Systems: Proceedings of the NeurIPS Conference*, 2024, vol. 37, pp. 91337–91372. doi:10.52202/079017-2898
9. Zhu K., Hu F., Ding Y., Zhou W., Wang R. A comprehensive review of network pruning based on pruning granularity and pruning time perspectives. *Neurocomputing*, 2025, vol. 626, Article 129382. doi:10.1016/j.neucom.2025.129382
10. Tatarnikova T. M., Mokretsov N. S. Optimizing knowledge distillation models for language models. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 4, pp. 737–743 (In Russian). doi:10.17586/2226-1494-2025-25-4-737-743, EDN: PSPNOU
11. Kuzmin V. N., Menisov A. B., Sabirov T. R. A method for optimizing neural networks based on structural distillation using a genetic algorithm. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 5, pp. 770–778 (In Russian). doi:10.17586/2226-1494-2024-24-5-770-778, EDN: SKBJQT
12. Park C., Park M., Moon H., Yoon M. K., Go S., Kim S., Ro W. W. DEPrune: Depth-wise separable convolution pruning for maximizing GPU parallelism. *Advances in Neural Information Processing Systems: Proceedings of NeurIPS Conference*, 2024, pp. 106906–106923. doi:10.52202/079017-3394
13. Sun X., Shi H. Towards better structured pruning saliency by reorganizing convolution. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 2193–2203. doi:10.1109/WACV57701.2024.00220
14. Wright D., Igel C., Selvan R. BMRS: Bayesian model reduction for structured pruning. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024, pp. 64119–64144. doi:10.52202/079017-2045
15. Dong Z., Duan Y., Zhou Y., Duan S., Hu X. Weight-adaptive channel pruning for CNNs based on closeness-centrality modeling. *Applied Intelligence*, 2024, vol. 54, pp. 201–215. doi:10.1007/s10489-023-05164-5
16. Khan N. A., Rafat A. M. S. Pruning convolution neural networks using filter clustering based on normalized cross-correlation similarity. *Journal of Information and Telecommunication*, 2024, vol. 9(2), pp. 190–208. doi:10.1080/24751839.2024.2415008
17. Duan Z., Lu M., Ma J., Huang Y., Ma Z., Zhu F. QARV: Quantization-aware ResNet VAE for lossy image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, pp. 436–450. doi:10.1109/TPAMI.2023.3322904
18. Tatarnikova T. M., Raskopina A. S. Analyzing the effectiveness of post-learning quantization for optimizing neural networks. *H&ES Reserch*, 2025, vol. 17, no. 2, pp. 4–10 (In Russian). doi:10.36724/2409-5419-2025-17-2-4-10, EDN: LUUUAK
19. Yu H., Zhang W., Ji M., Zhen C. ACP: Automatic channel pruning method by introducing additional loss for deep neural networks. *Neural Processing Letters*, 2022, vol. 55, pp. 1071–1085. doi:10.1007/s11063-022-10926-2
20. Liu Y., Wu D., Zhou W., Fan K., Zhou Z. EACP: An effective automatic channel pruning for neural networks. *Neurocomputing*, 2023, vol. 526, pp. 131–142. doi:10.1016/j.neucom.2023.01.014
21. Hu Y., Zhu J., Chen J. S-STE: Continuous pruning function for efficient 2:4 sparse pre-training. *Conference: Advances in Neural Information Processing Systems*, 2024, pp. 33756–33778. doi:10.52202/079017-1063
22. Shi Y., Tang A., Niu L., Zhou R. Sparse optimization guided pruning for neural networks. *Neurocomputing*, 2024, vol. 574, Article 127280. doi:10.1016/j.neucom.2024.127280
23. Krizhevsky A., Hinton G. *Learning Multiple Layers of Features from Tiny Images*: Technical report. University of Toronto, 2009. Available at: https://www.cs.toronto.edu/~kriz/cifar.html (accessed 14 January 2026).
24. Huang H., Pao H.-K. Interpretable deep model pruning. *Neurocomputing*, 2025, vol. 647, Article 130485. doi:10.1016/j.neucom.2025.130485
25. Ganguli T., Chong E. Activation-based pruning of neural networks. *Algorithms*, 2024, vol. 17, iss. 1, pp. 48. doi:10.3390/a17010048
26. Cui J., Wang Z., Yang Z., Guan X. A pruning method based on feature map similarity score. *Big Data and Cognitive Computing*, 2023, vol. 7, iss. 1, pp. 159. doi:10.3390/bdcc7040159
27. Zhao C., Zhang Y., Ni B. Exploiting channel similarity for network pruning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, vol. 33, iss. 9, pp. 5049–5061. doi:10.1109/TCSVT.2023.3248659

УДК 004.932.72
doi:10.31799/1684-8853-2026-3-14-22
EDN: ZWXEEH

Научные статьи
Articles

Алгоритм предварительной подготовки оптических обучающих данных, повышающий эффективность их использования в задачах нейросетевой обработки радиолокационных изображений

В. А. Павлов[a], *канд. техн. наук, доцент, orcid.org/0000-0003-0726-6613, pavlov_va@spbstu.ru*
А. А. Белов[a], *ведущий инженер, orcid.org/0000-0003-0617-4514*
Ф. Шариати[a], *старший преподаватель, orcid.org/0000-0002-7060-8826*
С. В. Волвенко[a], *старший научный сотрудник, orcid.org/0000-0001-7726-8492*
Л. Гу[a], *канд. техн. наук, инженер-исследователь, orcid.org/0000-0003-1105-5835*
[a]*Санкт-Петербургский политехнический университет Петра Великого, Политехническая ул., 29, Санкт-Петербург, 195251, РФ*

Введение: *несмотря на продемонстрированную в последние годы в ряде исследований возможность применения оптических изображений для обучения нейронных сетей, предназначенных для обработки данных радиолокаторов с синтезированной апертурой, существующие подходы все еще характеризуются недостаточно высокой эффективностью. Это связано с различиями в физической природе формирования изображений и, как следствие, с существенными отличиями классификационных признаков объектов, выделяемых нейронной сетью в процессе обучения.* ***Цель:*** *разработать алгоритм специализированной предварительной обработки оптических изображений для использования в трансферном междоменном обучении нейросетей, обрабатывающих радиолокационные изображения, а также оценить эффективность такой обработки по стандартным метрикам нейросетевого детектора.* ***Результаты:*** *разработан алгоритм специализированной предварительной обработки оптических снимков, включающий преобразование в полутоновое изображение, низкочастотную гауссовскую фильтрацию, выделение границ на основе лапласиана и последующее сглаживание для формирования псевдорадиолокационного представления. В экспериментах детектор YOLO11x обучался на предобработанных оптических изображениях, валидация проводилась на оптических изображениях, а тестирование — на реальных радиолокационных изображениях с видами морских судов. По сравнению с обучением на необработанных оптических изображениях лучшая конфигурация обработки повысила Precision с 0,675 до 0,813, Recall с 0,464 до 0,517, mAP50 с 0,518 до 0,613 и mAP50-95 с 0,202 до 0,342. Полученные результаты показывают, что подавление яркостно-цветовой информации, характерной для оптического диапазона, и выделение признаков, связанных с формой и границами объектов, повышают эффективность переноса между оптическим и радиолокационным доменами в рассмотренных условиях эксперимента.* ***Практическая значимость:*** *результаты работы позволяют использовать широкодоступные размеченные оптические изображения для обучения нейросетей, решающих задачи классификации, сегментации или детектирования объектов на радиолокационных изображениях.*

Ключевые слова *— радиолокатор с синтезированной апертурой, оптические изображения, радиолокационные изображения, обработка изображений, нейронная сеть, обучение, обнаружение объектов.*

Для цитирования: Павлов В. А., Белов А. А., Шариати Ф., Волвенко С. В., Гу Л. Алгоритм предварительной подготовки оптических обучающих данных, повышающий эффективность их использования в задачах нейросетевой обработки радиолокационных изображений. *Информационно-управляющие системы*, 2026, № 3, с. 14–22. doi:10.31799/1684-8853-2026-3-14-22, EDN: ZWXEEH

For citation: Pavlov V. A., Belov A. A., Shariaty F., Volvenko S. V., Gu L. Algorithm for preprocessing optical training data that improves the efficiency of their use in neural network processing tasks of radar images. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 14–22 (In Russian). doi:10.31799/1684-8853-2026-3-14-22, EDN: ZWXEEH

Введение

Значительная часть информации, получаемой средствами дистанционного зондирования Земли, формируется на основе результатов анализа изображений оптического диапазона, собранных спутниками и летательными аппаратами. Оптические изображения (ОИ) имеют высокое качество, их анализ и интерпретация опираются на отработанные методики, однако недостатком является невозможность их съемки ночью и при наличии облачности. Это приводит к проблемам с надежностью и оперативностью получения информации, важными при анализе быстро меняющихся событий, например при реагировании на стихийные бедствия.

Технология получения радиолокационных изображений (РЛИ) с помощью радиолокаторов с синтезированной апертурой также известна давно. Важнейшим преимуществом РЛИ по сравнению с ОИ является существенно меньшая зависимость их информационных характеристик от времени суток и метеорологических условий в районе съемки. Однако в течение длительного

времени широкодоступные РЛИ, как правило, уступали ОИ по пространственному разрешению, которое для массово используемых данных часто составляло десятки метров против единиц метров у оптических снимков. Такие разрешения были достаточны для анализа крупных площадных объектов типа сельскохозяйственных угодий, лесов, разливов нефти и т. п., но недостаточны для надежного обнаружения и анализа состояния более мелких объектов, в первую очередь искусственного происхождения: зданий и других сооружений, линий электропередач, кораблей и т. п.

В последнее время ситуация меняется: широкодоступными становятся РЛИ с разрешением порядка одного метра, близким к разрешению оптических снимков. Такие РЛИ предоставляют несколько компаний: «Роскосмос» (https://gptl.ru/), Capella Space (https://www.capellaspace.com/), ICEYE (https://www.iceye.com/) и др. Более высокое разрешение существенно повышает информационную ценность радиолокационных данных. Однако применение современных методов машинного обучения, особенно нейросетевых, к автоматизированной обработке РЛИ высокого разрешения осложняется рядом факторов, в частности: относительно малым количеством качественно аннотированных обучающих данных для РЛИ по сравнению с ОИ, их высокой стоимостью, а также специфическими особенностями формирования и интерпретации РЛИ.

Нехватка размеченных данных является одним из ключевых факторов, сдерживающих развитие нейросетевых методов обнаружения объектов в радиолокационном диапазоне. В связи с этим в последние годы активно развиваются подходы, позволяющие переносить знания, полученные при анализе ОИ (где аннотированных наборов данных существенно больше [1, 2]), на задачи обработки РЛИ. Этот вектор исследований направлен на сокращение междоменного разрыва и учет специфики РЛИ через многоуровневое согласование доменов и прогрессивный перенос на уровнях пикселей, признаков и предсказаний [3–5]. Наибольший практический интерес представляет сценарий, когда «источник» (ОИ) содержит достаточно большое количество качественно размеченных данных, а «цель» (РЛИ) представлена ограниченным числом размеченных примеров либо вовсе недоступна для обучения, что делает актуальными методы междоменного обучения и доменной адаптации.

В рамках перехода ОИ к РЛИ можно выделить несколько устойчивых направлений. Первое — это методы доменного выравнивания и адаптации, которые уменьшают расхождения распределений ОИ и РЛИ на разных уровнях представления данных. Достигается это за счет применения пиксельных преобразований, выравнивания признаковых пространств и калибровки предсказаний, как правило, в сочетании со схемами «учитель–ученик» (teacher–student) и другими методами полуконтролируемого обучения, включая самообучение (self-training) [3, 4, 6, 7]. Второе направление делает акцент на структурно-текстурных признаках как наиболее стабильных при переходе между оптическим и радиолокационным доменами. Здесь используются архитектурные и алгоритмические приемы, направленные на сохранение геометрии объектов, усиление границ и текстур при подавлении модально-специфичных артефактов, что позволяет эффективнее «совмещать» оптические и радиолокационные представления [6, 7].

Отдельно развивается направление синтетической генерации РЛИ-подобных данных из ОИ. В нем применяются генеративно-состязательные сети и более современные генеративные подходы для построения «промежуточного домена» или прямого перехода ОИ к РЛИ в целях расширения обучающих выборок [8, 9]. Однако в ряде обзоров подчеркивается, что такая трансляция часто является более неоднозначной, чем обратная задача, поскольку один и тот же оптический вид может соответствовать множеству РЛИ-проявлений в зависимости от геометрии съемки и характеристик рассеяния [10, 11]. Альтернативой генеративным подходам является физическое моделирование радиолокационных сигнатур с использованием электромагнитного численного анализа и их внедрение в статистически сгенерированный фон, что позволяет создавать неограниченное количество реалистичных обучающих примеров с контролируемыми параметрами [12]. Такой подход, включающий извлечение сигнатур из реальных снимков с неполной информацией автоматической идентификационной системы судов и синтез новых по 3D-моделям, продемонстрировал повышение точности обнаружения (mAP) на 15–20 % при ограниченной обучающей выборке [12]. Поэтому наряду с генеративными моделями сохраняют актуальность стратегии, основанные на переносе через представления (representation learning) и специализированное предобучение, ориентированное на повышение переносимости признаков между оптическим и радиолокационным доменами [13].

В статьях [14, 15] в целях обучения сверточной нейронной сети для работы с РЛИ (на примере обнаружения морских судов) использовались цветные ОИ, переведенные в градации серого. Однако очевидно, что такой простейший метод адаптации не является оптимальным.

Исследование нацелено на разработку алгоритма специализированной предварительной обработки ОИ для применения в междоменном обучении нейросетевого детектора РЛИ и на оценку его эффективности с помощью стандартных метрик обнаружения объектов.

Описание предлагаемого алгоритма

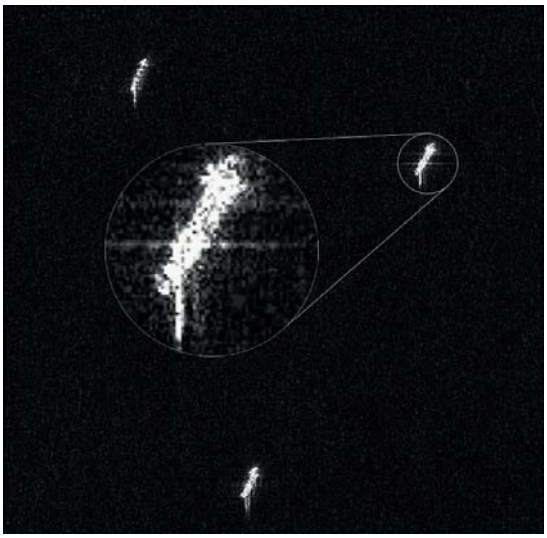
Для сравнения на рис. 1, *а* и *б* представлены оптическое и радиолокационное изображения, содержащие один класс объекта, — морское судно. Сопоставление и анализ оптических и радиолокационных сходных объектов позволяет сделать следующие выводы. В настоящее время типичные данные дистанционного зондирования Земли, получаемые с помощью оптических и радиолокационных средств, уже имеют сравнимое разрешение (порядка одного метра или более высокое). Соответственно, структурные особенности многих объектов (размеры, форма контуров, взаимная ориентация составных частей объекта) одинаково проявляются в обоих случаях. С другой стороны, яркостная (амплитудная) информация отличается существенно. Например, плоские поверхности типа дорожного покрытия, плоской крыши здания, поверхности водоема на РЛИ всегда имеют малую амплитуду (яркость). Яркость таких объектов, в зависимости от их материала и условий освещенности, на ОИ может быть практически произвольной.

Эти наблюдения подтверждаются анализом физических механизмов формирования отражений в оптическом и радиодиапазонах длин волн, связанных в основном с большим различием в длине волны. Известно, что характер рассеяния от неровной поверхности определяется отношением среднего размера ее неровностей (в первую очередь их высоты) к длине волны и может иметь диффузный или зеркальный характер [17]. Соответственно, многие поверхности, шероховатые и диффузно рассеивающие в оптическом диапазоне, являются гладкими и зеркально отражающими в радиодиапазоне, поскольку из-за различия длин волн в оптическом и радиодиапазоне одна и та же поверхность по критериям шероховатости может рассматриваться как шероховатая для ОИ и как гладкая для РЛИ.

Таким образом, при трансформации ОИ в РЛИ-подобные изображения следует ориентироваться на преобразования, сохраняющие структурную информацию и снижающие влияние яркостной и цветовой компонент ОИ. Для удаления цветовой информации могут использоваться известные методы перевода цветных изображений в оттенки серого. Однако они не решают проблему несоответствия поведения яркостных компонент РЛИ и ОИ. Как уже обсуждалось, практически все однородные гладкие поверхности на РЛИ имеют малую амплитуду отражений, а на ОИ — произвольную, и часто достаточно большую. Спектр однородных гладких участков на изображении сконцентрирован в зоне низких пространственных частот, амплитуда которых зависит от яркости данной зоны. Преобразование, сводящее яркость (амплитуду) любой однородной области на изображении к малой яркости, — это подавление низких пространственных частот изображения, достигаемое использованием двумерных фильтров верхних частот (ФВЧ), в частности дифференциаторов. ФВЧ обычно применяется после обработки изображения специально подобранным фильтром низких частот (ФНЧ). Это делается для снижения влияния шума на результаты фильтрации, так как ФВЧ будет дополнительно подчеркивать имеющиеся высокочастотные (ВЧ) компоненты шума. В качестве ФНЧ часто используется фильтр с конечной импульсной характеристикой с ядром Гаусса.

а)

б)



■ ***Рис. 1.*** Оптическое (https://www.kaggle.com/competitions/airbus-ship-detection) (*а*) и радиолокационное (*б*) изображения [16]

■ ***Fig. 1.*** Optical (https://www.kaggle.com/competitions/airbus-ship-detection) (*а*) and radar (*б*) images [16]

В соответствии с вышеизложенным можно предложить следующий алгоритм преобразования ОИ в псевдо-РЛИ.

1. Предварительная НЧ-фильтрация ОИ. Для снижения влияния шума перед дифференцированием цветное изображение поканально сворачивается с ядром Гаусса.

Пусть $G_{\sigma_1}(x,y)$ — ядро Гаусса с параметром σ_1 (стандартное отклонение):

$$G_{\sigma_1}(x,y) = \frac{1}{2\pi\sigma_1^2} e^{-\frac{x^2+y^2}{2\sigma_1^2}},$$

где x, y — целочисленные координаты, изменяющиеся в пределах $[-r_1, r_1]$, а размер ядра $k_1 = 2r_1 + 1$ выбирается как $k_1 = 2[3\sigma_1] + 1$ (что охватывает практически всю энергию гауссианы).

Ядро нормируется таким образом, чтобы сумма всех его элементов равнялась единице: $\sum\sum G_{\sigma_1}(x,y) = 1$, что гарантирует сохранение средней яркости однородных областей. Тогда сглаженное изображение $I_{smooth}(x, y)$ (цветное) представляет собой результат свертки (обозначается $*$) исходного изображения с этим ядром:

$$I_{smooth}(x,y) = G_{\sigma_1}(x,y) * I(x,y).$$

Обработка границ изображения осуществляется методом дополнения нулями (zero-padding), т. е. пиксели, выходящие за пределы исходного изображения, считаются равными нулю.

2. Преобразование цветного ОИ в уровни серого. Полученное сглаженное цветное изображение путем поканального усреднения преобразуется в полутоновое:

$$I_{gray}(x,y) = \\ = \frac{1}{3}\left(I^R_{smooth}(x,y) + I^G_{smooth}(x,y) + I^B_{smooth}(x,y)\right).$$

3. ВЧ-фильтрация полученного изображения (выделение границ). Детектор границ реализован на основе лапласиана:

$$I_{edge}[i, j] = \sum_{m=-2}^{2}\sum_{n=-2}^{2} L[m, n] I_{gray}[i-m, j-n],$$

где L — ядро лапласиана. В работе использовалось ядро 5×5

$$L = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 24 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix}.$$

4. Дополнительная НЧ-фильтрация изображения. Пусть $G_{\sigma_2}(x,y)$ — ядро Гаусса с параметром σ_2:

$$G_{\sigma_2}(x,y) = \frac{1}{2\pi\sigma_2^2} e^{-\frac{x^2+y^2}{2\sigma_2^2}}.$$

Размер ядра $k_2 = 2[3\sigma_2] + 1$, нормировка: $\sum\sum G_{\sigma_2}(x,y) = 1$. Тогда сглаженное изображение представляет собой результат свертки исходного изображения с этим ядром с заполнением нулями на границах:

$$I^{edge}_{smooth}(x,y) = G_{\sigma_2}(x,y) * I_{edge}(x,y).$$

Стоит отметить, что порядок выполнения пп. 1 и 2 не является существенным. Алгоритм выглядит несложным, но для его эффективного использования важен правильный подбор параметров фильтров, т. е. необходимо оценить влияние параметров применяемых фильтров на качество обнаружения объектов.

Ниже приведены примеры изображений, полученных в результате применения алгоритма с тремя различными наборами фильтров (рис. 2, *а–г*).

Комбинация фильтров 1: последовательное применение низкочастотной гауссовской фильтрации с параметром ядра $\sigma = 3$, преобразования в градации серого методом усреднения каналов, ВЧ-фильтрации детектором границ на основе лапласиана и финального низкочастотного гауссовского сглаживания с параметром $\sigma = 20$. Результат предобработки представлен на рис. 2, *б*.

Комбинация фильтров 2: предварительная низкочастотная гауссовская фильтрация с параметром $\sigma = 20$, преобразование в градации серого методом усреднения каналов, ВЧ-фильтрация детектором границ на основе лапласиана и финальное НЧ-сглаживание с параметром $\sigma = 20$. Таким образом, оба этапа НЧ-обработки выполняются с одинаковым и значительным радиусом размытия, что обеспечивает интенсивное сглаживание как исходного, так и промежуточного изображения. Результат предобработки представлен на рис. 2, *в*.

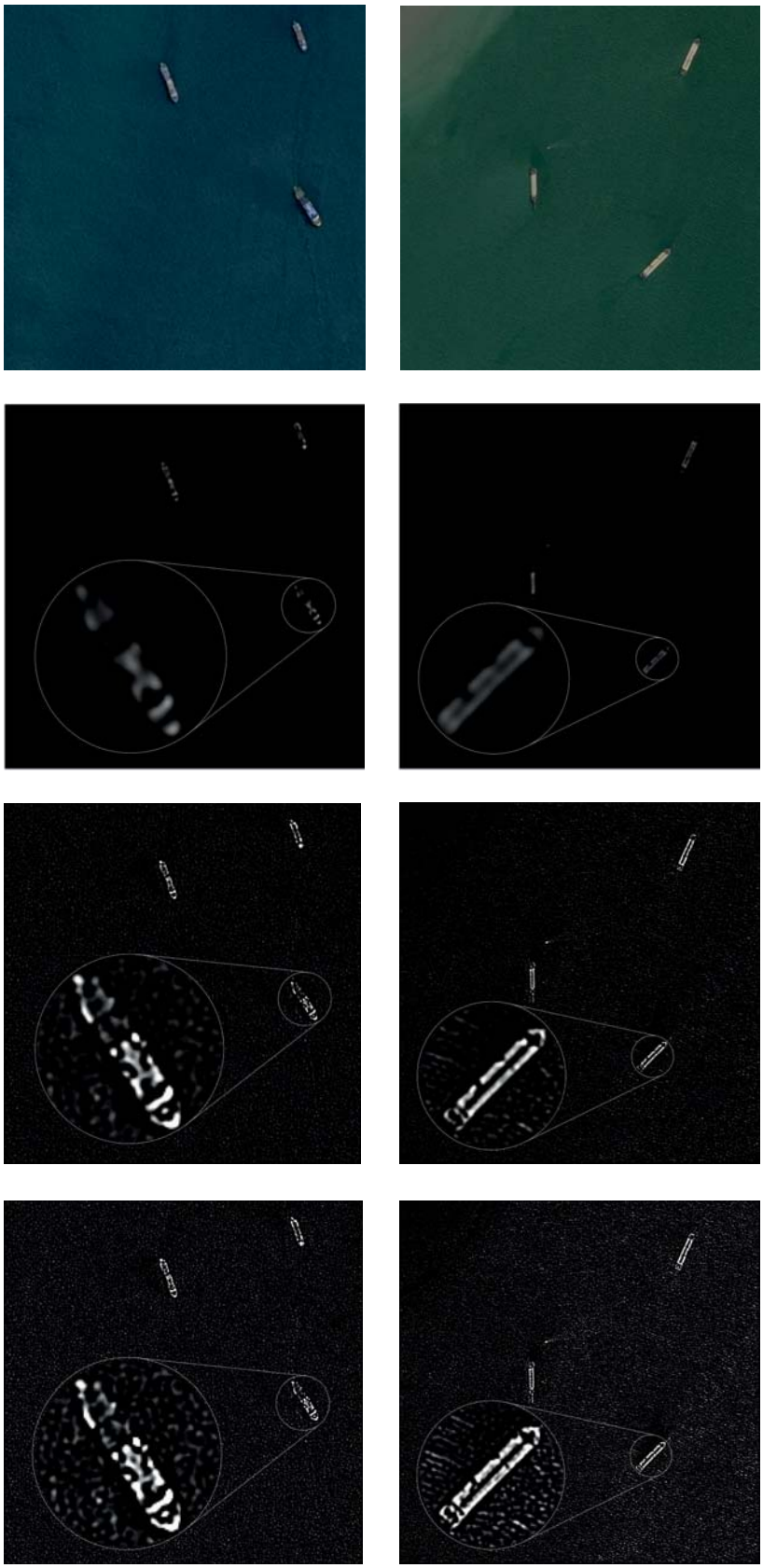
Комбинация фильтров 3: преобразование в градации серого методом усреднения каналов, ВЧ-фильтрация детектором границ на основе лапласиана и последующее низкочастотное гауссовское сглаживание с параметром $\sigma = 20$. В отличие от предыдущих вариантов, здесь отсутствует предварительная НЧ-фильтрация, что позволяет сохранить более мелкие детали исходного изображения перед выделением границ, однако финальное размытие со значительным радиусом обеспечивает итоговую «радароподобную» сглаженность. Результат предобработки представлен на рис. 2, *г*.

а)

б)

в)

г)



■ ***Рис. 2.*** Примеры оптических изображений: *а* – исходные; *б* – обработанные комбинацией фильтров 1; *в* – обработанные комбинацией фильтров 2; *г* – обработанные комбинацией фильтров 3

■ ***Fig. 2.*** Examples of optical images: *а* – originals; *б* – processed by a combination of filters 1; *в* – processed by a combination of filters 2; *г* – processed by a combination of filters 3

Результаты и обсуждение

При визуальном анализе наиболее соответствующим реальным РЛИ представляется набор, полученный комбинацией фильтров 2. Однако визуальная оценка результатов предварительной обработки позволяет судить только о внешнем сходстве полученных псевдо-РЛИ с реальными РЛИ и не является достаточным критерием эффективности подготовки обучающих данных. Поэтому в работе эффективность предварительной обработки оценивалась косвенно — по результатам обнаружения морских судов на тестовой выборке реальных РЛИ. Для этого нейросетевой детектор обучался на ОИ, обработанных различными комбинациями фильтров, а затем тестировался на реальных РЛИ; в качестве показателей эффективности использовались стандартные метрики детекторов объектов.

Для проведения экспериментов был выбран нейросетевой детектор YOLO11x [18], обладающий высокой точностью обнаружения объектов на ОИ. Для каждого набора фильтров выполнялось обучение YOLO11x: длительность 30 эпох, размер батча 8, входное разрешение 640×640. Обучение и валидация проводились на предобработанных ОИ, тестирование — на РЛИ. Размер наборов данных: обучающий — 1610 ОИ, валидационный — 403 ОИ, тестовый — 5604 РЛИ с изображениями видов морских судов. В качестве метрик качества обнаружения использовались стандартные показатели машинного обучения [19, 20]: точность (Precision), полнота (Recall), mAP50, mAP50-95. Указанные метрики применялись в качестве показателей эффективности предварительной обработки; более высокие значения этих метрик соответствуют лучшему переносу признаков с ОИ на РЛИ.

Обучающая и валидационная выборки сформированы из ОИ конкурса Airbus Ship Detection (источник — спутники Airbus SPOT 6/7, цветные изображения, разрешение 1,5 м) (https://www.kaggle.com/competitions/airbus-ship-detection). Тестовая выборка РЛИ взята из общедоступного набора данных HRSID [16], содержащего снимки со спутников Sentinel-1B (C-диапазон, разрешение 1,7–4,9 м, поляризации HH, HV, VV), TerraSAR-X и TanDEM-X (X-диапазон, разрешение 0,25–3 м, поляризации HH, VV). Все РЛИ имеют пространственное разрешение от одного до 5 м, представлены в виде кополяризационных (116 снимков) и кросс-поляризационных (20 снимков) режимов. Подробные характеристики исходных панорамных снимков (углы падения, режимы съемки, количество судов) приведены в [16]. Оба набора содержат изображения с видами морских судов.

■ ***Таблица 1.*** Результаты обнаружения морских судов на тестовой выборке РЛИ

■ ***Table 1.*** Ship detection results on the radar-image test set

Тип обработки	Precision	Recall	mAP50	mAP50-95
Без обработки	0,675	0,464	0,518	0,202
Комбинация фильтров 1	0,785	0,492	0,563	0,250
Комбинация фильтров 2	0,821	0,517	0,608	0,313
Комбинация фильтров 3	0,786	0,484	0,569	0,321

Результаты тестирования на реальных РЛИ, приведенные в табл. 1, показывают повышение значений метрик при использовании предложенной предварительной обработки. Все комбинации фильтров демонстрируют более высокие значения метрик по сравнению с базовым вариантом (без обработки) по всем метрикам. Наилучший баланс точности и полноты обеспечивает комбинация фильтров 2, увеличивающая Precision с 0,675 до 0,821, Recall с 0,464 до 0,517, а ключевую метрику mAP50 с 0,518 до 0,608. Это свидетельствует о том, что предварительная обработка ОИ, учитывающая отдельные характеристики РЛИ, повышает эффективность обучения сверточной нейронной сети при переносе на реальные радиолокационные данные. Модель, обученная на необработанных ОИ, демонстрирует более низкие значения метрик при переносе на целевую предметную область.

Исследование влияния размера ядра предварительной гауссовской фильтрации (табл. 2) позволило определить значение параметра, при котором в проведенных экспериментах были получены наибольшие значения метрик для настройки алгоритма преобразования. Увеличение размера ядра с 10 до 20 практически не влияет на Precision и Recall, но повышает mAP50-95 с 0,312 до 0,313. Дальнейшее увеличение до 30 пикселей приводит к росту обобщающей способности модели (mAP50-95 до 0,342) при сохранении высокой точности (0,813). При размере ядра 40 наблюдается резкое ухудшение всех метрик, что, вероятно, связано с чрезмерным сглаживанием и потерей мелких, но важных для обнаружения объектов деталей. Таким образом, размер ядра 30 пикселей является наиболее предпочтительным выбором для данного набора данных.

■ ***Таблица 2.*** Результаты тестирования на РЛИ при изменении размера гауссовского ядра предварительной НЧ-фильтрации

■ ***Table 2.*** Radar-image test results for different Gaussian kernel sizes of the preliminary low-pass filtering in filter combination 2

Размер ядра	Precision	Recall	mAP50	mAP50-95
10	0,803	0,518	0,607	0,312
20	0,821	0,517	0,608	0,313
30	0,813	0,517	0,613	0,342
40	0,805	0,488	0,580	0,317

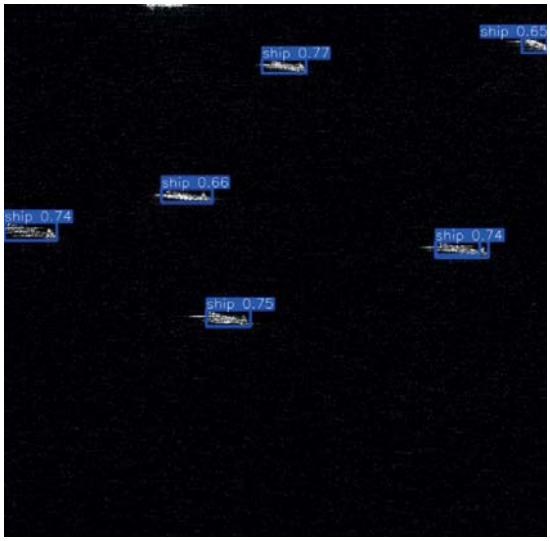
Два примера обнаружения морских судов на реальных тестовых РЛИ с использованием YOLOv11x, обученного на предобработанных ОИ, представлены на рис. 3, *а* и *б*.

Таким образом, проведенные эксперименты подтверждают, что предложенная методика предварительной обработки ОИ, адаптированная к характеристикам РЛИ, повышает эффективность переноса признаков при обучении нейросетевого детектора. Наилучшие результаты достигаются при использовании комбинации фильтров 2 с размером гауссовского ядра 30 пикселей, что обеспечивает оптимальный баланс между сглаживанием шумов и сохранением значимых деталей объектов. Полученные значения метрик Precision, Recall и mAP50 свидетельствуют о практической применимости подхода для решения задачи обнаружения морских судов на реальных РЛИ, в том числе в условиях прибрежной зоны и открытого моря.

Заключение

В ходе проведенного исследования было показано, что применение специализированной предварительной обработки ОИ позволяет повысить значения метрик обнаружения при обучении сверточных нейронных сетей, предназначенных для анализа реальных РЛИ, в рассмотренных условиях эксперимента. Поскольку физические механизмы формирования отражений в оптическом и радиодиапазонах существенно различаются, простой перевод цветных оптических снимков в градации серого является нерациональным подходом, не позволяющим нейронной сети эффективно переносить знания на целевую предметную область. Во всех рассмотренных экспериментальных случаях предложенный алгоритм, направленный на имитацию характеристик РЛИ путем подавления несущественных яркостных компонент и подчеркивания структурной информации об объектах, продемонстрировал значительное улучшение качества обнаружения морских судов по сравнению с использованием немодифицированных ОИ. Наилучший баланс метрик (включая Precision, Recall и mAP) был достигнут при использовании комбинации фильтров с интенсивным сглаживанием, где оптимальный размер ядра предварительной низкочастотной гауссовской фильтрации составил

а)



б)

■ ***Рис. 3.*** Примеры обнаружения морских судов на реальных тестовых РЛИ в открытом море (*а*) и в прибрежной зоне (*б*)

■ ***Fig. 3.*** Examples of ship detection on real test radar images: in open sea (*а*) and in a coastal area (*б*)

30 пикселей, что обеспечило высокую точность и рост обобщающей способности модели. Таким образом, в работе продемонстрирована практическая возможность эффективного вовлечения обширных наборов предварительно обработанных ОИ как непосредственно для процесса обучения нейронных сетей, так и в качестве инструмента аугментации уже имеющихся, но немногочисленных радиолокационных обучающих данных.

Литература

1. **Li K., Wan G., Cheng G.** DIOR: A large-scale dataset for object detection in optical remote sensing images. *IEEE Dataport*, April 12, 2025. doi:10.21227/tq1e-nx82
2. **Xia G. S., Bai X., Ding J., Zhu Z.** DOTA: A large-scale dataset for object detection in aerial images. *IEEE Dataport*, April 12, 2025. doi:10.21227/wwrj-3d46
3. **Jeong S., Kim Y., Kim S., Sohn K.** Enriching SAR ship detection via multistage domain alignment. *IEEE Geoscience and Remote Sensing Letters*, 2021, vol. 19, pp. 1–5. doi:10.1109/LGRS.2021.3115498
4. **Shi Y., Du L., Guo Y., Du Y.** Unsupervised domain adaptation based on progressive transfer for ship detection: From optical to SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, pp. 1–17. doi:10.1109/TGRS.2022.3185298
5. **Liu S., Wang Z., Wang R., Chen Y., Fan X., Li W.** SAR ship detection based on deep domain adaptation with limited samples. *Procedia Computer Science*, 2023, vol. 221, pp. 378–385. doi:10.1016/j.procs.2023.07.051
6. **Zhu Y., Ai J., Xue W., Wang Z., Zhao Z.** Cross-modal ship detection from optical to SAR images based on pixel-level and feature-level progressive transfer. *IEEE Sensors Journal*, 2025, vol. 25, no. 8, pp. 13344–13356. doi:10.1109/JSEN.2025.3543520
7. **Luo C., Zhang Y., Guo J., Hu Y., Zhou G., You H., Ning X.** SAR-CDSS: A semi-supervised cross-domain object detection from optical to SAR domain. *Remote Sensing*, 2024, vol. 16, no. 6, p. 940. doi:10.3390/rs16060940
8. **Wu B., Wang H., Zhang C., Chen J.** Optical-to-SAR translation based on CDA-GAN for high-quality training sample generation for ship detection in SAR amplitude images. *Remote Sensing*, 2024, vol. 16, no. 16, p. 3001. doi:10.3390/rs16163001
9. **Luo X., Lin Y., Xu K., Nam H., Peng D.** *OS-Ship-1K: A CycleGAN-Based Optical-SAR Multimodal Ship Detection Dataset*. 2025. https://openreview.net/forum?id=aIB8eQc9Au (дата обращения: 01.02.2026).
10. **Huang Z., Zhang X., Tang Z., Xu F., Datcu M., Han J.** Generative artificial intelligence meets synthetic aperture radar: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2024, vol. 14, no. 1, pp. 6–48. doi:10.1109/MGRS.2024.3483459
11. **Wang Z., Zhang Z., Shan X., Wei H. A., Tang P.** Generative models for SAR-optical image translation: A systematic review. *International Journal of Applied Earth Observation and Geoinformation*, 2026, vol. 146, Article 105009. doi:10.1016/j.jag.2025.105009
12. **Lee S.-J., Lee K.-J.** Efficient generation of artificial training DB for ship detection using satellite SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, vol. 14, pp. 11764–11774. doi:10.1109/JSTARS.2021.3128184
13. **Bao W., Huang M., Zhang Y., Xu Y., Liu X., Xiang X.** Boosting ship detection in SAR images with complementary pretraining techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, vol. 14, pp. 8941–8954. doi:10.1109/JSTARS.2021.3109002
14. **Павлов В. А., Белов А. А., Волвенко С. В., Рашич А. В.** Применение обученных на оптических изображениях сверточных нейронных сетей для обнаружения объектов на радиолокационных изображениях. *Компьютерная оптика*, 2024, т. 48, № 2, с. 253–259. doi:10.18287/2412-6179-CO-1316, EDN: RJDCTL
15. **Павлов В. А., Белов А. А.** Использование оптических изображений в обучении нейронных сетей, обрабатывающих радиолокационные изображения. *Радиотехника*, 2025, т. 89, № 3, с. 98–108. doi:https://doi.org/10.18127/j00338486-202503-09, EDN: XLSHRP
16. **Wei S., Zeng X., Qu Q., Wang M., Su H., Shi J.** HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access*, 2020, vol. 8, pp. 120234–120254. doi:10.1109/ACCESS.2020.3005861
17. **Верба В. С., Неронский Л. Б., Осипов И. Г., Турук В. Э.** *Радиолокационные системы землеобзора космического базирования*. М., Радиотехника, 2010. 680 с.
18. **Khanam R., Hussain M.** YOLOv11: An overview of the key architectural enhancements. *arXiv preprint*, 2024. doi:10.48550/arXiv.2410.17725
19. **Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A.** The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 2010, vol. 88, no. 2, pp. 303–338. doi:10.1007/s11263-009-0275-4
20. **Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L.** Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014: Proceedings of the 13th European Conference on Computer Vision*. Cham: Springer International Publishing, 2014, vol. 8693, pp. 740–755. doi:10.1007/978-3-319-10602-1_48

UDC 004.932.72
doi:10.31799/1684-8853-2026-3-14-22
EDN: ZWXEEH

Algorithm for preprocessing optical training data that improves the efficiency of their use in the tasks for neural network processing of radar images

V. A. Pavlov[a], PhD, Tech., Associate Professor, orcid.org/0000-0003-0726-6613, pavlov_va@spbstu.ru
A. A. Belov[a], Leading Engineer, orcid.org/0000-0003-0617-4514
F. Shariaty[a], Senior Lecturer, orcid.org/0000-0002-7060-8826
S. V. Volvenko[a], Senior Researcher, orcid.org/0000-0001-7726-8492
L. Gu[a], PhD, Tech., Research Engineer, orcid.org/0000-0003-1105-5835
[a]Peter the Great St. Petersburg Polytechnic University, 29, Politekhnicheskaia St., 195251, Saint-Petersburg, Russian Federation

Introduction: Although several studies have demonstrated in recent years the possibility of using optical images to train neural networks designed for processing synthetic aperture radar data, existing approaches still show limited efficiency. This is due to differences in the physical nature of image formation and, consequently, to substantial differences in the classification features extracted by a neural network during training. **Purpose:** To develop an algorithm for specialized preprocessing of optical images which can be used in cross-domain transfer learning of neural networks for radar image processing, and to evaluate the effectiveness of this preprocessing using standard object-detection metrics. **Results:** We develop an algorithm for specialized preprocessing of optical images. It includes grayscale conversion, low-frequency Gaussian filtering, boundary extraction using the Laplacian operator, and subsequent smoothing to form a pseudo-radar representation. In the experiments, YOLO11x has been trained on preprocessed optical images, validated on optical images, and tested on real radar images containing ships. Compared with training on raw optical images, the best preprocessing configuration increases Precision from 0.675 to 0.813, Recall from 0.464 to 0.517, mAP50 from 0.518 to 0.613, and mAP50-95 from 0.202 to 0.342. These results indicate that suppressing brightness and color information specific to the optical domain and emphasizing features related to object shape and boundaries improve transfer efficiency between the optical and radar domains under the considered experimental conditions. **Practical relevance:** The results of this work enable the use of widely accessible labeled optical images for training neural networks tasked with classifying, segmenting, or detecting objects in radar images..

Keywords — synthetic aperture radar, optical images, radar images, image processing, neural network, training, object detection.

For citation: Pavlov V. A., Belov A. A., Shariaty F., Volvenko S. V., Gu L. Algorithm for preprocessing optical training data that improves the efficiency of their use in neural network processing tasks of radar images. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 14–22 (In Russian). doi:10.31799/1684-8853-2026-3-14-22, EDN: ZWXEEH

References

1. Li K., Wan G., Cheng G. DIOR: A large-scale dataset for object detection in optical remote sensing images. *IEEE Dataport*, April 12, 2025. doi:10.21227/tq1e-nx82
2. Xia G. S., Bai X., Ding J., Zhu Z. DOTA: A large-scale dataset for object detection in aerial images. *IEEE Dataport*, April 12, 2025. doi:10.21227/wwrj-3d46
3. Jeong S., Kim Y., Kim S., Sohn K. Enriching SAR ship detection via multistage domain alignment. *IEEE Geoscience and Remote Sensing Letters*, 2021, vol. 19, pp. 1–5. doi:10.1109/LGRS.2021.3115498
4. Shi Y., Du L., Guo Y., Du Y. Unsupervised domain adaptation based on progressive transfer for ship detection: From optical to SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, pp. 1–17. doi:10.1109/TGRS.2022.3185298
5. Liu S., Wang Z., Wang R., Chen Y., Fan X., Li W. SAR ship detection based on deep domain adaptation with limited samples. *Procedia Computer Science*, 2023, vol. 221, pp. 378–385. doi:10.1016/j.procs.2023.07.051
6. Zhu Y., Ai J., Xue W., Wang Z., Zhao Z. Cross-modal ship detection from optical to SAR images based on pixel-level and feature-level progressive transfer. *IEEE Sensors Journal*, 2025, vol. 25, no. 8, pp. 13344–13356. doi:10.1109/JSEN.2025.3543520
7. Luo C., Zhang Y., Guo J., Hu Y., Zhou G., You H., Ning X. SAR-CDSS: A semi-supervised cross-domain object detection from optical to SAR domain. *Remote Sensing*, 2024, vol. 16, no. 6, p. 940. doi:10.3390/rs16060940
8. Wu B., Wang H., Zhang C., Chen J. Optical-to-SAR translation based on CDA-GAN for high-quality training sample generation for ship detection in SAR amplitude images. *Remote Sensing*, 2024, vol. 16, no. 16, p. 3001. doi:10.3390/rs16163001
9. Luo X., Lin Y., Xu K., Nam H., Peng D. *OS-Ship-1K: A Cycle-GAN-Based Optical-SAR Multimodal Ship Detection Dataset*. 2025. Available at: https://openreview.net/forum?id=aIB8eQc9Au (accessed 01 February 2026).
10. Huang Z., Zhang X., Tang Z., Xu F., Datcu M., Han J. Generative artificial intelligence meets synthetic aperture radar: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2024, vol. 14, no. 1, pp. 6–48. doi:10.1109/MGRS.2024.3483459
11. Wang Z., Zhang Z., Shan X., Wei H. A., Tang P. Generative models for SAR-optical image translation: A systematic review. *International Journal of Applied Earth Observation and Geoinformation*, 2026, vol. 146, Article 105009. doi:10.1016/j.jag.2025.105009
12. Lee S.-J., Lee K.-J. Efficient generation of artificial training DB for ship detection using satellite SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, vol. 14, pp. 11764–11774. doi:10.1109/JSTARS.2021.3128184
13. Bao W., Huang M., Zhang Y., Xu Y., Liu X., Xiang X. Boosting ship detection in SAR images with complementary pretraining techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, vol. 14, pp. 8941–8954. doi:10.1109/JSTARS.2021.3109002
14. Pavlov V. A., Belov A. A., Volvenko S. V., Rashich A. V. Application of convolutional neural networks trained on optical images for object detection in radar images. *Computer Optics*, 2024, vol. 48, no. 2, pp. 253–259 (In Russian). doi:10.18287/2412-6179-CO-1316, EDN: RJDCTL
15. Pavlov V. A., Belov A. A. On using optical images in training of neural networks for radar image processing. *Radioengineering*, 2025, vol. 89, no. 3, pp. 98–108 (In Russian). doi:10.18127/j00338486-202503-09, EDN: XLSHRP
16. Verba V. S., Neronskiy L. B., Osipov I. G., Turuk V. E. *Radiolokacionnye sistemy zemleobzora kosmicheskogo bazirovaniya* [Space-based radar systems for land survey]. Moscow, Radiotekhnika Publ., 2010. 680 p. (In Russian).
17. Wei S., Zeng X., Qu Q., Wang M., Su H., Shi J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access*, 2020, vol. 8, pp. 120234–120254. doi:10.1109/ACCESS.2020.3005861
18. Khanam R., Hussain M. YOLOv11: An overview of the key architectural enhancements. *arXiv preprint*, 2024. doi:10.48550/arXiv.2410.17725
19. Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 2010, vol. 88, no. 2, pp. 303–338. doi:10.1007/s11263-009-0275-4
20. Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L. *Microsoft COCO: Common Objects in Context*. In: *Computer Vision – ECCV 2014: Proceedings of the 13th European Conference on Computer Vision*. Cham, Springer International Publishing, 2014, vol. 8693, pp. 740–755. doi:10.1007/978-3-319-10602-1_48

УДК 004.853.3
doi:10.31799/1684-8853-2026-3-23-34
EDN: ZZVLVG

Научные статьи
Articles

Проектирование адаптивной архитектуры сверточной нейронной сети с использованием глубинных сепарабельных сверток для работы в реальном времени на встраиваемых устройствах

А. В. Сацюк[a], *канд. техн. наук, заведующий лабораторией, orcid.org/0009-0006-7228-8279, alexandrsatsuk@gmail.com*
С. А. Радковский[a], *канд. техн. наук, доцент, orcid.org/0009-0007-9411-949X*
А. И. Шеховцов[a], *канд. техн. наук, доцент, orcid.org/0009-0003-6963-815X*
С. Д. Сонина[a], *старший преподаватель, orcid.org/0009-0009-6114-4386*
А. А. Воробьев[a], *старший преподаватель, orcid.org/0009-0000-0938-6139*
[a]*Донецкий институт железнодорожного транспорта, Горная ул., 6, Донецк, Донецкая Народная Республика, 283018, РФ*

Введение: *развертывание современных систем компьютерного зрения на встраиваемых устройствах (например, Raspberry Pi 5) сталкивается с проблемой достижения реального времени (целевой порог 20 кадров/с) при анализе потока с MIPI-камеры из-за жестких ограничений ресурсов. Базовые однопроходные детекторы, такие как YOLOv8n, обладают избыточной вычислительной сложностью для таких условий.* ***Цель:*** *разработать адаптированную, легковесную архитектуру сверточной нейронной сети путем модификации YOLOv8n, сфокусировав вычислительные ресурсы на объектах среднего и крупного масштаба.* ***Результаты:*** *основное внимание уделено радикальному снижению FLOPs и параметров через замену стандартных сверток на глубинные сепарабельные свертки и исключение избыточных уровней детекции (сокращение выходных голов). Исследование проводилось на платформе Raspberry Pi 5 с входным разрешением 1280 × 720. Архитектурная реконфигурация привела к созданию модели OptiYOLO с минимальными показателями: 1,98 млн параметров и 4,1 GFLOPs. После квантования до INT8 достигнута частота обработки 28,7 кадров/с, что значительно превосходит требование реального времени. Общая точность (mAP0.5) составила 73,4 %, что признано приемлемым компромиссом. Точность для критических классов («автомобиль» — 86,1 %, «человек» — 82,5 %) подтвердила успешное перераспределение ресурсов на более крупные объекты.* ***Практическая значимость:*** *разработанная OptiYOLO предоставляет сбалансированное решение для автономных, энергоэффективных систем видеонаблюдения, где приоритетом является скорость обработки, а не детекция мелких объектов.*

Ключевые слова *— YOLO, глубинные сепарабельные свертки, Raspberry Pi 5, оптимизация нейронных сетей, реальное время, сверточная нейронная сеть, компьютерное зрение.*

Для цитирования: Сацюк А. В., Радковский С. А., Шеховцов А. И., Сонина С. Д., Воробьев А. А. Проектирование адаптивной архитектуры сверточной нейронной сети с использованием глубинных сепарабельных сверток для работы в реальном времени на встраиваемых устройствах. *Информационно-управляющие системы*, 2026, № 3, с. 23–34. doi:10.31799/1684-8853-2026-3-23-34, EDN: ZZVLVG

For citation: Satsiuk A. V., Radkovsky S. A., Shekhovtsov A. I., Sonina S. D., Vorobyov A. A. Designing an adaptive convolutional neural network architecture using depthwise separable convolutions for real-time operation on embedded devices. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 23–34 (In Russian). doi:10.31799/1684-8853-2026-3-23-34, EDN: ZZVLVG

Введение

Разработка эффективных систем компьютерного зрения для работы в режиме реального времени на устройствах с ограниченными вычислительными ресурсами представляет одну из наиболее актуальных задач современного искусственного интеллекта. Особую практическую значимость приобретают автономные системы видеомониторинга, способные детектировать и классифицировать потенциально опасные объекты (такие как несанкционированный транспорт, люди в запрещенных зонах или оставленные предметы) в охраняемом периметре [1–3]. Ключевым требованием к подобным системам является не только высокая точность распознавания, но и способность обрабатывать видеопоток с минимальной задержкой для своевременного оповещения оператора-диспетчера. Важным условием для данной задачи, подтвержденным предварительным анализом сцен с целевой MIPI-камеры, является тот факт, что целевые объекты детекции (например, человек, транспортное средство или крупная сумка) в среднем занимают в кадре область не менее 80 × 80 пикселей. Это условие позволяет проводить дальнейшую архитектурную оптимизацию, ориентируясь на детекцию объектов среднего и крупного масштаба.

Несмотря на выдающиеся успехи глубокого обучения, в частности архитектур сверточных нейронных сетей семейства YOLO (You Only Look Once), которые установили новые стандарты в задачах детекции объектов в реальном времени [4–8], их прямое применение на маломощных embedded-платформах, таких как Raspberry Pi, сопряжено с существенными трудностями. Эти модели часто требуют значительных вычислительных затрат и объема памяти, что делает их непригодными для длительной автономной работы в составе компактных систем, оснащенных ресурсоэффективными компонентами, например MIPI-камерами [9, 10]. В связи с этим значимыми направлениями исследований стали архитектурная оптимизация и сжатие нейронных сетей. Перспективным подходом является замена стандартных сверточных слоев на глубинные сепарабельные свертки (Depthwise Separable Convolutions, DSC), которые позволяют радикально снизить вычислительную сложность и количество параметров модели при незначительной потере точности, что детально исследовано в работах [11, 12]. Другим важным направлением является разработка и использование легковесных Backbone-архитектур (например, MobileNet, ShuffleNet), специально спроектированных для мобильных устройств [13–15]. Однако их интеграция в архитектуру однопроходной детекции, сохраняющую высокую точность локализации объектов, остается комплексной задачей.

Следовательно, существует острая необходимость в создании специализированных, оптимизированных архитектур нейронных сетей, наследующих принципы скоростной однопроходной детекции от моделей-доноров, но адаптированных под строгие аппаратные ограничения целевой платформы.

Основной задачей данного исследования является разработка архитектуры сверточной нейронной сети для системы мониторинга в реальном времени, способной детектировать и классифицировать ограниченный набор потенциально опасных объектов на видеопотоке с MIPI-камеры, установленной на платформе Raspberry Pi 5.

Предлагаемый подход основывается на модификации и адаптивном упрощении архитектурных принципов YOLO. Главное внимание уделяется оптимизации глубины и ширины сети, реконфигурации слоев и выбору эффективных операций, включая DSC и анализ чувствительности слоев к pruning-операциям [16], для достижения оптимального баланса между скоростью инференса и точностью.

Системные требования

Успешная реализация системы мониторинга в реальном времени на платформе Raspberry Pi 5 потребовала формулирования ключевых условий и целевых показателей. Выбор данной платформы обусловлен необходимостью найти оптимальный компромисс между доступной ценой и минимально необходимой производительностью для автономной обработки видео с требуемой скоростью [17–19], обеспечивая немедленное оповещение диспетчера при обнаружении угрозы.

Для корректного проектирования были определены следующие технические требования и характеристики:

1) аппаратная платформа: Raspberry Pi 5;

2) входные данные: видеопоток с MIPI-камеры с разрешением 1280 × 720 пикселей;

3) целевые объекты: четыре класса («автомобиль», «человек», «животное», «сумка») с минимальным размером объекта ~80 × 80 пикселей;

4) критерии эффективности: частота обработки не менее 20 кадров/с при сохранении среднего уровня точности (mAP0.5) не ниже 0,70;

5) оптимизационный базис: использование методологии адаптации модели-донора YOLOv8n с применением DSC для радикального снижения вычислительной сложности.

Необходимо было определить системные ограничения и характеристики целевой сцены для правильного формулирования требований к архитектуре нейронной сети. После проведенных предварительных экспериментов по балансировке между качеством изображения, вычислительной нагрузкой и пропускной способностью шины установлено оптимальное входное разрешение для нейронной сети 1280 × 720 пикселей. Данное разрешение является приемлемым с точки зрения информативности для задачи детекции и одновременно допустимым для обработки в реальном времени на целевом аппаратном обеспечении при использовании MIPI-камеры [20].

Целевыми объектами для детекции и классификации выбраны автомобили, люди, животные и сумки. Критическим системным допущением, основанным на планируемой физической конфигурации системы (высота и угол установки камеры, заданный периметр), является ограничение максимальной дальности наблюдения до 50 м.

При таких сценарии и разрешении кадра минимальный ожидаемый размер любого целевого объекта в пикселях составляет в среднем не менее 80 × 80 [1] (рис. 1). Это допущение позволяет существенно упростить задачу детекции, сфокусировав архитектурные оптимизации на работе с объектами среднего и крупного масштаба, и отказаться от детектирования крайне мелких объектов, что является одной из наиболее ресурсоемких подзадач [21–25].

■ ***Рис. 1.*** Размер объектов наблюдения на кадре с разрешением 1280 × 720 пикселей

■ ***Fig. 1.*** Observed object dimensions within the 1280 × 720 frame

Таким образом, задача формализуется как разработка легковесной однопроходной архитектуры детекции (one-stage detector), адаптированной под следующие ключевые условия:

1) аппаратная платформа: Raspberry Pi 5 (ограниченные ресурсы CPU/GPU, оперативной памяти);

2) входные данные: видеопоток с разрешением 1280 × 720 пикселей;

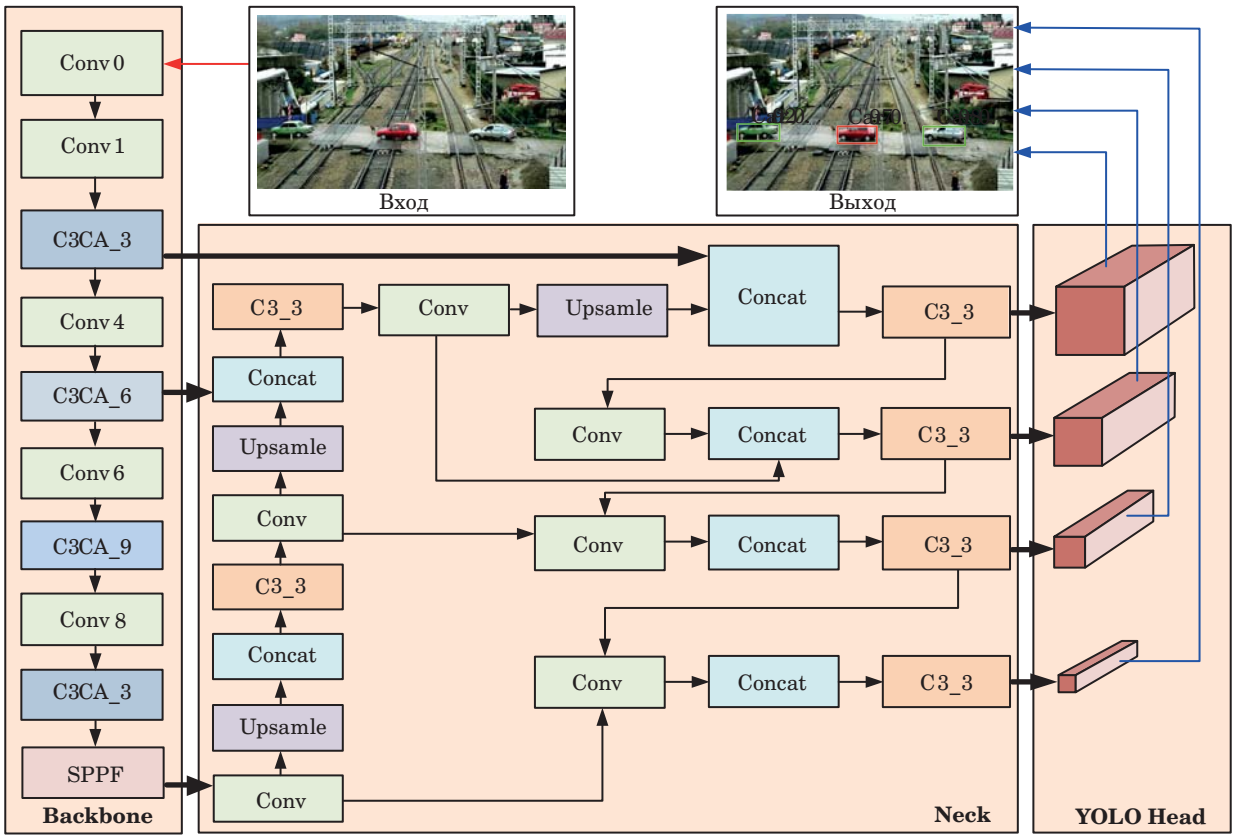
3) целевые объекты: четыре класса («автомобиль», «человек», «животное», «сумка») с минимальным ожидаемым размером ~80 × 80 пикселей;

4) критерий эффективности: частота обработки не менее 20 кадров/с для обеспечения работы в режиме, воспринимаемом как реальное время, при сохранении средней точности (mAP0.5) на уровне, достаточном для практического применения (не ниже 0,70);

5) методология: стратегическая адаптация и оптимизация модели-донора YOLOv8n, направленная на радикальное снижение вычислительной сложности при контролируемой потере точности, прежде всего на мелких объектах, исключенных из постановки задачи.

Общая концепция предлагаемой архитектуры

Предлагаемая архитектура нейронной сети, ориентированная на детекцию объектов в реальном времени на платформе Raspberry Pi 5, базируется на стратегии адаптации и оптимизации модели-донора YOLOv8n (рис. 2).



■ ***Рис. 2.*** Классическое представление архитектуры YOLO

■ ***Fig. 2.*** Standard representation of the YOLO architecture

Данная модель была выбрана в качестве исходного прототипа благодаря ее сбалансированному соотношению скорости и точности в рамках семейства YOLO, а также изначально заложенной ориентации на применение в ресурсоограниченных средах [26]. Однако, как показано выше, прямое использование базовой модели для целевого сценария классификации и детекции с разрешением 1280 × 720 и частотой не менее 20 кадров/с на Raspberry Pi 5 невозможно. Даже эта «легковесная» версия (YOLOv8n, YOLOv8s) не может быть напрямую развернута на целевом оборудовании без существенных доработок, направленных на достижение необходимой частоты кадров при работе с потоком данных от MIPI-камеры с сохранением точности на заданных классах объектов.

Принципом предлагаемой адаптации является учет специфики целевой сцены, в частности отсутствия необходимости детектировать объекты размером менее 80 × 80 пикселей. Это позволяет провести целенаправленную редукцию архитектуры, фокусируясь на эффективном извлечении признаков для объектов среднего и крупного масштаба.

Главная идея разработки заключается в последовательном и целенаправленном упрощении архитектуры-донора при сохранении ее ключевых функциональных преимуществ: однопроходной схемы детекции и эффективной многоуровневой пирамиды признаков. Упрощение реализуется по трем основным направлениям.

Во-первых, проводится модификация и редукция Backbone-сети (хребта архитектуры), ответственной за извлечение признаков. Стандартные сверточные блоки C2f в YOLOv8n, несмотря на их эффективность за счет обильных остаточных связей, содержат значительное количество стандартных 3 × 3 сверток и сложную ветвистую структуру, что делает их вычислительно дорогими для целевой платформы. Вместо них предлагается использовать оптимизированные блоки, основанные на DSC [11] и линейных bottleneck-структурах, что позволяет радикально сократить количество параметров и операций умножения-сложения. Глубина Backbone целенаправленно уменьшается для снижения латентности, при этом для компенсации потенциальной потери контекста усиливается роль механизмов агрегации признаков на разных уровнях. Учитывая, что входное изображение имеет фиксированное разрешение 1280 × 720, а целевые объекты не считаются экстремально мелкими, изначальная глубокая пирамида признаков донорской модели является избыточной. В этом случае сокращение числа этапов пространственного сжатия в Backbone позволяет сохранить более детальную информацию на финальных картах признаков, что критически важно для локализации объектов заданного размера.

Во-вторых, модификации подвергается секция агрегации признаков (Neck). Стандартная архитектура Neck в YOLOv8n, основанная на принципах FPN+PAN, включает сложные блоки C2f и множество операций upsampling, что создает значительную нагрузку на память и вычислительные ядра. В предлагаемой модели структура агрегации признаков оптимизирована: количество каналов в узлах feature fusion сокращено, а вычислительно затратные блоки заменены на облегченные skip-connections с конкатенацией признаков [27–29]. Это обеспечивает эффективную передачу семантической и пространственной информации при существенно меньших вычислительных затратах. Кроме того, на основе анализа размеров целевых объектов один из верхних уровней пирамиды, ответственный за детекцию самых мелких объектов, исключен, что упрощает топологию сети и сокращает количество предсказательных голов.

В-третьих, Head-часть (голова), выполняющая окончательное прогнозирование ограничивающих рамок, классов объектов и коэффициентов уверенности, подвергается двунаправленной структурной оптимизации:

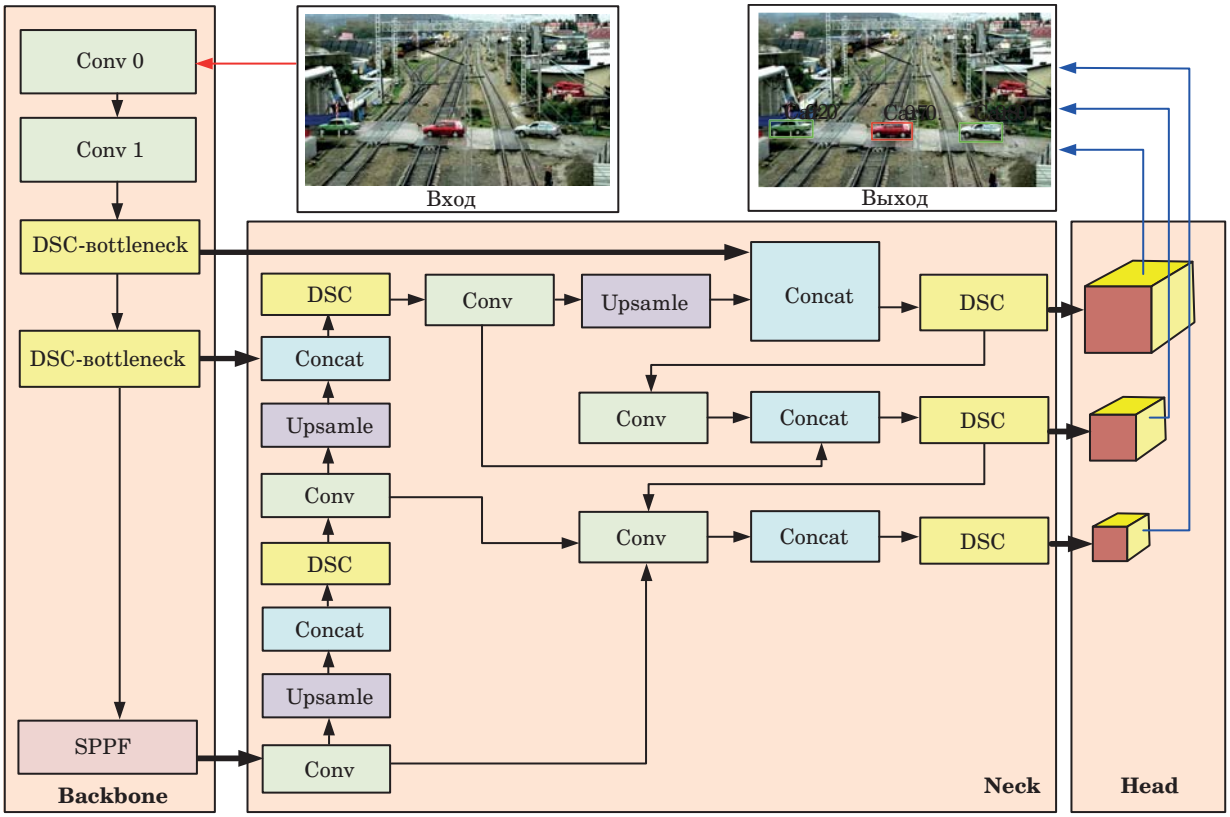
1) удаляется самый нижний уровень детекции (самая мелкая голова). Поскольку постановка задачи исключает необходимость детекции объектов размером менее 80 × 80 пикселей, сокращение числа голов с четырех до трех является обоснованным, обеспечивая максимальную экономию ресурсов на финальной стадии, где высокая точность для мелких объектов не является приоритетной;

2) финальный слой, проецирующий агрегированные признаки на выходные тензоры, также реализуется с использованием DSC, что гарантирует минимизацию потребления ресурсов Raspberry Pi 5 даже на стадии генерации предсказаний.

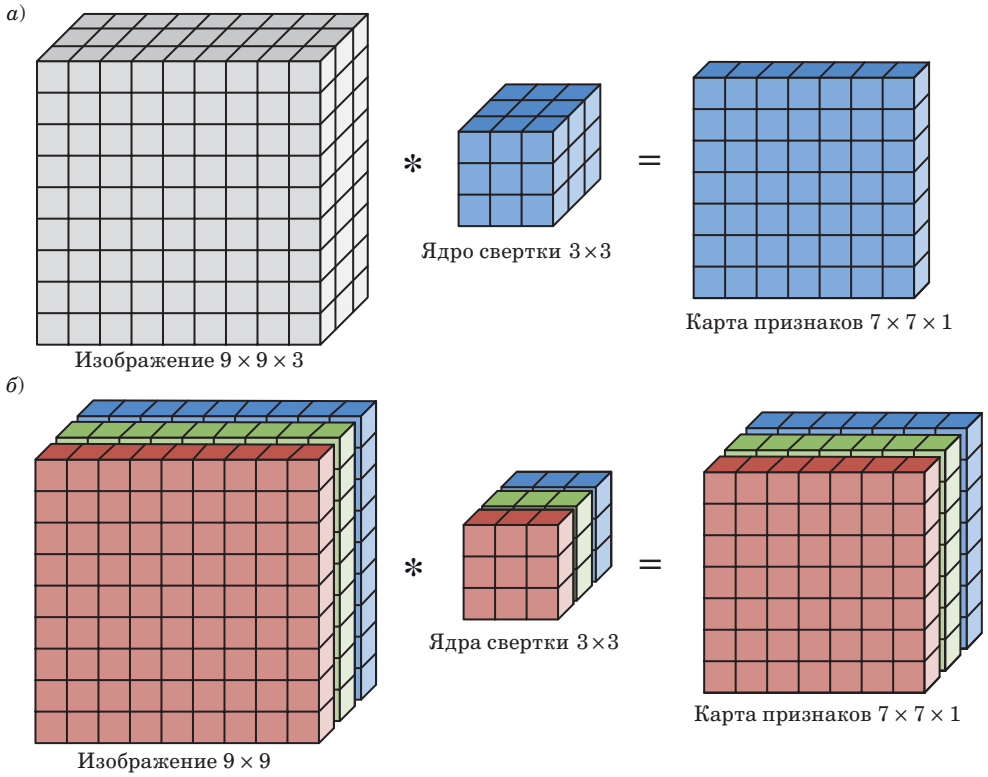
Общая схема предлагаемой архитектуры, наглядно демонстрирующая эволюцию от модели-донора к конечному дизайну, представлена на рис. 3. Желтым цветом выделены узлы структуры, которые были добавлены или заменены.

Визуальное сравнение двух архитектур (см. рис. 2 и 3) позволяет оценить сокращение сложности при сохранении ключевых модулей. Детализированная структура предложенного легковесного блока на основе сепарабельных сверток, который является строительным элементом модифицированных Backbone и Neck, показана на рис. 4.

Этот блок сочетает глубинные свертки для пространственного фильтрования и точечные свертки (1 × 1) для комбинирования каналов, что обеспечивает существенный выигрыш в производительности по сравнению со стандартными сверточными слоями.



■ ***Рис. 3.*** Предлагаемая архитектура сверточной нейронной сети
■ ***Fig. 3.*** Proposed CNN architecture



■ ***Рис. 4.*** Применения ядра свертки 3 × 3 с использованием классической сверточной обработки (*а*) и DSC-обработки (*б*)
■ ***Fig. 4.*** Application of a 3 × 3 convolution kernel using classical convolution processing (*а*) and DSC processing (*б*)

Разработка целевой архитектуры (см. рис. 3) сфокусирована на замене стандартных, вычислительно затратных слоев на их ресурсоэффективные аналоги, что обеспечивает достижение требуемой частоты кадров при обработке потока с разрешением 1280 × 780.

Блочная модификация архитектуры

Основной вклад в снижение вычислительной нагрузки заложен в оптимизации Backbone-сети. В донорской архитектуре (см. рис. 2) блоки C3CA_x и их аналоги (например, C2f) являются основным потребителем FLOPs. В предлагаемой архитектуре эти блоки полностью заменены на DSC-bottleneck.

Блок C3CA_x основан на стандартных свертках, которые выполняют пространственное сканирование и слияние каналов в одном шаге. Блок DSC-bottleneck реализует принцип DSC, разделяя эти две операции [11].

Рассмотрим стандартную свертку с входным и выходным количеством каналов C_{in} и C_{out} и ядром размера $K \times K$ на входе с размером карты признаков $H \times W$.

В этом случае сложность стандартной свертки (в блоке C3CA_x)

$$FLOPs_{std} = H \cdot W \cdot C_{in} \cdot C_{out} \cdot K^2,$$

а сложность DSC

$$FLOPs_{DSC} = H \cdot W \cdot (K^2 \cdot C_{in} + C_{in} \cdot C_{out}),$$

где первое слагаемое — глубинная свертка, второе — точечная свертка 1×1.

При условии, что K^2 значительно больше 1 (например, $3^2 = 9$), а C_{out} относительно велико, сокращение FLOPs по сравнению со стандартной сверткой приближается к сокращению вычислительной сложности в K^2 раз, сохраняя при этом эффективность комбинирования каналов благодаря точечной свертке [26, 30].

Так, например, для типичных параметров, встречающихся в блоках C3CA, когда $C_{in} = 64$, $C_{out} = 128$ и ядро $K = 3$, сложность стандартной свертки составляет порядка 737 280 операций. В то же время DSC требует всего ≈8768 операций с учетом точечной свертки. Таким образом, замена стандартных сверточных слоев на DSC-блоки обеспечивает локальное сокращение вычислительной сложности в ≈84,1 раза.

Для дальнейшей оптимизации вычислительной сложности предложенный блок (DSC-bottleneck) реализует архитектуру сжатия признаков [31]. Данный подход предполагает предварительное сокращение размерности (числа каналов) перед выполнением операции глубинной свертки. Это позволяет эффективно сжать данные и существенно снизить объем вычислительных операций при сохранении репрезентативности признаков.

Аналогичный принцип оптимизации применен и в секции Neck, отвечающей за мультимасштабную агрегацию (PANet-подобная структура), где в базовой модели использованы блоки C3_3 (сверточный блок с тремя сверхточными слоями) после операции слияния (Concat). Эти блоки выполняют как пространственную фильтрацию, так и смешивание каналов.

В модифицированной архитектуре (см. рис. 3) блоки C3_3 заменены на блоки, реализующие DSC (желтые блоки на выходе каждой ветви Neck). Эта замена направлена на снижение вычислительных требований в процессе обработки признаков, извлекаемых с разных уровней детализации. Использование одиночного DSC-слоя вместо полного блока C3_3 (который сам может содержать несколько стандартных сверток) обеспечивает линейное или сублинейное снижение FLOPs на данном этапе.

Ключевым архитектурным решением стало сокращение числа выходных голов в секции Head с четырех до трех (см. рис. 3). Удаление самой мелкой выходной головы, отвечающей за детекцию наименьших объектов, архитектурно обосновано требованием задачи исключать объекты размером менее 80 × 80 пикселей. Это делает соответствующую шкалу детекции избыточной и, как следствие, устраняет необходимость в самом нижнем пути агрегации в Neck. В результате удаляется целый каскад сверточных и объединяющих слоев, что обеспечивает значительный выигрыш в общей вычислительной сложности и позволяет сфокусировать вычислительные ресурсы на объектах среднего и крупного масштаба.

Теоретически сокращение числа голов с четырех до трех приводит к прямому снижению количества параметров и FLOPs в выходном блоке на величину, пропорциональную полному вкладу удаленной головы H_{small}:

$$K_{FLOPs} = \frac{FLOPs(H_{small})}{FLOPs(H_{all})} \cdot 100,$$

где K_{FLOPs} — выигрыш; $FLOPs(H_{small})$ — вычислительная сложность в канале малой головы; $FLOPs(H_{all})$ — вычислительная сложность всей модели.

По результатам расчетов в исходной архитектуре (с четырьмя головами) самая мелкая голова вносила вклад (при параметрах $H = 16$, $W = 16$, $C_{in} = 512$ для четырех классов) 0,85 GFLOPs при общем объеме 14,2 GFLOPs. Таким образом, исключение этого компонента обеспечило теоретическое снижение вычислительной сложности на 5,981 %, что в сочетании с заменой блоков на DSC позволило достичь роста итогового ускорения в 1,56 раза.

Дополнительно финальный слой в секции Head, проецирующий агрегированные признаки на выходные тензоры, был заменен на блоки, реализующие DSC, как и в других секциях. Это гарантирует, что даже на стадии постобработки и генерации предсказаний минимизируется потребление ресурсов целевой платформы Raspberry Pi 5.

Экспериментальная часть: проведение и анализ результатов

Для валидации разработанной легковесной архитектуры OptiYOLO были проведены сравнительные испытания на целевом оборудовании — одноплатном компьютере Raspberry Pi 5 с оперативной памятью 8 ГБ, на видеопотоке от MIPI-камеры с разрешением 1280 × 720.

Обучение и валидация проводились на модифицированном наборе данных, включающем расширенные подмножества COCO [32], дополненные сценами, релевантными для задач периметрального мониторинга (транспорт на железнодорожных путях, объекты в зонах ограниченного доступа). В датасет входят четыре целевых класса: «автомобиль», «человек», «животное», «сумка». В программном коде и при визуализации результатов детекции используются англоязычные обозначения (car, person, animal, bag). Общий объем датасета составил примерно 6500 изображений на каждый класс. Для повышения робастности моделей применялись методы аугментации, сфокусированные на изменении масштаба и освещенности. Обучение проводилось с использованием оптимизатора Adam (скорость обучения 0,001), размером батча 32 в течение 150 эпох. После обучения веса были подвергнуты квантованию до формата INT8 для оптимизации развертывания на встраиваемых системах [33–35].

В качестве базовых моделей-доноров использовались официальные, оптимизированные веса YOLOv8n, YOLOv8s, а также более крупная версия YOLOv8m для оценки того, насколько эффективно редукция обходит необходимость использования более мощных моделей.

Для достижения максимальной скорости на Raspberry Pi 5 все модели (базовые и OptiYOLO) были конвертированы в целочисленный формат INT8. Стоит заметить, что процесс квантования оказал минимальное негативное влияние на общую точность OptiYOLO, снизив mAP0.5 всего на 0,2 пункта (с 73,6 до 73,4 %), что значительно меньше, чем наблюдалось у более сложной YOLOv8s (потеря 1,1 пункта). Этот результат подтверждает большую устойчивость к сжатию архитектуры OptiYOLO, разработанной с акцентом на минимизацию числа каналов и параметров, что благоприятствует эффективному целочисленному инференсу на нейронных сопроцессорах (NPU) или оптимизированных ядрах RPi 5.

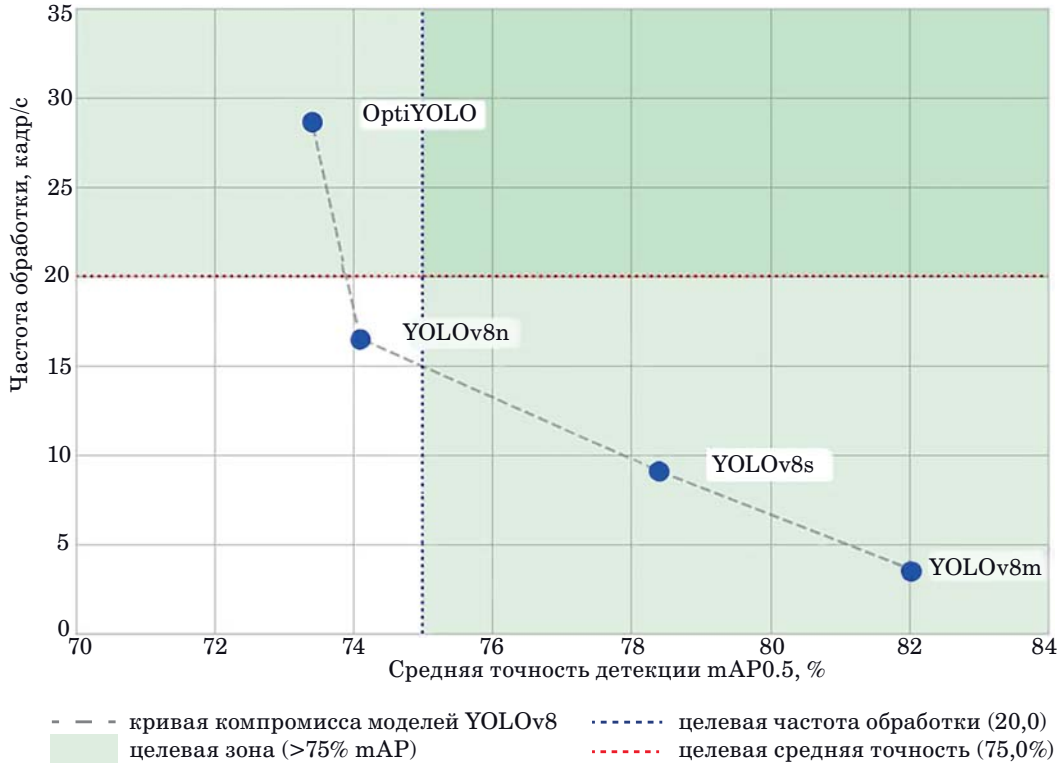
Оценка производительности проводилась по критериям, заданным при постановке задачи: точность детекции (mAP0.5) и частота обработки при инференсе на Raspberry Pi 5. Особое внимание уделялось количеству параметров как ключевому индикатору ресурсоемкости модели. Измерения частоты обработки проводились путем усреднения результатов 500 последовательных кадров.

Результаты тестирования OptiYOLO сопоставлены с тремя базовыми моделями YOLOv8 при инференсе в условиях, максимально приближенных к реальным. Компромисс между скоростью и точностью наглядно иллюстрирует рис. 5.

Разработанная архитектура OptiYOLO демонстрирует уникальное позиционирование: она достигает скорости, значительно превышающей требуемый порог 20 кадров/с, сохраняя при этом точность, сопоставимую с самой легковесной базовой моделью YOLOv8n. Основные количественные показатели производительности моделей, протестированных с входным разрешением 1280 × 720 и после квантования в формат INT8 [36], представлены в таблице.

■ Сравнительные характеристики моделей на Raspberry Pi 5
■ Comparative Model Characteristics on Raspberry Pi 5

Модель (INT8)	Параметры, млн	GFLOPs	Частота обработки, кадр/с	mAP0.5, %			
				общая	car	person	bag
YOLOv8m	25,9	88,6	3,5±0,1	82,1	91,5	87,8	65,4
YOLOv8s	7,22	15,8	9,1±0,3	78,4	88,9	85,1	61,2
YOLOv8n	3,25	8,7	16,5±0,5	74,1	85,2	81,0	55,9
OptiYOLO	1,98	4,1	28,7±0,7	73,4	86,1	82,5	56,1



■ ***Рис. 5.*** Зависимость частоты обработки от mAP0.5
■ ***Fig. 5.*** Comparative plot of processing frequency vs. mAP0.5

■ ***Рис. 6.*** Примеры детекции OptiYOLO на реальных кадрах
■ ***Fig. 6.*** OptiYOLO detection examples on real-world frames

Качественные результаты работы разработанной модели OptiYOLO на тестовых сценах, включая детекцию объектов на разных дистанциях, представлены на рис. 6. Как видно из рисунка, модель OptiYOLO устойчиво и с достаточной уверенностью классифицирует объекты (в данном примере класс «car») в зоне контроля.

Проведенный эксперимент убедительно демонстрирует преимущества архитектурного упрощения, основанного на специфике задачи.

1. Модель OptiYOLO достигла частоты 28,7 кадра/с, что обеспечивает работу в режиме реального времени с существенным запасом производительности на Raspberry Pi 5. При этом она является самой легковесной по числу параметров (1,98 млн против 3,2 млн у YOLOv8n) и вычислительной сложности (4,1 GFLOPs против 8,7 GFLOPs у YOLOv8n).

2. Несмотря на общее снижение mAP0.5 на 0,7 единицы относительно YOLOv8n, точность для классов «автомобиль» и «человек» демонстрирует улучшение (например, «автомобиль» — 86,1). Это подтверждает, что реконфигурация архитектуры, направленная на сохранение признаков среднего и крупного масштаба, привела к перераспределению точности в пользу наиболее критичных для безопасности классов.

3. Особо следует отметить, что при валидации на сценах с удалением объектов до 60 м (что соответствует границе целевого периметра) точность для классов «автомобиль» (86,1 %) и «человек» (82,5 %) остается на высоком уровне. Это доказывает, что прореживание в секции Backbone оказалось эффективным для сохранения пространственной информации, критичной для дальней детекции.

4. Модель YOLOv8m показывает наилучшую общую точность, однако требует значительно больших вычислительных ресурсов. По сравнению с OptiYOLO, она имеет в 13 раз больше параметров (25,9 млн против 1,98 млн) и в 21,6 раза выше вычислительную сложность (88,6 GFLOPs против 4,1 GFLOPs), что делает ее неприменимой для автономного использования на Raspberry Pi 5. Напротив, предлагаемая модель OptiYOLO предоставляет сбалансированное решение, недостижимое для стандартных версий YOLO.

Результаты подтверждают, что разработанная OptiYOLO успешно решает поставленную задачу, обеспечивая требуемую производительность при минимальных аппаратных затратах.

Заключение

Проведенное исследование позволило разработать и валидировать легковесную однопроходную архитектуру сверточной нейронной сети, получившую наименование OptiYOLO, специально адаптированную для задач детекции объектов в реальном времени на ресурсоограниченной платформе Raspberry Pi 5. Обоснованная стратегия оптимизации, базирующаяся на замене стандартных сверточных блоков на DSC и структурном упрощении пирамиды признаков, позволила радикально снизить вычислительную сложность модели.

Ключевым архитектурным решением стало удаление головы детекции, ориентированной на наименьшие объекты (менее 80 × 80 пикселей), что оказалось возможным благодаря специфике постановки задачи по периметральному мониторингу. Этот шаг в сочетании с заменой блоков C3CA_x и C3_3 на DSC-bottleneck привел к созданию самой ресурсоэффективной модели среди протестированных: OptiYOLO требует всего 1,98 млн параметров и 4,1 GFLOPs, что означает снижение числа параметров на 39 % и операций FLOPs на 53 % по сравнению с базовой YOLOv8n.

Экспериментальные результаты, полученные при инференсе на аппаратной платформе с использованием видеопотока 1280 × 720 после квантования до INT8, свидетельствуют о достижении поставленных целей. Модель OptiYOLO демонстрирует частоту обработки 28,7 кадров/с, что существенно превышает минимально требуемый порог 20 кадров/с, обеспечивая функционирование системы в режиме, воспринимаемом как реальное время. При этом наблюдаемое снижение общей точности (mAP0.5 на уровне 73,4 %) является приемлемым компромиссом. Следует отметить сохранение высокой точности для критически важных классов — «автомобиль» (86,1 %) и «человек» (82,5 %), что подтверждает эффективность сохранения пространственной информации на средних и крупных масштабах.

С разработанным подходом целенаправленная архитектурная модификация, сфокусированная на устранении избыточной сложности, связанной с детектированием мелких объектов, позволяет эффективно перенести передовые методы однопроходной детекции на бюджетные встраиваемые системы, открывая перспективы для их широкого практического применения в системах безопасности и автоматизации.

Финансовая поддержка

Исследование выполнено в рамках государственного задания EKTT-2026-0001 за счет средств федерального бюджета.

Литература

1. **Сацюк А. В.** Мониторинг инфраструктуры на основе искусственного интеллекта. *Автоматика, связь, информатика*, 2025, № 9, с. 32–34. doi:10.62994/AT.2025.9.9.005, EDN: OXBPFP
2. **Сацюк А. В., Воевода Е. Г., Каминский Р. Я.** Разработка алгоритма поиска заданного объекта в видеопотоке реального времени на основе метода корреляционного поиска. *Сборник научных трудов Донецкого института железнодорожного транспорта*, 2025, № 1(76), с. 15–23. EDN: VKNJZZ
3. **Сацюк А. В., Воевода Е. Г.** Система автоматического контроля безопасности на железнодорожных переездах. *Сборник научных трудов Донецкого института железнодорожного транспорта*, 2024, № 2(73), с. 39–45. EDN: LDLPWU
4. **Васильев М. Е., Шалимов А. С., Савина О. А.** Обзор версий YOLO: одноэтапная модель сверточной нейронной сети. *Universum: технические науки*, 2025, № 6-1(135), с. 36–46. EDN: HLDKKY
5. **Redmon J., Farhadi A**. YOLOv3: An incremental improvement. *arXiv preprint*. arXiv:1804.02767. 2018.
6. **Zeng Y., Guo D. J., He W. K., Zhang T., Liu Z. T.** ARF-YOLOv8: A novel real-time object detection model for UAV-captured images detection. *J Real-Time Image Proc*, 2024, vol. 21, Article 107. https://doi.org/10.1007/s11554-024-01483-z
7. **Bochkovskiy A., Wang C. Y., Liao H. Y. M.** YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint*. arXiv:2004.10934. 2020.
8. **Клековкин В. А., Марков Н. Г., Небаба С. Г.** Модели сверточных нейронных сетей YOLO с механизмом внимания для систем компьютерного зрения реального времени. *Вестник Томского государственного университета. Управление, вычислительная техника и информатика*, 2025, № 72, с. 39–50. doi:10.17223/19988605/72/4, EDN: KEWUSH
9. **Егунов В. А., Королева И. Ю., Типаев Д. В.** Алгоритмы движения мобильного робота с построением карты местности в реальном времени. *Инженерный вестник Дона*, 2022, № 4(88), с. 126–132. EDN: YPFPSM. http://www.ivdon.ru (дата обращения: 10.01.2026).
10. **Павленко Д. А., Ковалев В. А., Снежко Э. В., Левчук В. А., Печковский В. И.** Распознавание подстилающей поверхности Земли с помощью сверточной нейронной сети на одноплатном микрокомпьютере. *Информатика*, 2020, т. 17, № 3, с. 36–43. doi:10.37661/1816-0301-2020-17-3-36-43, EDN: DIUEIM
11. **Сацюк А. В., Володарец Н. В.** Модификация модели YOLO для гибридной системы детекции и трекинга в БПЛА с автоматическим наведением. *Информационно-управляющие системы*, 2025, № 4, с. 36–44. doi:10.31799/1684-8853-2025-4-36-44, EDN: YKQVJU
12. **Лимонова Е. Е., Шешкус А. В., Николаев Д. П., Иванова А. А., Ильин Д. А., Арлазаров В. Л.** Оптимизация быстродействия первых слоев глубоких сверточных нейронных сетей. *Вестник Российского фонда фундаментальных исследований*, 2016, № 4, с. 84–96. doi:10.22204/2410-4639-2016-092-04-84-96, EDN: YOSCDX
13. **Qin D., Leichner C., Delakis M., Fornoni M., Luo S., Yang F., Wang W., Banbury C., Ye C., Akin B., Aggarwal V., Zhu T., Moro D., Howard A.** MobileNetV4: Universal models for the mobile ecosystem. *European Conference on Computer Vision*, Cham, Springer Nature Switzerland, 2024, pp. 78–96. doi:10.1007/978-3-031-73661-2_5
14. **Zhang X., Zhou X., Lin M., Sun J.** Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856. doi:10.1109/CVPR.2018.00716
15. **Tan M., Le Q.** Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (PMLR)*, 2019, pp. 6105–6114. doi:10.48550/arXiv.1905.11946
16. **Han S., Pool J., Tran J., Dally W. J.** Learning both weights and connections for efficient neural network. arXiv. 2015. https://doi.org/10.48550/ARXIV.1506.02626
17. **Chen C., Wang T., Liu C., Liu Y., Cheng L.** Lightweight convolutional transformers enhanced meta-learning for compound fault diagnosis of industrial robot. *IEEE Transactions on Instrumentation and Measurement*, 2023, vol. 72, pp. 1–12. doi:10.1109/TIM.2023.3277956
18. **Фролов С. В., Коробов А. А., Савинова К. С., Потлов А. Ю.** Нейросетевая система управления микроклиматом и освещенностью в неонатальном инкубаторе. *Датчики и системы*, 2025, № 1(279), с. 62–68. doi:10.24412/1992-7185-2025-1-62-68, EDN: IIZLFV
19. **Аксенов Д. С., Жиляев В. А., Маркин Н. И., Титов И. А.** Система распознавания объектов на базе Raspberry Pi 4 и Intel Neural Computer Stick 2. *Информационные системы и технологии*, 2023, № 4(138), с. 10–16. EDN: FHOYQH
20. **Сацюк А. В., Белый Р. В., Ищенко А. Е.** Оценка эффективности алгоритмов YOLO для обнаружения объектов в реальном времени во встраиваемых системах беспилотных транспортных средств. *Сборник научных трудов Донецкого института железнодорожного транспорта*, 2024, № 4(75), с. 73–82. EDN:OCNLZZ
21. **Burggraaff O., Schmidt N., Zamorano J., Pauly K., Pascual S., Tapia C., Spyrakos E., Snik F.** Standardized spectral and radiometric calibration of consumer cameras. *Optics Express*, 2019, vol. 27, no. 14, pp. 19075–19101. doi:10.1364/OE.27.019075
22. **Lin T. Y., Dollar Pi., Girshick R., He K., Hariharan B., Belongie S.** Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125. doi:10.48550/arXiv.1612.03144
23. **Kisantal M., Wojna Z., Murawski J., Naruniecj J., Cho K.** Augmentation for small object detection. *arXiv preprint*. arXiv:1902.07296. 2019.
24. **Banakh T., Głąb S., Jabłońska E., Swaczyna J.** Haar I sets: Looking at small sets in Polish groups through compact glasses. *arXiv preprint*. arXiv:1803.06712. 2018.
25. **Wang A., Chen H., Liu L., Chen K., Lin Z., Han J., Ding G.** YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 107984–108011. doi:10.48550/arXiv.2405.14458
26. **Howard A. G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H.** Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*. arXiv:1704.04861. 2017.
27. **Liu S., Qi L., Qin H., Shi J., Jia J.** Path aggregation network for instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759-8768. doi:10.1109/CVPR.2018.00913

28. **Zhang Z., Li X., Yang Y.** Scaling YOLOv8 for high-precision real-time object detection. *arXiv preprint*. arXiv:2408.15857. 2024.
29. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.49
30. **Chollet F.** Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258. doi:10.1109/CVPR.2017.195
31. **Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C.** Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474
32. **Desai M., Mewada H., Pires I. M., Roy S.** Evaluating the performance of the YOLO object detection framework on COCO dataset and real-world scenarios. *Procedia Computer Science*, 2024, vol. 251, pp. 157–163. doi:10.1016/j.procs.2024.11.096
33. **Сацюк А. В.** Оптимизация архитектуры YOLOv8 для задач захвата объекта БПЛА: анализ компромисса между точностью, скоростью и вычислительными ресурсами. *Вестник Ростовского государственного университета путей сообщения*, 2025, № 2(98), с. 35–42. doi:10.46973/0201-727X_2025_2_35, EDN: LCAOOM
34. **Jacob B., Kligys S., Chen B., Zhu M., Tang M., Howard A., Adam H., Kalenichenko D.** Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713. doi:10.1109/CVPR.2018.00286
35. **Nagel M., Nagel M., Amjad R., Fournarakis M., Bondarenko Y., Baalen M., Blankevoort T.** A white paper on neural network quantization. *arXiv preprint*. arXiv:2106.08295. 2021.
36. **Cherepanov N. I., Stepina N. O., Nikiforov I. V.** Improving image analysis and processing performance on the RISC-V platform with Lichee Pi 4A. *Proceedings of the Institute for System Programming of the RAS*, 2025, vol. 37, no. 5, pp. 157–172. doi:10.15514/ISPRAS-2025-37(5)-12, EDN: UDVIEH

UDC 004.853.3
doi:10.31799/1684-8853-2026-3-23-34
EDN: ZZVLVG

Designing an adaptive convolutional neural network architecture using depthwise separable convolutions for real-time operation on embedded devices

A. V. Satsiuk[a], PhD, Tech., Acting Head of Labs, orcid.org/0009-0006-7228-8279, alexandrsatsuk@gmail.com
S. A. Radkovsky[a], PhD, Tech., Associate Professor, orcid.org/0009-0007-9411-949X
A. I. Shekhovtsov[a], PhD, Tech., Associate Professor, orcid.org/0009-0003-6963-815X
S. D. Sonina[a], Senior Lecturer, orcid.org/0009-0009-6114-4386
A. A. Vorobyov[a], Senior Lecturer, orcid.org/0009-0000-0938-6139
[a]Donetsk Institute of Railway Transport, 6, Gornaya St., 283018, Donetsk, Russian Federation

Introduction: Deploying modern computer vision systems on embedded devices (such as Raspberry Pi 5) faces the challenge of achieving real-time performance (target threshold of 20 FPS) when analyzing the stream from a MIPI camera, due to severe resource constraints. Baseline single-stage detectors, like YOLOv8n, possess excessive computational complexity for these conditions. **Purpose:** To develop an adapted, lightweight convolutional neural network architecture by modifying YOLOv8n, focusing computational resources primarily on medium and large-scale objects. **Results:** We place the main focus on the radical reduction of FLOPs and parameters through the replacement of standard convolutions with depthwise separable convolutions and the elimination of redundant detection levels (reduction of output heads). The study has been conducted on the Raspberry Pi 5 platform with an input resolution of 1280×720. This architectural reconfiguration has led to the OptiYOLO model achieving minimal metrics: 1.98 million parameters and 4.1 GFLOPs. After INT8 quantization, a processing speed of 28.7 FPS has been achieved, significantly exceeding the real-time requirement. The overall accuracy has reached 73.4% (mAP0.5), which is considered an acceptable trade-off. The precision for critical classes ("car" – 86.1%, "person" – 82.5%) confirms the successful redistribution of resources towards larger objects. **Practical relevance:** The developed OptiYOLO provides a balanced solution for autonomous, energy-efficient video surveillance systems where processing speed, rather than the detection of small objects, is the priority.

Keywords — YOLO, depthwise separable convolutions, Raspberry Pi 5, neural network optimization, real-time, convolutional neural network, computer vision.

For citation: Satsiuk A. V., Radkovsky S. A., Shekhovtsov A. I., Sonina S. D., Vorobyov A. A. Designing an adaptive convolutional neural network architecture using depthwise separable convolutions for real-time operation on embedded devices. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 23–34 (In Russian). doi:10.31799/1684-8853-2026-3-23-34, EDN: ZZVLVG

Financial support
The study was carried out within the framework of state assignment EKTT-2026-0001 at the expense of the federal budget.

References

1. Satsiuk A. V. Automated monitoring system for railway infrastructure based on artificial intelligence. *Automation, Communications, Informatics*, 2025, no. 9, pp. 32–34 (In Russian). doi:10.62994/AT.2025.9.9.005, EDN: OXBPFP
2. Satsuk A. V., Voevoda E. G., Kaminskiy R. Ya. Development of an algorithm for searching a specified object in a real-time video stream based on the correlation search method. *Sbornik nauchnykh trudov Donetckogo instituta zheleznodorozhnogo transporta*, 2025, no. 1(76), pp. 15–23 (In Russian). EDN: VKNJZZ
3. Satsuk A. V., Voevoda E. G. Automated safety control system at railway crossings. *Sbornik nauchnykh trudov Donetckogo instituta zheleznodorozhnogo transporta*, 2024, no. 2(73), pp. 39–45 (In Russian). EDN: LDLPWU
4. Vasilyev M. E., Shalimov A. S., Savina O. A. Overview of YOLO versions: Single-stage convolutional neural network model. *Universum: tekhnicheskie nauki*, 2025, no. 6-1(135), pp. 36–46 (In Russian). EDN: HLDKKY
5. Redmon J., Farhadi A. YOLOv3: An incremental improvement. *arXiv preprint*. arXiv:1804.02767. 2018.
6. Zeng Y., Guo D. J., He W. K., Zhang T., Liu Z. T. ARF-YOLOv8: A novel real-time object detection model for UAV-captured images detection. *J Real-Time Image Proc*, 2024, vol. 21, Article 107. https://doi.org/10.1007/s11554-024-01483-z
7. Bochkovskiy A., Wang C. Y., Liao H. Y. M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint*. arXiv:2004.10934. 2020.
8. Klekovkin V. A., Markov N. G., Nebaba S. G. YOLO convolutional neural network models with attention mechanism for real-time computer vision systems. *Tomsk State University Journal of Control and Computer Science*, 2025, no. 72, pp. 39–50 (In Russian). doi:10.17223/19988605/72/4, EDN: KEWUSH
9. Egunov V. A., Koroleva I. Y., Tipaev D. V. Algorithms for the movement of a mobile robot with the construction of a real-time terrain map. *Engineering Journal of Don*, 2022, no. 4(88), pp. 126–132. EDN: YPFPSM. Available at: http://www.ivdon.ru (accessed 10 January 2026). (In Russian).
10. Paulenko D. A., Kovalev V. A., Snezhko E. V., Liauchuk V. A., Pechkovsky E. I. Recognition of underlying surface using a convolutional neural network on a single-board computer. *Informatics*, 2020, vol. 17, no. 3, pp. 36–43 (In Russian). doi:10.37661/1816-0301-2020-17-3-36-43, EDN: DIUEIM
11. Satsiuk A. V., Volodarets N. V. Modification of the YOLO model for a hybrid detection and trackingsystem in UAVs with an automatic guidance system. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 4, pp. 36–44 (In Russian). doi:10.31799/1684-8853-2025-4-36-44, EDN: YKQVJU
12. Limonova E. E., Sheshkus A. V., Nikolaev D. P., Ivanova A. A., Ilin D. A., Arlazarov V. L. Performance optimization of the initial layers of deep convolutional neural networks. *Russian Foundation for Basic Research Journal*, 2016, no. 4, pp. 84–96 (In Russian). doi:10.22204/2410-4639-2016-092-04-84-96, EDN: YOSCDX
13. Qin D., Leichner C., Delakis M., Fornoni M., Luo S., Yang F., Wang W., Banbury C., Ye C., Akin B., Aggarwal V., Zhu T., Moro D., Howard A. MobileNetV4: Universal models for the mobile ecosystem. *European Conference on Computer Vision*, Cham, Springer Nature Switzerland, 2024, pp. 78–96. doi:10.1007/978-3-031-73661-2_5
14. Zhang X., Zhou X., Lin M., Sun J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856. doi:10.1109/CVPR.2018.00716
15. Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (PMLR)*, 2019, pp. 6105–6114. doi:10.48550/arXiv.1905.11946
16. Han S., Pool J., Tran J., Dally W. J. Learning both weights and connections for efficient neural network. arXiv. 2015. https://doi.org/10.48550/ARXIV.1506.02626
17. Chen C., Wang T., Liu C., Liu Y., Cheng L. Lightweight convolutional transformers enhanced meta-learning for compound fault diagnosis of industrial robot. *IEEE Transactions on Instrumentation and Measurement*, 2023, vol. 72, pp. 1–12. doi:10.1109/TIM.2023.3277956
18. Frolov S. V., Korobov A. A., Savinova K. S., Potlov A. Yu. Neural network based system for control of microclimate and lighting in a neonatal incubator. *Sensors & Systems*, 2025, no. 1(279), pp. 62–68 (In Russian). doi:10.24412/1992-7185-2025-1-62-68, EDN: IIZLFV
19. Aksyonov D. S., Zhilyaev V. A., Markin N. I., Titov I. A. Object recognition system based on Raspberry Pi 4 and Intel Neural Computer Stick 2. *Information Systems and Technologies*, 2023, no. 4(138), pp. 10–16 (In Russian). EDN: FHOYQH
20. Satsuk A. V., Belyy R. V., Ishchenko A. E. Performance evaluation of YOLO algorithms for real-time object detection in embedded systems for autonomous vehicles. *Sbornik nauchnykh trudov Donetckogo instituta zheleznodorozhnogo transporta*, 2024, no. 4(75), pp. 73–82 (In Russian). EDN: OCNLZZ
21. Burggraaff O., Schmidt N., Zamorano J., Pauly K., Pascual S., Tapia C., Spyrakos E., Snik F. Standardized spectral and radiometric calibration of consumer cameras. *Optics Express*, 2019, vol. 27, no. 14, pp. 19075–19101. doi:10.1364/OE.27.019075
22. Lin T. Y., Dollar Pi., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125. doi:10.48550/arXiv.1612.03144
23. Kisantal M., Wojna Z., Murawski J., Naruniecj J., Cho K. Augmentation for small object detection. *arXiv preprint*. arXiv:1902.07296. 2019.
24. Banakh T., Głąb S., Jabłońska E., Swaczyna J. Haar I sets: Looking at small sets in Polish groups through compact glasses. *arXiv preprint*. arXiv:1803.06712. 2018.
25. Wang A., Chen H., Liu L., Chen K., Lin Z., Han J., Ding G. YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 107984–108011. doi:10.48550/arXiv.2405.14458
26. Howard A. G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*. arXiv:1704.04861. 2017.
27. Liu S., Qi L., Qin H., Shi J., Jia J. Path aggregation network for instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759-8768. doi:10.1109/CVPR.2018.00913
28. Zhang Z., Li X., Yang Y. Scaling YOLOv8 for high-precision real-time object detection. *arXiv preprint*. arXiv:2408.15857. 2024.
29. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.49
30. Chollet F. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258. doi:10.1109/CVPR.2017.195
31. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. doi:10.1109/CVPR.2018.00474
32. Desai M., Mewada H., Pires I. M., Roy S. Evaluating the performance of the YOLO object detection framework on COCO dataset and real-world scenarios. *Procedia Computer Science,* 2024, vol. 251, pp. 157–163. doi:10.1016/j.procs.2024.11.096
33. Satsiuk A. V. Optimization of the YOLOv8 architecture for UAV object capture tasks: Analysis of the trade-off between accuracy, speed and computational resources. *Vestnik Rostovskogo gosudarstvennogo universiteta putey soobshcheniya*, 2025, no. 2(98), pp. 35–42 (In Russian). doi:10.46973/0201-727X_2025_2_35, EDN: LCAOOM
34. Jacob B., Kligys S., Chen B., Zhu M., Tang M., Howard A., Adam H., Kalenichenko D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713. doi:10.1109/CVPR.2018.00286
35. Nagel M., Nagel M., Amjad R., Fournarakis M., Bondarenko Y., Baalen M., Blankevoort T. A white paper on neural network quantization. *arXiv preprint*. arXiv:2106.08295. 2021.
36. Cherepanov N. I., Stepina N. O., Nikiforov I. V. Improving image analysis and processing performance on the RISC-V platform with Lichee Pi 4A. *Proceedings of the Institute for System Programming of the RAS*, 2025, vol. 37, no. 5, p. 157–172. doi:10.15514/ISPRAS-2025-37(5)-12, EDN: UDVIEH

УДК 004.8
doi:10.31799/1684-8853-2026-3-35-48
EDN: XBAJEX

Научные статьи
Articles

Метод обучения моделей компьютерного зрения на основе кросс-модальной дистилляции знаний с применением больших визуальных моделей

А. В. Кучков[а], *аспирант, orcid.org/0009-0007-7508-3348*
А. М. Кашевник[а,б], *канд. техн. наук, доцент, orcid.org/0000-0001-6503-1447, alexey.kashevnik@iias.spb.su*
[а]*Университет ИТМО, Кронверкский пр., 49, Санкт-Петербург, 197101, РФ*
[б]*Санкт-Петербургский институт информатики и автоматизации РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *имеющиеся методы обучения мультимодальных моделей компьютерного зрения в большинстве случаев представляют собой отдельные ветви выделения распознавания признаков с поздним смешением результатов.* ***Цель:*** *разработать метод создания мультимодальных моделей компьютерного зрения с использованием единого представления мультимодальных данных для упрощения процессов смешения данных и дистилляции знаний.* ***Методы:*** *сериализация разреженных и плотных типов данных; кросс-модальная дистилляция знаний для архитектур компьютерного зрения; применение больших визуальных моделей для дистилляции знаний в сериализованном формате.* ***Результаты:*** *разработан метод обучения моделей компьютерного зрения на основе кривых Пеано с использованием дистилляции знаний из больших визуальных моделей. Метод позволяет производить смешение данных различных размерностей с помощью кросс-модального внимания в реальном времени посредством применения одномерных кривых Пеано (кривых Гильберта и Мортона) для сериализации многомерных данных. Предложенный метод показал задержку 50 мс против 40 мс в одномодальном режиме (Point Transformer v3), что свидетельствует о низких накладных расходах при кросс-модальной дистилляции на сериализованных картах признаков. Метод протестирован в режиме предобучения на датасете nuScenes с обращением к большой визуальной модели DINOv3. В режиме дистилляции использование 25 % от общего набора данных обеспечило 79,2 mIoU по сравнению с 82,1 mIoU при 100 % набора данных с функцией потерь — коэффициентом Отиаи.* ***Практическая значимость:*** *с использованием сериализованного представления данных методы кросс-модального смешения станут менее ресурсозатратными.* ***Обсуждение:*** *предложенный метод позволяет унифицировать декодер в модели сегментации смешанных данных благодаря кросс-модальному смешению сериализованных признаков после энкодеров изображений и облаков точек. При этом ранняя сериализация изображений показала себя нецелесообразной ввиду изначально плотной структуры изображений. Реализация метода сериализации изображений с меньшим временем выполнения даст возможность отказаться от раздельных энкодеров для ветвей облаков точек и изображений, что может существенно упростить архитектуру.*

Ключевые слова *— мультимодальные модели, дистилляция знаний, большие визуальные модели, кривые Пеано, кросс-модальное внимание.*

Для цитирования: Кучков А. В., Кашевник А. М. Метод обучения моделей компьютерного зрения на основе кросс-модальной дистилляции знаний с применением больших визуальных моделей. *Информационно-управляющие системы*, 2026, № 3, с. 35–48. doi:10.31799/1684-8853-2026-3-35-48, EDN: XBAJEX
For citation: Kuchkov A. V., Kashevnik A. M. Method for training computer vision models based on cross-modal knowledge distillation using large visual models. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 35–48 (In Russian). doi:10.31799/1684-8853-2026-3-35-48, EDN: XBAJEX

Введение

Мультимодальные модели компьютерного зрения предполагают использование двух и более сенсоров машинного зрения для восприятия окружающей обстановки в целях реагирования на изменения во внешней среде. Примерами подобных средств являются системы автономного вождения и системы-ассистенты предупреждения о столкновениях. Мультимодальные модели ввиду большой емкости требуют большого количества данных для обучения. Настоящая проблема стала еще более актуальной с появлением сетей-трансформеров, демонстрирующих высокие показатели точности, но требующих больше данных для их обучения. Для обучения мультимодальных моделей используются датасеты, содержащие данные от нескольких сенсоров, таких как лидары, камеры, радары и иные датчики. Среди публично доступных датасетов наибольшим объемом обладают такие наборы данных, как KITTI, nuScenes, SemanticKITTI и Waymo Open Dataset. Независимо от обилия датасетов количество взаимно аннотированных данных для нескольких сенсоров в рамках одной сцены ограничено. Несмотря на то, что датасеты KITTI, nuScenes и SemanticKITTI представляют облака точек и изображения для каждой из сцен, приоритетным сенсором остается лидар, данные которого, как правило, являются основным источником аннотаций.

Метки, явно заданные в датасете и содержащие в себе информацию о классе, форме и положении объекта, являются «жесткими» метками, количество которых зачастую ограниченно для определенных типов данных.

В общем случае подходы к уменьшению зависимости от «жестких» меток можно разделить на четыре категории [1, 2]:

— weakly-supervised-методы — использование меток, сгенерированных другой моделью-учителем («мягких» меток) [3];

— semi-supervised-методы — совместное использование небольшого размеченного набора данных и большого неразмеченного набора [4];

— self-supervised-методы — обучение с генерацией псевдометок самой моделью без привлечения внешней разметки [5];

— unsupervised-методы — полный отказ от использования как «жестких», так и «мягких» меток в пользу иных критериев оценки качества предсказаний [6].

Модель учителя представляет собой сеть, из которой происходит дистилляция знаний. Модель студента (ученика) при этом является сетью, в которую дистиллируются данные [7].

Weakly-supervised-методы используют метки, сгенерированные иной моделью и называемые «мягкими» метками. Качество сетей, обученных weakly-supervsied-методом, зависит как от модели учителя, так и от используемой функции ошибки дистилляции. Как правило, модель учителя обучена на большем наборе данных по сравнению с сетью ученика. Преимуществом метода является то, что отсутствуют какие-либо ограничения на архитектуру модели — ученика, что позволяет использовать как устоявшиеся подходы (CNN), так и более новые архитектуры, такие как визуальные трансформеры (ViT) [3].

Self-supervised-методы обычно реализуют архитектуру с двумя ветвями, в которой одна ветвь является моделью учителя, а другая — моделью студента. Их архитектуры могут как быть полностью идентичными, так и иметь некоторые особенности. Преимуществом данного метода является то, что он позволяет произвести обучение в один этап, а также отлично масштабируется вплоть до моделей с 7 млрд параметров, что допускает создание таких больших визуальных моделей (VFM, Visual Foundation Models), как DINOv3 (Distillation with No Labels) [5], SAM (Segment Anything Mode) [4], SegGPT (Segmenting Everything In Context) [8]. Однако в то время, как данный метод очень хорошо подходит для обучения больших моделей, в том числе визуальных, на текущий момент он не адаптирован под задачи реального времени, что особенно критично при выполнении нейронных сетей на встраиваемом оборудовании с ограниченными ресурсами.

Unsupervised-методы в свою очередь подразумевают полное отсутствие контроля над обучением посредством каких-либо меток. В классическом понимании unsupervise-обучение применяется для кластеризации, оценки плотности вероятности и снижения размерности пространства признаков.

В настоящей работе основное внимание уделяется weakly-supervised-методу со сжатием признаков модели учителя в пространство меньшей размерности сети ученика. Преимущество указанного подхода заключается в отсутствии необходимости применять «жесткие» метки для вычисления критериев точности работы моделей, так как на этапе предобучения происходит дистилляция знаний из cls-токена сети VFM в сеть PTv3, что позволяет ускорить процесс обучения. Таким образом, сеть DINOv3 выступает в качестве генератора self-supervised признаков для нашей модели.

Целью данной работы является создание метода обучения мультимодальных моделей, использующих изображения и облака точек, на основе сериализации данных и привлечения VFM – DINOv3. Научная значимость настоящего исследования заключается в следующем:

1) предложен метод совместного представления признаков облаков точек изображений в сериализованном виде с помощью кривых Пеано;

2) разработан метод смешения признаков изображений и облаков точек в сериализованном пространстве;

3) разработан метод дистилляции знаний в сериализованное пространство признаков с использованием VFM – DINOv3.

Обзор исследований обучения моделей компьютерного зрения на основе дистилляции знаний

Непосредственный метод дистилляции знаний зависит от модальности дистиллируемых данных. Методы дистилляции знаний для различных модальностей представлены в табл. 1. Дистилляция знаний внутри одной модальности, как правило, осуществляется в целях сжатия знаний исходной модели в веса меньшей.

Наибольший интерес представляют методы переноса знаний в направлениях Image → Lidar и Image → Image ввиду разнообразия таких VFM, как DINOv3, SAM и SegGPT, обученных self-supervsied-методом на больших наборах данных. VFM могут служить в качестве учителя для генерации качественных карт признаков без необходимости дообучения. Вследствие большой емкости указанных моделей область их применения не ограничена лишь RGB-изображениями; они могут

■ ***Таблица 1.*** Методы кросс-модальной дистилляции знаний
■ ***Table 1.*** Cross-modal knowledge distillation methods

Данные	Типы дистилляции знаний	Примеры	Функции потерь	Применение
Изображения и облака точек	Выравнивание признаков, кросс-внимание, контрастная дистилляция знаний	DITR [9], Modal Mimicking [10], OLIVINE [3]	Функция ошибки InfoNCE между сетями 2D и 3D, коэффициент Отиаи, расстояние Кульбака – Лейблера	Предобучение 3D-моделей при помощи 2D-учителей
Облака точек радара и лидара	Выравнивание признаков, кросс-внимание, контрастная дистилляция знаний	RadarDistill [11]	InfoNCE, коэффициент Отиаи, расстояние Кульбака – Лейблера	Предобучение радарных моделей при помощи облаков 3D-учителей
Текст и облака точек	Дистилляция с учетом интра- и кросс-модальных отношений	Multimodal Relation Distillation [12], CLIP2Point [13]	NT-Xent Loss, коэффициент Отиаи, расстояние Джефри, нормализованное сходство	Предобучение 3D-моделей с LLM (большой языковой моделью)
Текст и изображения	Контрастная дистилляция знаний	CLIP [14], TinyCLIP [15], SCKD [16]	InfoNCE, MSE	Предобучение 2D-моделей с LLM

■ ***Таблица 2.*** Большие визуальные модели
■ ***Table 2.*** Visual Foundation Models

Модель	Модальность	Параметры	Задачи
DALL-E [17]	Текст + изображения	~12 млрд (v1), 3,5 млрд (v2)	Генерация изображений
SAM [4]	Текст + изображения	До ~636 млн	Сегментация изображений
Stable Diffusion [18]	Текст + изображения	До ~8 млрд	Генерация изображений
CLIP ViT-L/14 [14]	Текст + изображения	~ 428 млн	Сопоставление изображений и текста
Florence-2 [19]	Текст + изображения	230 млн (Base), 770 млн (Large)	Детектирование, сегментация
DINOv3 [5]	Изображения	До ~7 млрд	Извлечение универсальных визуальных признаков
CogVLM-17B [20]	Текст + изображения	17 млрд	Визуальные ответы на вопросы
OpenVLA [21]	Текст + изображения	7 млрд	Генерация действий для манипуляций
Qwen2.5-VL-32B-Instruct [22]	Текст + изображения	32 млрд	Мультимодальные задачи (текст + изображение)
Gemma 3 [6]	Текст + изображения	1/4/12/27 млрд	Мультиязычная VLM

применяться для задач дистилляции знаний и для модальностей, отличных от RGB-изображений, таких как лидарные и радарные данные.

Поскольку обучающий набор не обязан содержать истинные метки, то для процесса дистилляции могут использоваться неразмеченные экспериментальные данные с последующим получением «слабых» меток с помощью VFM. Использование VFM зачастую подразумевает генерацию «слабых» меток до начала процесса предобучения. Выбор VFM определяется исходя из требований к емкости модели, точности, а также ее доступности. В табл. 2 представлены актуальные на текущий момент VFM.

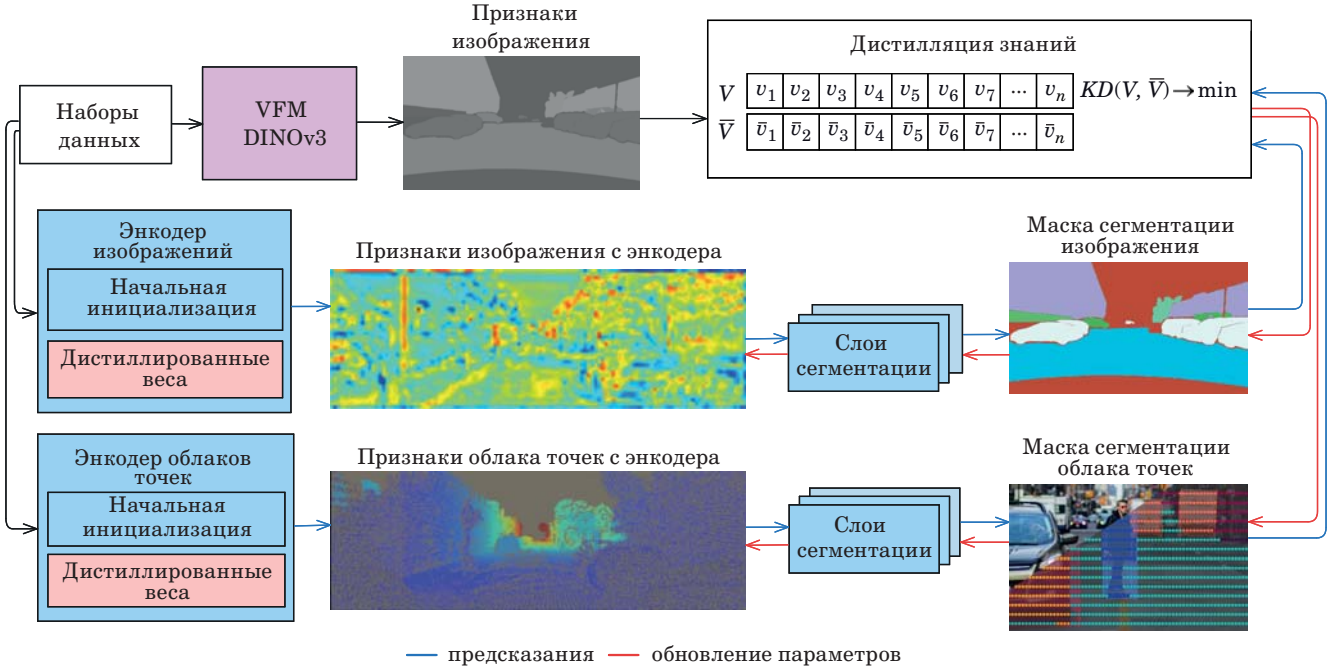
Как правило, для получения «слабых» меток используется специальный тип моделей VFM, обученных на выполнение общих задач компьютерного зрения (сегментация, детектирование). В данную группу входят такие модели, как DINOv1/2/3, SAM, CLIP и Qwen2.5-VL.

Предобучение мультимодальных моделей с помощью VFM подразумевает индивидуальный подход к дистилляции знаний для данных каждого из сенсоров. Более того, дистилляция может осложняться применением смешения данных в модели. Для дистилляции знаний в подобных моделях необходимо решить несколько вопросов: на каком этапе модели происходит дистилляция, из какой модальности дистиллируются данные и какую функцию ошибки дистилляции использовать?

Этап дистилляции, как правило, происходит в поздних слоях сети ввиду наличия качественных карт признаков [3, 9]. Дистилляция на ранних этапах малоэффективна, так как первые слои не содержат глубоких признаков. Модальность дистилляции определяется наличием VFM, обученной на большом объеме данных. В связи с тем, что большинство современных VFM обучены на RGB-изображениях, использование данной модальности на текущий момент является общепринятым подходом.

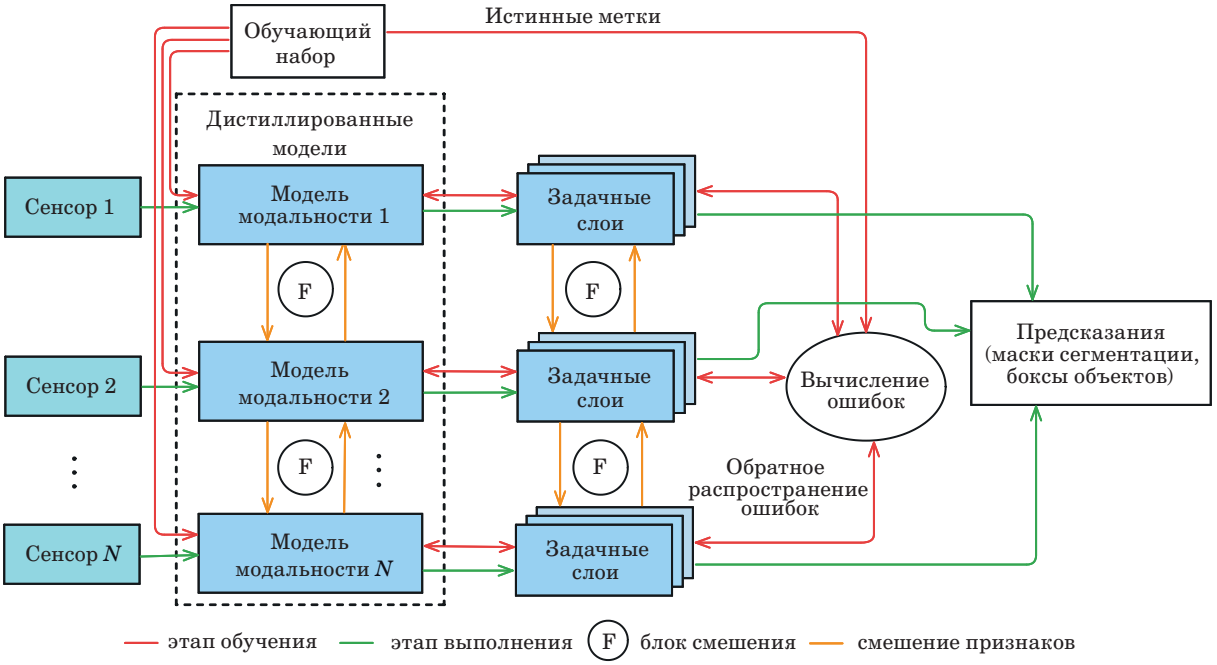
Для выбора функции ошибки дистилляции необходимо учитывать, что не все функции ошибок позволяют оценить локальные и глобальные признаки. В связи с чем для дистилляции, как правило, применяются сразу несколько функций ошибок. Схема обучения двухмодальной модели с применением механизма дистилляции знаний показана на рис. 1. Применение подобных моделей охватывает три этапа: предобучение, обучение и выполнение. Дистилляция знаний происходит исключительно на первом этапе, который занимает большую часть времени. Источником данных на стадиях предобучения и обучения служат такие наборы данных, как KITTI, nuScenes, Waymo Open Dataset и др. 2D- и 3D-модели ответственны за выделение признаков из облаков точек и изображений. VFM в свою очередь (DINOv3 на схеме) выделяет признаки из изображений. Дистилляция происходит на двух масштабах: на семантическом уровне и на уровне признаков.

Далее дистиллированные модели проходят через этап дообучения (fine-tuning), где используются истинные метки из обучающего набора для необходимых модальностей. По причине того, что модель уже является предобученной, количество данных в обучающем наборе может быть сокращено вплоть до 1 % [3]. Более того, возможен сценарий Linear Probing (LP), в котором замораживается сеть выделения признаков, а обучению подвергаются лишь задачно-ориентированные слои (рис. 2). Данный метод позволяет дообучить модель под конкретную задачу гораздо быстрее, так как требуется не полное обновление параметров модели, а лишь нескольких слоев. Сценарий LP может быть применен к произвольному количеству источников данных.

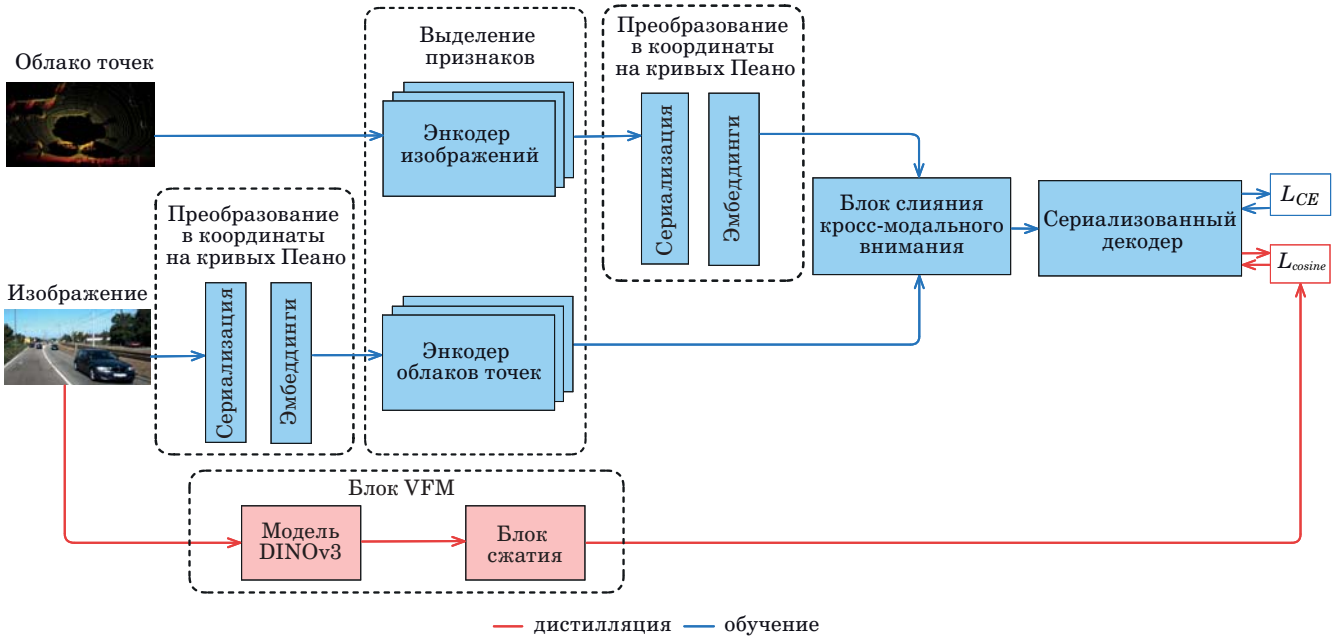


■ ***Рис. 1.*** Пример обучения моделей обработки облаков точек и RGB-изображений с предложенным методом дистилляции знаний

■ ***Fig. 1.*** An example of multimodal model pretraining with LiDAR and RGB input using the proposed knowledge distillation method



■ ***Рис. 2.*** Смешение данных в мультисенсорной системе с дистилляцией знаний
■ ***Fig. 2.*** Data fusion in a multisensory system with knowledge distillation



■ ***Рис. 3.*** Обучение моделей компьютерного зрения на основе дистилляции знаний с использованием VFM
■ ***Fig. 3.*** Training computer vision models based on knowledge distillation using VFM

Метод дистилляции знаний в мультисенсорной модели

В настоящем разделе предлагается метод создания и обучения мультимодальных моделей с использованием кривых Пеано и дистилляции знаний с применением VFM. Метод предполагает наличие следующих компонентов: источника данных (наборов данных для обучения), сетей выделения признаков для каждого сенсора (энкодеров), блока VFM, блока смешения данных на основе кросс-модальной дистилляции, а также энкодера смешанных признаков (рис. 3). Сначала облака точек подвергаются процессу сериализации,

который трансформирует 3D-координаты в координаты на одномерной кривой Пеано, проходящей через точки координатной системы с заданным размером сетки. Данному процессу подвергается лишь облако точек ввиду его разреженной структуры. Затем происходит этап выделения признаков, на котором из облака точек и изображений формируются карты признаков. Дальнейшим шагом является преобразование всех карт признаков в один формат — координаты на кривых Пеано. Для этого признаки изображений сериализируются и разбиваются на части, формируя понятную для декодера-трансформера последовательность. Перед декодером стоит блок смешения данных на основе кросс-модального внимания, преобразующий две одномерные последовательности на входе в одну с помощью двух блоков внимания.

Сериализация позволяет сократить расходы на вычисление матриц внимания путем устранения необходимости построения больших деревьев для вычисления взаимного расположения элементов облаков точек и изображений. Смешанные данные попадают в декодер, структура которого является зеркальной к энкодеру. Декодер на выходе формирует классы для каждой точки из облака.

Во время обучения в качестве основного критерия используется ошибка кросс-энтропии (Cross Entropy, CE), в то время как для предобучения используется коэффициент Отиаи.

Сериализация карт признаков

Для представления признаков облаков точек и изображений в едином пространстве используется техника сериализации многомерных данных [23]. Сериализация осуществляется с помощью одной или нескольких кривых Пеано, преобразующих облака точек и признаки изображений в набор координат. Обучение моделей компьютерного зрения на основе дистилляции знаний с использованием VFM одномерной кривой:

$$\varphi^{-1} : Z^n \to Z^1,$$

где φ^{-1} — функция отображения n-мерной точки в определенное положение на линии.

Необходимость в сериализации облака точек обусловлена неупорядоченной природой облаков точек, которая мешает применению механизмов внимания, использующихся в сетях-трансформерах. Следующая причина необходимости в иной форме представления данных возникает из-за того, что мультимодальные модели компьютерного зрения обладают высокой вычислительной сложностью ввиду высокой размерности данных. Как показано в работе [23], комбинация различных механизмов сериализации улучшает производительность моделей.

В случае с изображениями сериализация также может иметь место, что показано в пилотном исследовании, однако вычисление признаков на одномерной кривой не имеет никаких преимуществ по сравнению со стандартными методами выделения признаков, так как изображения по своей природе уже упорядочены.

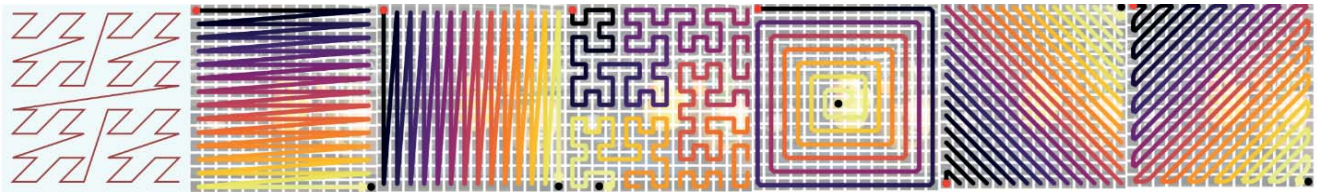
В настоящей работе делается акцент на том, что использование кривых Пеано позволяет реализовать механизм кросс-модального внимания с последующим смешением данных более эффективно, чем это можно было бы сделать с помощью вокселизации или проецирования облаков точек на 2D-плоскость. Альтернативным подходом является использование поточечной (pointwise) свертки для ускорения вычисления слоя самовнимания (self-attention) [24]. Однако данный подход слабо применим для трехмерных признаков облака точек.

Сериализованное представление многомерных координат на кривой характеризуется дискретным шагом g, представляющим размер сетки. В текущей работе подобной трансформации подвергаются не только облака точек, но и признаки изображения, т. е. для каждого признака ищется соответствие на кривой Пеано. Также, следуя работе [23], для кодирования батчей используются k нулевых старших разрядов позиции пикселя:

$$Encode(\mathbf{p}, b, g) = (b << k) \mid \varphi^{-1}(|\mathbf{p}/g|),$$

где $\mathbf{p}$ — координаты пикселя; b — размер батча; g — размер ячейки; << — битовый сдвиг влево.

Для сохранения возможности обратного преобразования, а также для ускорения вычислений вместо трансформации облаков и изображений непосредственно в формат кривых Пеано используются лишь индексы, ставящие соответствие каждой точке или пикселю координату на кривой. В работе [25] авторы выделили наиболее подходящие типы кривых (рис. 4). В настоящей работе для сериализации используются кривые Мортона и Гильберта. Обе кривые обладают свойством сохранять взаимное расположение двух точек, т. е. индексы точек, лежащих близко в N-мерном пространстве, также будут лежать близко друг к другу. При этом кривая Гильберта обладает данным свойством в большей степени, чем кривая Мортона. Остальные типы пространственных кривых таким свойством не обладают и, соответственно, не подходят для вычисления self-attention на признаках, полученных с облаков точек, ввиду наличия разрывов при переходе через строки или столбцы входного тензора. В работе [25] авторы показали, что в зависимости от датасета и используемой модели форма оптимальной аппроксимирующей кривой может



■ ***Рис. 4.*** Кривые Пеано. Слева направо: кривая Мортона (z-order), разбиение по строкам, разбиение по столбцам, кривая Гильберта, спираль, диагональ, змеевидная кривая

■ ***Fig. 4.*** Peano curves. From left to the right: Morton (z-order), row order, column order, Hilbert, spiral, diagonal, snake-like curves

меняться. Учитывая данную информацию, для сравнения точности модели в текущей работе были протестированы как отдельные кривые Мортона и Гильберта, так и их комбинации.

Важным свойством любой кривой является сохранение пространственной близости точек, имеющих близкие координаты. Не все кривые обладают данным свойством и имеют периодические разрывы. Так, на кривых, построенных методом разбиения по строкам и столбцам, отмечаются разрывы через каждые N элементов.

Дистилляция сериализованных данных

На этапе дистилляции применяется метод, описанный в DITR (DINO in the Room) [9]. Данный подход заключается в вычислении коэффициента Отиаи в качестве функции ошибки для направления процесса дистилляции в сторону схожести карт признаков PTv3 и DINOv3. Для дистилляции из DINOv3 были взяты карты признаков из iBOT [5] подсети, так как данные признаки содержат информацию на уровне патчей изображения. Таким образом, минимизировалась следующая ошибка:

$$L_{cosine} = \sum [1 - x_i^{pred}\, x_i^{2D} \div (||x_i^{pred}||\ \ ||x_i^{2D}||)] \div \upsilon,$$

где υ — набор точек, видимый как минимум на одной камере.

Смешение данных

Наиболее распространенным подходом к смешению данных является раздельная обработка данных различных сенсоров ввиду простоты адаптации существующих решений [26]. Авторы [27] подтвердили, что такие методы смешения мультимодальных данных, как конкатенация признаков и сложение признаков, хоть и могут работать при определенных сценариях, не дают существенного превосходства над одномодальной сетью из-за слабой корреляции признаков. При этом в последнее время все чаще применяются механизмы смешения данных на основе кросс-модального внимания [28, 29]. В настоящей работе для смешения данных также используется метод кросс-модального внимания с двумя Attention-ветвями с матрицами ключей **K** от модальностей изображений и облаков точек. Блок данной архитектуры показал свою эффективность в кросс-модальном смешении данных, включая геометрическую и визуальную последовательности [28]. Предварительно для избежания лишних вычислений облака точек фильтруются по полю зрения камер.

Блок кросс-модального смешения (рис. 5) представляет собой комбинирование двух блоков Multi-Head Attention MHA с вычислением произведения $\mathbf{QK}^{\mathrm{T}}$, где матрицы **Q** и **K** принадлежат разным модальностям. Вычисление Attention в данном случае является двунаправленным, так как используется два блока с вычислением коэффициентов схожести от лидара к камере и от камеры к лидару. Полученные таким образом карты признаков конкатенируются и подаются на вход декодера сети, который является общим для ветвей лидара и камеры. Получение смешанной карты признаков может быть описано следующим выражением:

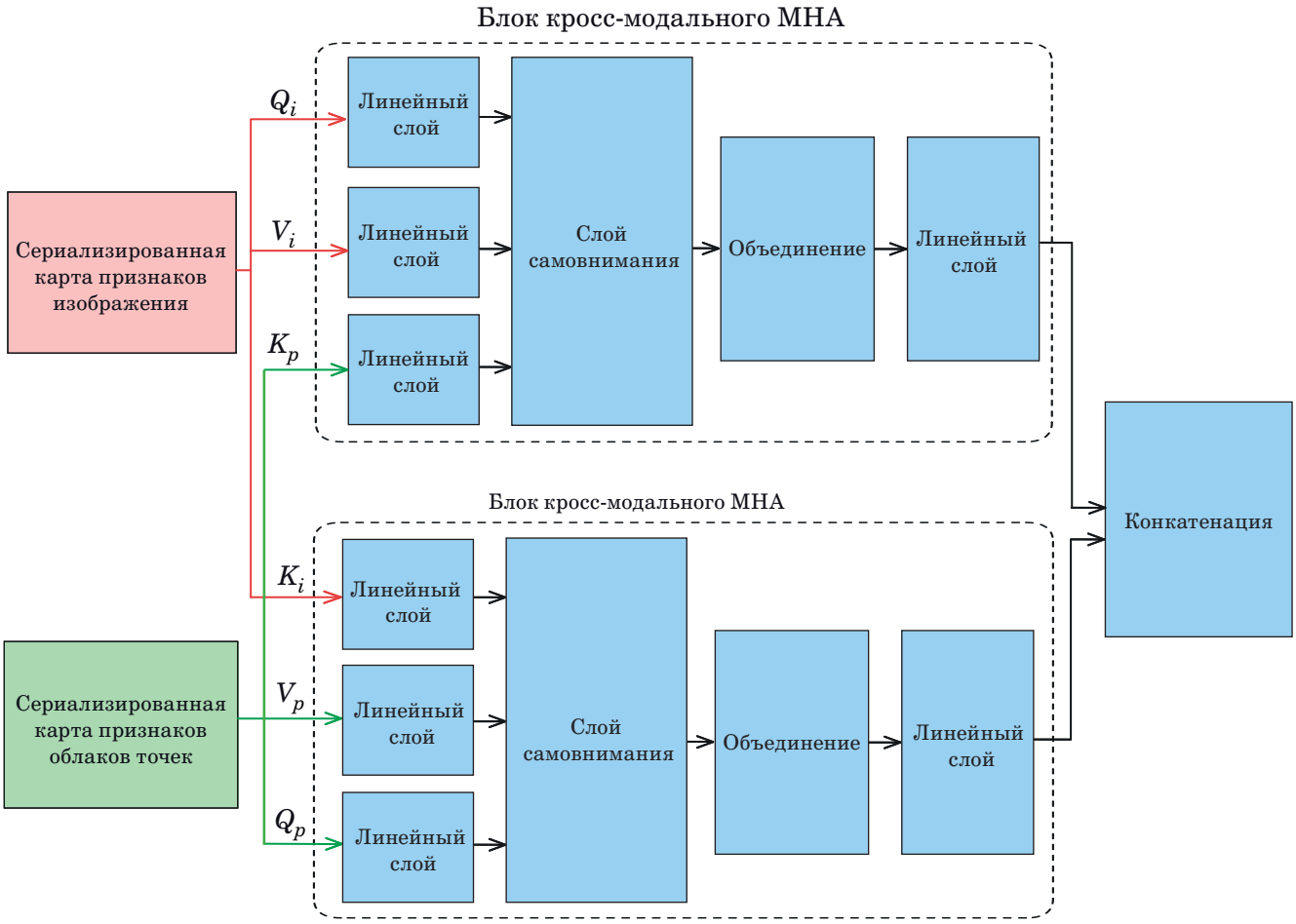
$$X_f = Concatenation(Attend_i(X_i, X_p), Attend_p(X_i, X_p)),$$

где $Attend_i$ и $Attend_p$ — внимание между изображениями (X_i) и облаками точек (X_p).

Эксперименты

В качестве базовой модели в текущей работе принята сеть 3D-трансформер PTv3, являющаяся SOTA-моделью на датасете nuScenes. Для дистилляции знаний применялась большая визуальная модель DINOv3 с ее функцией ошибки iBOT, позволяющей получить признаки на уровне входных частей изображений. Для сериализации облаков точек и изображений использован блок сериализации, преобразующий исходные облака точек и признаки изображений в координаты на кривой Пеано.

Эксперимент поставлен следующим образом. В первой части проводилось пилотное исследование на предмет возможности применения сети PTv3 для выделения признаков из изображений.

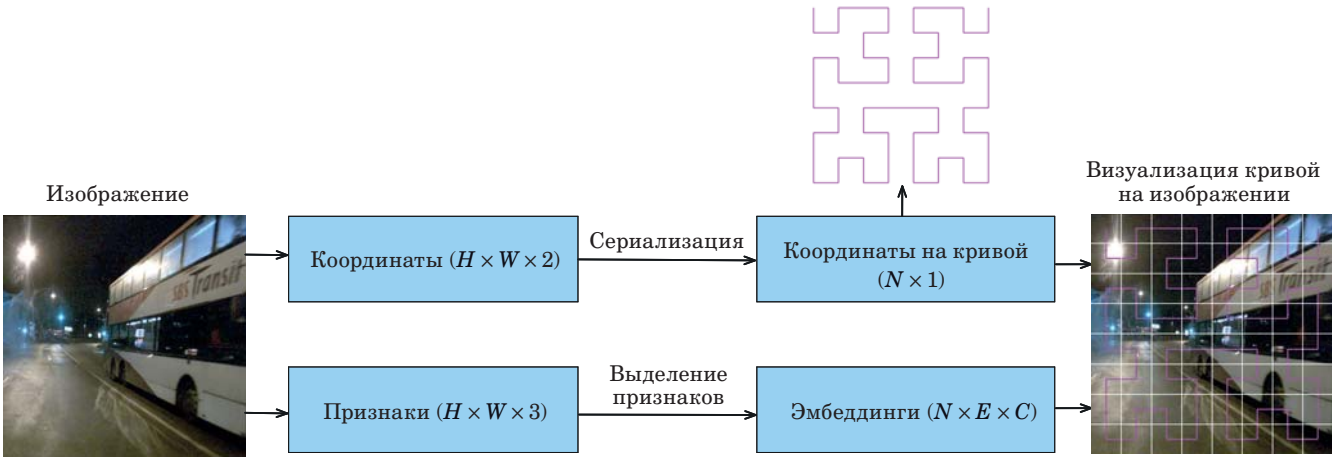


■ ***Рис. 5.*** Кросс-модальное смешение с помощью метода кросс-внимания
■ ***Fig. 5.*** Cross-modal data fusion with cross-modality attention

Во второй части исследовалась применимость механизма дистилляции знаний для обучения PTv3 с одномодальным мультимодальным входом.

Для проверки возможности выделения признаков из изображений был взят датасет CIFAR-10 [30], состоящий из 60 000 цветных изображений размером 32 × 32 с 10 классами, по 6000 изображений на каждый класс. Датасет поделен на 50 000 изображений в тренировочном наборе и 10 000 изображений в тестовом наборе. Из сети PTv3 был убран декодер, вместо которого использованы полносвязные слои для классификации изображений. Для каждого изображения построены координаты пикселей и каналы признаков (цвета RGB). Процесс сериализации продемонстрирован на рис. 6. Исследованы следующие варианты сериализации: сериализация кривыми Гильберта, Мортона, инвертированной кривой Гильберта, инвертированной кривой Мортона, а также сериализация с последовательным применением данных кривых.

В качестве каналов, содержащих признаки данных, использованы интенсивности изображения вместо интенсивности облака точек. Все вычисления выполнялись на NVIDIA RTX 3080. Размер набора данных — 32, количество эпох — 25, размер изображения — 32 × 32. Для лучшей визуализации координаты кривой отложены поверх оригинального изображения. В сравнение были включены различные комбинации из кривых Гильберта и Мортона, а также из их транспонированных вариантов. Результаты обучения трансформера PTv3 на изображениях представлены в табл. 3. Проведено сравнение метрик качества сверточных сетей VGG16 [31], ResNet50 [32] и Resnet18 [32] (которые зачастую используются как базовые для извлечения признаков) с результатами работы PTv3 при различных методах сериализации. Результаты пилотного исследования показали недостаточную эффективность метода сериализации для выделения признаков с изображения ввиду низкой скорости выполнения.



■ ***Рис. 6.*** Последовательность действий сериализации изображения
■ ***Fig. 6.*** Image serialization process. For visualization purposes, the curve is drawn over the original image

В то же время этот метод может применяться для смешения данных уже после процесса выделения признаков. По итогам первой части экспериментов на текущий момент было решено отказаться от PTv3 энкодера для ветви изображений. Для выделения признаков с изображений выбрана сеть Resnet50, потому как она обладает компромиссными характеристиками по скорости работы и точности. При сериализации признаков применялась комбинация кривых Мортона, Гильберта и их транспонированных вариантов, поскольку она продемонстрировала наивысшие показатели точности классификации как в пилотном исследовании, так и в самой работе PTv3.

■ ***Таблица 3.*** Метрики PTv3 с модальностью изображений и сравнение с другими моделями
■ ***Table 3.*** PTv3 metrics trained with image modality and comparison with other models

Модель	Точность, %	Параметры	Задержка, мс
PTv3 ($H + H_{trans} + Z + Z_{trans}$)	0,989	25,6 M	~31
PTv3 (H)	0,859	25,6 M	~28
PTv3 ($H + H_{trans}$)	0,890	25,6 M	~28
PTv3 ($H + Z$)	0,883	25,6 M	~28
PTv3 (Z)	0,969	25,6 M	~26
PTv3 ($Z + Z_{trans}$)	0,860	25,6 M	~26
VGG16 (32×32)	0,933	138 M	~9,9
ResNet18 (32×32)	0,813	11,5 M	~6,4
ResNet50 (32×32)	**0,922**	**~25 M**	**~8,3**

Примечание: H, Z и H_{trans}, Z_{trans} — кривые Гильберта, Мортона и их транспонированные варианты соответственно.

Облака точек подвергались сериализации с последующим выделением признаков с помощью сериализованного энкодера сети PTv3. Дальнейшее смешение данных происходило уже в сериализованном виде. Сравнение проводилось с предыдущими версиями Point Transformer, а также с моделями, основанными на схожем подходе. DITR [9] использует PTv3 и DINOv2 для генерации мультимодальных карт признаков и производит многоступенчатое смешение данных в декодере модели. Transformer Based Lidar Camera Fusion [33] использует Multi-Head Self-Attention на смешанной последовательности для повышения точности сегментации. LidarFormer [34] внедрил Cross-Space transformer для извлечения глобальной информации из BEV-карты признаков, а также Cross-Task transformer для извлечения семантической и объектной информации из смешанной карты признаков. Для предлагаемой в настоящей работе модели обучение производилось в двух сценариях (табл. 4): Linear Probing и с дообучением на частях данных из обучающего датасета nuScenes. В сценарии LP энкодер модели оставался замороженным, в то время как декодер подвергался обучению. Данные приведены для валидационной (val) и тренировочной (train) выборок.

Для каждой модели были произведены измерения времени выполнения одного батча на 16 элементов. Время задержки, приведенное в последнем столбце, является полным временем выполнения

■ ***Таблица 4.*** Результаты сравнительного анализа
■ ***Table 4.*** Results of the comparative analysis

Модель	LP		Дообучение на количестве данных, %						Веса	Время, мс
			1	5	10	25	100			
	val	train					val	train		
DITR [9]	–	–	–	–	–	–	**84,2**	**85,1**	7 B+25,6 M	~800
PTv3 [23]	–	–	–	–	–	–	81,2	83,0	25,6 M	**~ 40**
BEVFusion [35]	–	–	–	–	–	–	62,7	–	–	~120
T, B, LiDAR-Camera Fusion [33]	–	–	–	–	–	–	80,6	–	–	~83
LidarFormer [34]	–	–	–	–	–	–	82,7	81,5	77 M	~530
PTv3 Fusion	57,4	58,2	60,1	65,3	70,1	79,2	82,1	84,1	31,1 M	~50

(включая процесс сериализации для моделей PTv3 и PTv3 Fusion и копирования данных из CPU в GPU и обратно). Из представленных моделей DITR обладает наибольшим mIoU как на валидационном, так и на тренировочном датасете. При этом DITR также имеет самое большое время выполнения ввиду использования VFM DINOv2. LidarFormer, занимая второе место, обладает меньшей задержкой в 530 мс, хотя и недостаточной для задач реального времени.

Представленная в настоящей работе модель PTv3 Fusion превосходит оригинальную модель по точности на валидационном датасете (82,1 против 81,2 у PTv3). При этом, в отличие от DITR, она обладает временем, достаточным для выполнения задач реального времени. Дополнительные накладные расходы, связанные с модулем кросс-модального внимания, компенсируются применением сериализации с комбинацией кривых Пеано в виде кривых Гильберта, Мортона и их транспонированных вариантов.

Заключение

Предложенный в настоящей работе метод обучения мультимодальных моделей компьютерного зрения показал преимущество применения кривых Пеано для сериализации признаков мультимодальных данных. Использование комбинаций кривых Мортона, Гильберта позволило производить кросс-модальное внимание со смешанными признаками облаков точек и изображений быстрее по сравнению с альтернативными подходами. Проведенный анализ возможности ранней сериализации изображений в кривые Пеано показал избыточные затраты на обработку без особых преимуществ в точности. Данный результат объясняется плотной структурой изображений, для которых преобразование координат из плоскости в одномерную кривую не снижает вычислительные затраты. По результатам эксперимента было принято решение производить позднюю сериализацию, на этапе смешения данных с признаками облаков точек. В будущей работе планируется адаптировать представление изображений в более эффективном формате для ранней сериализации, что позволит унифицировать процесс извлечения признаков и провести тестирование на бóльших, по сравнению с CIFAR-10, датасетах.

Эксперименты показали, что VFM DINOv3 может служить учителем в задаче дистилляции знаний для ускорения процесса обучения модели. При этом коэффициент Отиаи является основной функцией потерь для дистилляции знаний из локальных карт признаков DINOv3 в энкодер и декодер Point Transformer v3. Поскольку DINOv3 используется исключительно на стадии предобучения, в режиме исполнения, VFM не вносит никаких дополнительных вычислительных задержек.

Финансовая поддержка

Исследование выполнено в рамках бюджетной темы FFZF-2025-0003.

Литература

1. **Shen W., Peng Z., Wang X., Wang H., Cen J., Jiang D., Xie L., Yang X., Tian Q.** A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 8, pp. 9284–9305. doi:10.1109/TPAMI.2023.3246102
2. **Zhu X., Goldberg A. B.** *Introduction to semi-supervised learning*. Ser.: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3. Morgan & Claypool Publishers, 2009, pp. 1–130. doi:10.2200/S00196ED1V01Y200906AIM006
3. **Zhang Y., Hou J.** Fine-grained Image-to-LiDAR contrastive distillation with Visual Foundation Models. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. doi:10.48550/arXiv.2405.14271
4. **Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., Girshick R.** Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
5. **Siméoni O., Vo H. V., Seitzer M., Baldassarre F., Oquab M., Jose C., Khalidov V., Szafraniec M., Yi S., Ramamonjisoa M., Massa F., Haziza D., Wehrstedt L., Wang J., Darcet T., Moutakanni T., Sentana L., Roberts C., Vedaldi A., Tolan J., Brandt J., Couprie C., Mairal J., Jégou H., Labatut P., Bojanowski P.** DINOv3. 2025. arXiv:2508.10104. doi:10.48550/arXiv.2508.10104
6. **Kamath A., Ferret J., Pathak S., Vieillard N., Merhej R., Perrin S., Matejovicova T., Ramé A., Riviére M., Rouillard L., Mesnard T., Cideron G., Grill J.-B., Ramos S., Yvinec E., Casbon M., Pot E., Penchev I., Liu G., Visin F., Kenealy K., Beyer L., Zhai X., Tsitsulin A., Busa-Fekete R., Feng A., Sachdeva N., Coleman B., Gao Y., Mustafa B., Barr I., Parisotto E., Tian D., Eyal M., Cherry C., Peter J.-T., Sinopalnikov D., Bhupatiraju S., Agarwal R., Kazemi M., Malkin D., Kumar R., Vilar D., Brusilovsky I., Luo J., Steiner A.** Gemma 3 Technical Report. 2025. arXiv:2503.19786. doi:10.48550/arXiv.2503.19786
7. **Татарникова Т. М., Мокрецов Н. С.** Метод дистилляции знаний для языковых моделей на основе выборочного вмешательства в обучение. *Кибернетика и системный анализ*, 2025, т. 150, № 2, с. 361–365. doi:10.15827/0236-235X.150.361-365
8. **Wang X., Zhang X., Cao Y., Wang W., Shen C., Huang T.** SegGPT: Towards segmenting everything in context. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140. doi:10.1109/ICCV51070.2023.00110
9. **Zeid K. A., Yilmaz K., de Geus D., Hermans A., Adrian D., Linder T., Leibe B.** DINO in the Room: Leveraging 2D foundation models for 3D segmentation. 2025. arXiv:2503.18944. doi:10.48550/arXiv.2503.18944
10. **Yang M., Qi Y., Xiong B., Zhang Z., Liu Y.** Modal mimicking knowledge distillation for monocular three-dimensional object detection. *Engineering Applications of Artificial Intelligence*, 2025, vol. 160, Article 111821. doi:10.1016/j.engappai.2025.111821
11. **Bang G., Choi K., Kim J., Kum D., Choi J. W.** RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15491–15500. doi:10.1109/CVPR52733.2024.01467
12. **Wang H., Bao Y., Pan P., Li Z., Liu X., Yang R., Huang D.** Multi-modal relation distillation for unified 3D representation learning. *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Cham, Springer Nature Switzerland, 2025, pp. 364–381. doi:10.1007/978-3-031-73414-4_21
13. **Huang T., Dong B., Yang Y., Huang X., Lau Rynson W. H., Ouyang W., Zuo W.** CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22100–22110. doi:10.1109/ICCV51070.2023.02025
14. **Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I.** Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*; *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html (дата обращения: 11.06.2025).
15. **Wu K., Peng H., Zhou Z., Xiao B., Liu M., Yuan L., Xuan H., Valenzuela M., Chen X., Wang X., Chao H., Hu H.** TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21970–21980. doi:10.1109/ICCV51070.2023.02008
16. **Xu R., Xiang Z., Zhang C., Zhong H., Zhao X., Dang R., Xu P., Pu T., Liu E.** SCKD: Semi-supervised cross-modality knowledge distillation for 4D radar object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 9, pp. 8933–8941. doi:10.1609/aaai.v39i9.32966
17. **Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I.** Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning*; *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html (дата обращения: 11.06.2025).
18. **Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B.** High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695. doi:10.1109/cvpr52688. 2022.01042

19. **Xiao B., Wu H., Xu W., Dai X., Hu H., Lu Y., Zeng M., Liu C., Yuan L.** Florence-2: Advancing a unified representation for a variety of vision tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, June 2024, pp. 4818–4829. doi:10.1109/CVPR52733.2024.00461
20. **Wang W., Lv Q., Yu W., Hong W., Qi J., Wang Y., Ji J., Yang Z., Zhao L., Song X., Xu J., Xu B., Li J., Dong Y., Ding M., Tang J.** CogVLM: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Curran Associates, Inc., 2024, pp. 121475–121499. doi:10.52202/079017-3860
21. **Kim M. J., Pertsch K., Karamcheti S., Xiao T., Balakrishna A., Nair S., Rafailov R., Foster E. P., Sanketi P. R., Vuong Q., Kollar T., Burchfiel B., Tedrake R., Sadigh D., Levine S., Liang P., Finn C.** OpenVLA: An open-source vision-language-action model. *Proceedings of the 8th Conference on Robot Learning*; *Proceedings of Machine Learning Research (PMLR)*, P. Agrawal, O. Kroemer, W. Burgard (Eds.), November 2025, vol. 270, pp. 2679–2713. https://proceedings.mlr.press/v270/kim25c.html (дата обращения: 11.03.2025).
22. **Bai S., Chen K., Liu X., Wang J., Ge W., Song S., Dang K., Wang P., Wang S., Tang J., Zhong H., Zhu Y., Yang M., Li Z., Wan J., Wang P., Ding W., Fu Z., Xu Y., Ye J., Zhang X., Xie T., Cheng Z., Zhang H., Yang Z., Xu H., Lin J.** Qwen2.5-VL Technical Report. 2025. arXiv:2502.13923. doi:10.48550/arXiv.2502.13923
23. **Wu X., Jiang L., Wang P.-S., Liu Z., Liu X., Qiao Y., Ouyang W., He T., Zhao H.** Point Transformer V3: Simpler, faster, stronger. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4840–4851. doi:10.1109/cvpr52733.2024.00463
24. **Бережнов Н. И., Сирота А. А.** Совершенствование механизмов внимания для архитектуры трансформер в задачах повышения качества изображений. *Компьютерная оптика*, 2024, т. 48, № 5, с. 726–733. doi:10.18287/2412-6179-CO-1393
25. **Kutscher D., Chan D. M., Bai Y., Darrell T., Gupta R.** REOrdering patches improves vision models. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. https://openreview.net/forum?id=g56WiaXKGF (дата обращения: 12.01.2026).
26. **Иванов В. Ф., Охотников А. Л., Градусов А. Н.** Алгоритм комплексирования сенсорных данных для задач автоматического управления подвижным составом. *Автоматика на транспорте*, 2024, т. 10, № 4, с. 360–371. doi:10.20295/2412-9186-2024-10-04-360-371, EDN: QWNIRH
27. **Yang Y., Gao X., Wang T., Hao X., Shi Y., Tan X., Ye X., Wang J.** Explore the LiDAR-camera dynamic adjustment fusion for 3D object detection. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 6291–6298. doi:10.48550/arxiv.2407.15334
28. **Daou S., Ben-Hamadou A., Rekik A., Kallel A.** Cross-attention fusion of visual and geometric features for large-vocabulary Arabic lipreading. *Technologies*, 2025, vol. 13, no. 1. doi:10.3390/technologies13010026
29. **Shi H., Wang X., Zhao J., Hua X.** A cross-modal attention-driven multi-sensor fusion method for semantic segmentation of point clouds. *Sensors*, 2025, vol. 25, no. 8. doi:10.3390/s25082474
30. **Krizhevsky A., Nair V., Hinton G.** *CIFAR-10 and CIFAR-100 datasets.* https://www.cs.toronto.edu/~kriz/cifar.html (дата обращения: 9.12.2025).
31. **Simonyan K., Zisserman A.** Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. arXiv:1409.1556v6. https://doi.org/1048550/arXiv:1409.1556
32. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90
33. **Peng Z., Zheng Y., Cheng Y., Li Y.** Transformer-based LiDAR-camera fusion for semantic segmentation. *2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 2025, pp. 241–245. doi:10.1109/ICBASE66587.2025.11181322
34. **Zhou Z., Ye D., Chen W., Xie Y., Wang Y., Panqu Wang, Foroosh H.** LiDARFormer: A unified transformer-based multi-task network for LiDAR perception. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14740–14747. doi:10.1109/ICRA57147.2024.10610374
35. **Liang T., Xie H., Yu K., Xia Z., Lin Z., Wang Y., Wang B., Tang Z.** BEVFusion: A simple and robust LiDAR-camera fusion framework. *Neural Information Processing Systems (NeurIPS)*, 2022. doi:10.13140/rg.2.2.31757.20966/1

UDC 004.8
doi:10.31799/1684-8853-2026-3-35-48
EDN: XBAJEX

Method for training computer vision models based on cross-modal knowledge distillation using large visual models

A. V. Kuchkov[a], Post-Graduate Student, orcid.org/0009-0007-7508-3348
A. M. Kashevnik[a,b], PhD, Tech., Associate Professor, orcid.org/0000-0001-6503-1447, alexey.kashevnik@iias.spb.su
[a]ITMO University, 49, Kronverksky Pr., 197101, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation

Introduction: Modern methods for training multimodal computer vision models mostly utilize separate feature extraction branches with late fusion. This approach is well suited for integrating existing networks into multimodal systems; however, it is resource-intensive at runtime due to the duplication of processing branches. **Purpose:** To develop a method for building multimodal computer vision models that employs a unified representation of multimodal data in order to simplify data fusion and knowledge distillation. **Methods:** Serialization of sparse and dense data types; cross-modal knowledge distillation for computer vision architectures; utilization of visual foundation models for knowledge distillation in a serialized feature space. **Results:** We develop a method for training computer vision models based on Peano curves using knowledge distillation from large visual models. The method enables blending data of different dimensions using cross-modal attention in real time by applying one-dimensional Peano curves (Gilbert and Morton curves) to serialize multidimensional data. The proposed method demonstrates a latency of 50 ms compared to 40 ms in the single-modal mode (Point Transformer v3), indicating low overhead when using cross-modal distillation on serialized feature maps. The method has been tested in pretraining mode on the nuScenes dataset using the large DINOv3 visual model. In distillation mode, using 25% of the total dataset has yielded 79.2 mIoU compared to 82.1 mIoU on 100% of the dataset using the cosine similarity distillation error function. **Practical relevance:** The use of a serialized data representation allows for accelerated computations on sparse spatial data and makes cross-modal data blending methods less resource-intensive. **Discussion:** The proposed method enables the implementation of cross-modal attention without significant additional computational costs. However, experiments conducted to date have shown that using a serialized encoder for images is impractical due to the inherently dense structure of images. Implementing an image serialization method with a faster execution time would eliminate the need for separate encoders for point cloud and image branches, significantly simplifying the architecture..

Keywords — multimodal models, knowledge distillation, visual foundation models, Peano curves, cross-modal attention.

For citation: Kuchkov A. V., Kashevnik A. M. Method for training computer vision models based on cross-modal knowledge distillation using large visual models. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 35–48 (In Russian). doi:10.31799/1684-8853-2026-3-35-48, EDN: XBAJEX

Financial support
This work was funded by the Russian State Research FFZF-2025-0003.

References

1. Shen W., Peng Z., Wang X., Wang H., Cen J., Jiang D., Xie L., Yang X., Tian Q. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 8, pp. 9284–9305. doi:10.1109/TPAMI.2023.3246102
2. Zhu X., Goldberg A. B. *Introduction to semi-supervised learning*. Ser.: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3. Morgan & Claypool Publishers, 2009, pp. 1–130. doi:10.2200/S00196ED1V01Y200906AIM006
3. Zhang Y., Hou J. Fine-grained Image-to-LiDAR contrastive distillation with Visual Foundation Models. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. doi:10.48550/arXiv.2405.14271
4. Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., Girshick R. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
5. Siméoni O., Vo H. V., Seitzer M., Baldassarre F., Oquab M., Jose C., Khalidov V., Szafraniec M., Yi S., Ramamonjisoa M., Massa F., Haziza D., Wehrstedt L., Wang J., Darcet T., Moutakanni T., Sentana L., Roberts C., Vedaldi A., Tolan J., Brandt J., Couprie C., Mairal J., Jégou H., Labatut P., Bojanowski P. DINOv3. 2025. arXiv:2508.10104. doi:10.48550/arXiv.2508.10104
6. Kamath A., Ferret J., Pathak S., Vieillard N., Merhej R., Perrin S., Matejovicova T., Ramé A., Riviére M., Rouillard L., Mesnard T., Cideron G., Grill J.-B., Ramos S., Yvinec E., Casbon M., Pot E., Penchev I., Liu G., Visin F., Kenealy K., Beyer L., Zhai X., Tsitsulin A., Busa-Fekete R., Feng A., Sachdeva N., Coleman B., Gao Y., Mustafa B., Barr I., Parisotto E., Tian D., Eyal M., Cherry C., Peter J.-T., Sinopalnikov D., Bhupatiraju S., Agarwal R., Kazemi M., Malkin D., Kumar R., Vilar D., Brusilovsky I., Luo J., Steiner A. Gemma 3 Technical Report. 2025. arXiv:2503.19786. doi:10.48550/arXiv.2503.19786
7. Tatarnikova T. M., Mokretsov N. S. Knowledge distillation method for language models based on selective intervention in training. *Cybernetics and Systems Analysis*, 2025, vol. 150, no. 2, pp. 361–365 (In Russian). doi:10.15827/0236-235X.150.361-365
8. Wang X., Zhang X., Cao Y., Wang W., Shen C., Huang T. SegGPT: Towards segmenting everything in context. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140. doi:10.1109/ICCV51070.2023.00110
9. Zeid K. A., Yilmaz K., de Geus D., Hermans A., Adrian D., Linder T., Leibe B. DINO in the Room: Leveraging 2D foundation models for 3D segmentation. 2025. arXiv:2503.18944. doi:10.48550/arXiv.2503.18944
10. Yang M., Qi Y., Xiong B., Zhang Z., Liu Y. Modal mimicking knowledge distillation for monocular three-dimensional object detection. *Engineering Applications of Artificial Intelligence*, 2025, vol. 160, Article 111821. doi:10.1016/j.engappai.2025.111821
11. Bang G., Choi K., Kim J., Kum D., Choi J. W. RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15491–15500. doi:10.1109/CVPR52733.2024.01467
12. Wang H., Bao Y., Pan P., Li Z., Liu X., Yang R., Huang D. Multi-modal relation distillation for unified 3D representation learning. *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Cham, Springer Nature Switzerland, 2025, pp. 364–381. doi:10.1007/978-3-031-73414-4_21
13. Huang T., Dong B., Yang Y., Huang X., Lau Rynson W. H., Ouyang W., Zuo W. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22100–22110. doi:10.1109/ICCV51070.2023.02025
14. Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*; *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8748–8763. Available at: https://proceedings.mlr.press/v139/radford21a.html http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part4/op (accessed 11 June 2025).
15. Wu K., Peng H., Zhou Z., Xiao B., Liu M., Yuan L., Xuan H., Valenzuela M., Chen X., Wang X., Chao H., Hu H. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21970–21980. doi:10.1109/ICCV51070.2023.02008
16. Xu R., Xiang Z., Zhang C., Zhong H., Zhao X., Dang R., Xu P., Pu T., Liu E. SCKD: Semi-supervised cross-modality knowledge distillation for 4D radar object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 9, pp. 8933–8941. doi:10.1609/aaai.v39i9.32966
17. Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I. Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning*. In: *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8821–8831. Available at: https://proceedings.mlr.press/v139/ramesh21a.html (accessed 11 June 2025).

18. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695. doi:10.1109/cvpr52688.2022.01042
19. Xiao B., Wu H., Xu W., Dai X., Hu H., Lu Y., Zeng M., Liu C., Yuan L. Florence-2: Advancing a unified representation for a variety of vision tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, June 2024, pp. 4818–4829. doi:10.1109/CVPR52733.2024.00461
20. Wang W., Lv Q., Yu W., Hong W., Qi J., Wang Y., Ji J., Yang Z., Zhao L., Song X., Xu J., Xu B., Li J., Dong Y., Ding M., Tang J. CogVLM: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Curran Associates, Inc., 2024, pp. 121475–121499. doi:10.52202/079017-3860
21. Kim M. J., Pertsch K., Karamcheti S., Xiao T., Balakrishna A., Nair S., Rafailov R., Foster E. P., Sanketi P. R., Vuong Q., Kollar T., Burchfiel B., Tedrake R., Sadigh D., Levine S., Liang P., Finn C. OpenVLA: An open-source vision-language-action model. *Proceedings of the 8th Conference on Robot Learning*. In: *Proceedings of Machine Learning Research (PMLR)*, P. Agrawal, O. Kroemer, W. Burgard (Eds.), November 2025, vol. 270, pp. 2679–2713. Available at: https://proceedings.mlr.press/v270/kim25c.html (accessed 11 March 2025).
22. Bai S., Chen K., Liu X., Wang J., Ge W., Song S., Dang K., Wang P., Wang S., Tang J., Zhong H., Zhu Y., Yang M., Li Z., Wan J., Wang P., Ding W., Fu Z., Xu Y., Ye J., Zhang X., Xie T., Cheng Z., Zhang H., Yang Z., Xu H., Lin J. Qwen2.5-VL Technical Report. 2025. arXiv:2502.13923. doi:10.48550/arXiv.2502.13923
23. Wu X., Jiang L., Wang P.-S., Liu Z., Liu X., Qiao Y., Ouyang W., He T., Zhao H. Point Transformer V3: Simpler, faster, stronger. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4840–4851. doi:10.1109/cvpr52733.2024.00463
24. Berezhnov N. I., Sirota A. A. Improving attention mechanisms in transformer architecture in image restoration. *Computer Optics*, 2024, vol. 48, no. 5, pp. 726–733 (In Russian). doi:10.18287/2412-6179-CO-1393
25. Kutscher D., Chan D. M., Bai Y., Darrell T., Gupta R. REOrdering patches improves vision models. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. Available at: https://openreview.net/forum?id=g-56WiaXKGF (accessed 12 January 2026).
26. Ivanov V. F., Ohotnikov A. L., Gradusov A. N. Algorithm of complexing sensor data for tasks of automatic control of rolling stock. *Transport Automation Research*, 2024, vol. 10, no. 4, pp. 360–371 (In Russian). doi:10.20295/2412-9186-2024-10-04-360-371, EDN: QWNIRH
27. Yang Y., Gao X., Wang T., Hao X., Shi Y., Tan X., Ye X., Wang J. Explore the LiDAR-camera dynamic adjustment fusion for 3D object detection. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 6291–6298. doi:10.48550/arxiv.2407.15334
28. Daou S., Ben-Hamadou A., Rekik A., Kallel A. Cross-attention fusion of visual and geometric features for large-vocabulary Arabic lipreading. *Technologies*, 2025, vol. 13, no. 1. doi:10.3390/technologies13010026
29. Shi H., Wang X., Zhao J., Hua X. A cross-modal attention-driven multi-sensor fusion method for semantic segmentation of point clouds. *Sensors*, 2025, vol. 25, no. 8. doi:10.3390/s25082474
30. Krizhevsky A., Nair V., Hinton G. *CIFAR-10 and CIFAR-100 datasets*. Available at: https://www.cs.toronto.edu/~kriz/cifar.html (accessed 9 December 2025).
31. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. arXiv: 1409.1556v6. https://doi.org/1048550/ arXiv:1409.1556
32. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90
33. Peng Z., Zheng Y., Cheng Y., Li Y. Transformer-based LiDAR-camera fusion for semantic segmentation. *2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 2025, pp. 241–245. doi:10.1109/ICBASE66587.2025.11181322
34. Zhou Z., Ye D., Chen W., Xie Y., Wang Y., Panqu Wang, Foroosh H. LiDARFormer: A unified transformer-based multi-task network for LiDAR perception. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14740–14747. doi:10.1109/ICRA57147.2024.10610374
35. Liang T., Xie H., Yu K., Xia Z., Lin Z., Wang Y., Wang B., Tang Z. BEVFusion: A simple and robust LiDAR-camera fusion framework. *Neural Information Processing Systems (NeurIPS)*, 2022. doi:10.13140/rg.2.2.31757.20966/1

УДК 621.391
doi:10.31799/1684-8853-2026-3-49-62
EDN: ZJAEPG

Научные статьи
Articles

Кодовые конструкции на базе обобщенных каскадных кодов для систем связи, использующих прием на основе порядковых статистик

Д. С. Осипов[a,б], *доктор техн. наук, старший научный сотрудник, orcid.org/0000-0003-0400-7181, d_osipov@iitp.ru*
[a]*Институт проблем передачи информации им. А. А. Харкевича РАН, Б. Каретный пер., 19, стр. 1, Москва, 127051, РФ*
[б]*Национальный исследовательский университет «Высшая школа экономики», Таллинская ул., 34, Москва, 123458, РФ*

Введение: *во многих проектируемых в настоящее время и перспективных системах связи методы оценивания характеристик канала и управления мощностью сигнала, разработанные для систем связи предыдущих поколений, не могут обеспечить требуемую точность оценивания и выравнивания мощности сигналов на приемном конце. Одним из вариантов решения этой проблемы является использование методов приема на основе порядковых статистик, которые не требуют управления мощностью или оценивания. Для обеспечения требуемых вероятностных характеристик сигнально-кодовые конструкции на основе таких методов приема должны дополняться внешними кодами.* ***Цель:*** *разработать практически значимые конструкции внешних кодов, способные обеспечить как заданные вероятностные характеристики, так и сравнительно высокую скорость передачи в системе связи рассматриваемого типа.* ***Результаты:*** *разработан общий подход к выбору внешних кодов на основе теории обобщенных каскадных кодов, в рамках которого недвоичные коды в частотной области, интерпретируемые как внутренние коды, дополняются внешними кодами во временной области, параметры которых могут оптимизироваться на основе аналитического расчета и конкретных требований, обусловленных спецификой проектируемых систем. Предложены алгоритмы декодирования с «мягким» входом для внутреннего кода и тем самым для обобщенного каскадного кода в целом. Приведены примеры практически значимых кодов, полученные в результате применения предложенного подхода, основанного на совместном использовании вероятностей ошибки и отказа от декодирования внутренних кодов, полученных с помощью имитационного моделирования, и аналитического метода расчета параметров внешних кодов.* ***Практическая значимость:*** *применение предложенного подхода позволяет получать конструкции внешних кодов, оптимизированных по целому ряду критериев и пригодных для аппаратной реализации.*

Ключевые слова *— межмашинная связь, псевдослучайно переключаемые частоты, прием на основе порядковых статистик, обобщенные каскадные коды, внутренние коды, декодирование с «мягким» входом.*

Для цитирования: Осипов Д. С. Кодовые конструкции на базе обобщенных каскадных кодов для систем связи, использующих прием на основе порядковых статистик. *Информационно-управляющие системы*, 2026, № 3, с. 49–62. doi:10.31799/1684-8853-2026-3-49-62, EDN: ZJAEPG

For citation: Osipov D. S. Generalized concatenated code-based constructions for communication systems with order statistics-based reception. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 49–62 (In Russian). doi:10.31799/1684-8853-2026-3-49-62, EDN: ZJAEPG

Введение

В настоящее время многие вновь проектируемые системы связи должны работать в условиях воздействия интенсивных аддитивных помех различного типа. Это в частности справедливо для систем связи, использующих неортогональный множественный доступ, таких как системы массового межмашинного взаимодействия и системы сверхнадежной связи с низкой задержкой. Сравнительно высокая мощность помех объясняется как большим числом устройств, одновременно ведущих передачу, так и несовершенством механизмов управления мощностью, традиционно используемых в таких системах. Наряду с тем, что устройства передают сравнительно короткие сообщения (зачастую спорадически) и во многих сценариях находятся в движении, мощность и спектральный состав помехи существенно меняются во времени, что в значительной степени снижает эффективность механизмов оценивания характеристик канала. В научной литературе предложены различные методы решения проблем управления мощностью и оценивания [1–5]. В настоящей работе рассматривается принципиально иной подход: однопользовательский некогерентный прием на основе порядковых статистик, не использующий оценки характеристик канала и не требующий выравнивания мощностей сигналов на приемном конце. Системы с приемниками на основе этого принципа могут работать в условиях интенсивных аддитивных помех

различного типа, не нуждаются в механизмах управления мощностью передачи и процедурах оценивания характеристик канала, не требуют блоковой синхронизации, допускают некоординированную передачу и способны обеспечить устойчивый прием без информации о таких параметрах системы связи, как число активных пользователей в сети или отношение сигнал/шум на приемном конце. В литературе были предложены различные однопользовательские приемники на основе порядковых статистик и сигнально-кодовые конструкции для таких приемников [6–10]. Настоящая работа посвящена проблеме построения практически значимых внешних кодов для сигнально-кодовой конструкции и системы связи, описанных в [10], и оптимизации параметров таких кодов.

Система многотональной связи, использующая подкоды МДР-кодов, динамически выделяемый поддиапазон и декодирование на основе порядковых статистик: передача, прием и стратегии декодирования

Будем считать, что в рассматриваемой системе связи $U+1$ пользователей одновременно ведут передачу по каналу «вверх» и все пользователи используют одну и ту же сигнально-кодовую конструкцию. Здесь и далее будем рассматривать прием данных от одного пользователя, которого будем называть «рассматриваемым», а сигналы, переданные другими пользователями, которых будем называть «мешающими», будут рассматриваться как источник помех.

Предполагается, что каждый из пользователей передает кодовые слова $C_q(N_r, k_I, d_I)$-кода над полем $GF(q)$, который называется внешним кодом в частотной области. Рассмотрим передачу i-м пользователем t-го по счету кодового слова $\mathbf{v}_{i,t} = \left[v_{i,t,1}, \ldots, v_{i,t,N_r}\right]$. Каждый символ кодового слова кодируется отдельным кодом (эти коды будем называть внутренними кодами во временной области):

$$\mathbf{c}_{i,t,\rho} = v_{i,t,\rho} \cdot \mathbf{g}_\rho + \beta^{b_\rho} \cdot \mathbf{g}_{N_r+1},$$

где $\mathbf{g}_\rho$ — строка порождающей матрицы вида

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_{N_r} \\ \mathbf{g}_{N_r+1} \end{bmatrix} = \begin{bmatrix} \beta^0 & \beta^0 & \cdots & \beta^0 \\ \beta^1 & \beta^2 & \cdots & \beta^n \\ \vdots & \vdots & \ddots & \vdots \\ \beta^{N_r} & \beta^{2\cdot N_r} & \cdots & \beta^{(N_r\cdot n)} \end{bmatrix};$$

$b_1, \ldots, b_{N_r}$ – натуральные числа такие, что $\forall\ s, \rho$: $1 \le \rho \le N_r$, $1 \le s \le N_r$, $1 \le b_s \le n$, $1 \le b_\rho \le n$, $b_s \ne b_\rho$; β — примитивный элемент поля.

Таким образом, каждый из N_r внутренних кодов во временной области является подкодом $C_q(n, N_r+1, n-N_r)$-кода, полученного выкалыванием из $C_q(q, N_r+1, q-N_r)$-кода Рида — Соломона.

Затем каждый символ каждого из N_r кодовых слов внутренних кодов во временной области отображается в вектор-кодовое слово кода, который взаимно однозначно сопоставляет каждому символу поля вектор веса 1 (такой нелинейный равновесный эквидистантный код называется внутренним кодом в частотной области). В результате каждое слово внутреннего кода во временной области (и, соответственно, каждый символ внешнего кода в частотной области) отображается в двоичную матрицу $\mathbf{M}_{i,t,\rho}$, в каждом столбце которой стоит единственный ненулевой элемент. Все полученные матрицы складываются в поле действительных чисел, полученная в результате матрица имеет вид $\mathbf{M}_{i,t} = \sum_{\rho=1}^{N_r} \mathbf{M}_{i,t,\rho}$.

Именно эта матрица размера $q \times n$ будет передаваться рассматриваемым пользователем по каналу связи. Для этого пользователь дописывает к полученной матрице матрицу из всех нулей $\mathbf{Z}$ размера $(Q-q) \times n$. Каждый (пусть для определенности s-й) столбец полученной матрицы

$$\tilde{\mathbf{M}}_{i,t} = [\tilde{\mathbf{m}}_{i,t,1}, \tilde{\mathbf{m}}_{i,t,2}, \ldots, \tilde{\mathbf{m}}_{i,t,n}] = \begin{bmatrix} \mathbf{M}_{i,t} \\ \mathbf{Z} \end{bmatrix}$$

подвергается псевдослучайной перестановке $\pi_{i,s}$ (где s — номер символов внутренних кодов во временной области, все перестановки независимы). Каждый столбец полученной матрицы

$$\breve{\mathbf{M}}_{i,t} = \left[\pi_{1,1}\left(\tilde{\mathbf{m}}_{i,t,1}\right), \pi_{1,2}\left(\tilde{\mathbf{m}}_{i,t,2}\right), \ldots, \pi_{1,N}\left(\tilde{\mathbf{m}}_{i,t,n}\right)\right]$$

последовательно передается в канал, причем каждый символ передается на отдельной поднесущей. Таким образом, для передачи сообщений в описываемой системе связи требуется метод передачи, обеспечивающий возможность передачи на Q независимых поднесущих. Здесь и далее без потери общности будем полагать, что для решения этой задачи используется OFDM с циклическим префиксом. Совокупность поднесущих, которые соответствуют первым q строкам матрицы $\tilde{\mathbf{M}}_{i,t}$ (матрице $\mathbf{M}_{i,t}$), т. е. тем поднесущим, которые могут использоваться для передачи s-х символов слов внутреннего кода во временной области, будем называть частотным поддиапазоном. Так как благодаря псевдослучайным перестановкам номера поднесущих, входящих в частотный поддиапазон, в котором передаются s-е символы, меняются при передаче символов для каждого нового значения s, как и в работе [11], будем говорить, что символы, соответствующие внутренним

кодам во временной области, передаются в динамически выделяемом частотном поддиапазоне. Каждый ненулевой символ матрицы $\mathbf{M}_{i,t}$ (и, следовательно, матриц $\tilde{\mathbf{M}}_{i,t}$ и $\bar{\mathbf{M}}_{i,t}$) формирует сигнал на одной из поднесущих, поэтому внутренний код в частотной области является, по существу, способом отображения символов внутреннего кода во временной области на созвездие частотно-позиционной модуляции.

При приеме для каждого принятого из канала вектора-столбца высоты Q выполняется обратная перестановка $\pi_{i,s}^{-1}$. В результате измерения, соответствующие сигналам, переданным рассматриваемым пользователем, оказываются в первых q строках. Подматрица, соответствующая этим строкам, задается выражением вида

$$\tilde{\mathbf{Y}}_{i,t} = \mathbf{M}_{i,t} * \mathbf{A}_{i,t} * \mathbf{\Psi}_{i,t} + \sum_{k\neq i, i=1}^{U+1} \mathbf{W}_{i,k,t} * \mathbf{A}_{k,t} * \mathbf{\Psi}_{k,t} + \boldsymbol{\eta}.$$

Здесь $*$ — знак произведения Адамара (Шура); $\mathbf{A}_{i,t}$ — матрица амплитуд коэффициентов передачи канала «вверх» от i-го пользователя к центральному узлу; $\mathbf{\Psi}_{i,t}$ — матрица вида

$$\mathbf{\Psi}_{i,t} = \begin{bmatrix} e^{j\psi_{1,1}} & \dots & e^{j\psi_{1,n}} \\ \vdots & \ddots & \vdots \\ e^{j\psi_{q,1}} & \dots & e^{j\psi_{q,n}} \end{bmatrix},$$

где j — мнимая единица и $\psi_{a,b}$ — случайная величина, равномерно распределенная на окружности $[0, 2\pi]$; $\boldsymbol{\eta}$ — матрица, соответствующая отсчетам фонового аддитивного шума; $\mathbf{W}_{i,k,t}$ — подматрица, соответствующая q первым строкам матрицы

$$\tilde{\mathbf{W}}_{i,k,t} = \left[\pi_{i,1}^{-1}\left(\pi_{k,\omega_1}\left(\tilde{\mathbf{m}}_{k,T(t,\tau_k,1),S(t,\tau_k,1)}\right)\right), \dots \right.$$
$$\left. \dots, \pi_{i,n}^{-1}\left(\pi_{k,\omega_n}\left(\tilde{\mathbf{m}}_{k,T(t,\tau_k,n),S(t,\tau_k,n)}\right)\right)\right],$$

где ω_s — номер перестановки, которая использовалась при передаче столбца, соответствующего $S(t, \tau_k, s)$-му символу $T(t, \tau_k, s)$-го кодового слова k-го пользователя:

$$S(t,\tau_k,s) = (n + s - \tau_k + 1) \bmod n + 1;$$

$$T(t,\tau_k,s) = \begin{cases} t & \tau_k = 0 \\ t & s \le |\tau_k|, \tau_k < 0 \\ t+1 & s > |\tau_k|, \tau_k < 0 \\ t-1 & s \le |\tau_k|, \tau_k > 0 \\ t & s > |\tau_k|, \tau_k > 0 \end{cases},$$

τ_k — задержка сигнала от k-го «мешающего» пользователя относительно сигнала от рассматриваемого (i-го) пользователя; s — номер символа t-го кодового слова.

Приемник извлекает и возводит каждый элемент матрицы в квадрат, вычисляя энергии принятых сигналов. Каждый столбец полученной матрицы $\mathbf{\Omega}_{i,t} = \tilde{\mathbf{Y}}_{i,t} \cdot \tilde{\mathbf{Y}}_{i,t}^{H}$ сортируется по убыванию. Обозначим матрицу, полученную сортировкой столбцов матрицы $\mathbf{\Omega}_{i,t}$ (по убыванию), как $\mathbf{\Omega}_{i,t}^{\downarrow}$. Используя матрицу $\mathbf{\Omega}_{i,t}^{\downarrow}$, приемник вычисляет $\mathbf{D}_{i,t}^{\alpha}$ — матрицу оценок достоверности каждого символа кодового слова как

$$\mathbf{D}_{i,t}^{\alpha}(s,z) = \begin{cases} \lambda_1 & \mathbf{\Omega}_{i,t}(s,z) \ge \mathbf{\Omega}_{i,t}^{\downarrow}(\alpha,z) \\ \lambda_0 & \mathbf{\Omega}_{i,t}(s,z) < \mathbf{\Omega}_{i,t}^{\downarrow}(\alpha,z) \end{cases},$$

где λ_1 и λ_0 — заранее выбранные действительные константы ($0 \le \lambda_0 \le \lambda_1$); $\mathbf{\Omega}_{i,t}^{\downarrow}(\alpha,z)$ — элемент матрицы $\mathbf{\Omega}_{i,t}^{\downarrow}$, стоящий на пересечении строки с номером α и столбца с номером z (т. е. элемент с номером α в вариационном ряду, составленном путем упорядочивания элементов столбца с номером z матрицы $\mathbf{\Omega}_{i,t}$ по убыванию); α — заранее выбранное натуральное число ($1 < \alpha < q$). Так как каждый символ кодового слова внешнего кода кодируется отдельным внешним кодом, приемник использует N_r декодеров, при этом входом каждого из декодеров является матрица оценок достоверности $\mathbf{D}_{i,t}^{\alpha}$. В работе [11] показано, что вероятностные характеристики декодеров не зависят от выбора констант λ_0 и λ_1, поэтому здесь и далее будем полагать, что используются наиболее удобные с точки зрения аппаратной реализации значения $\lambda_0 = 0$ и $\lambda_1 = 1$.

Пусть каждое слово каждого из внутренних кодов во временной области пронумеровано числами от 1 до q. Так как внутренний код в частотной области задает взаимно однозначное соответствие, в котором каждому символу конечного поля ставится в соответствие единственный двоичный вектор веса 1, каждому слову с номером m внутреннего кода во временной области с номером ρ соответствует единственная матрица $\mathbf{X}_{\rho,m}$. Каждый декодер вычисляет решающие статистики

$$S_{\rho,m}^{i,t} = \sum_{s=1}^{q}\sum_{z=1}^{n} \mathbf{D}_{i,t}^{\alpha}(s,z) * \mathbf{X}_{\rho,m}(s,z)$$

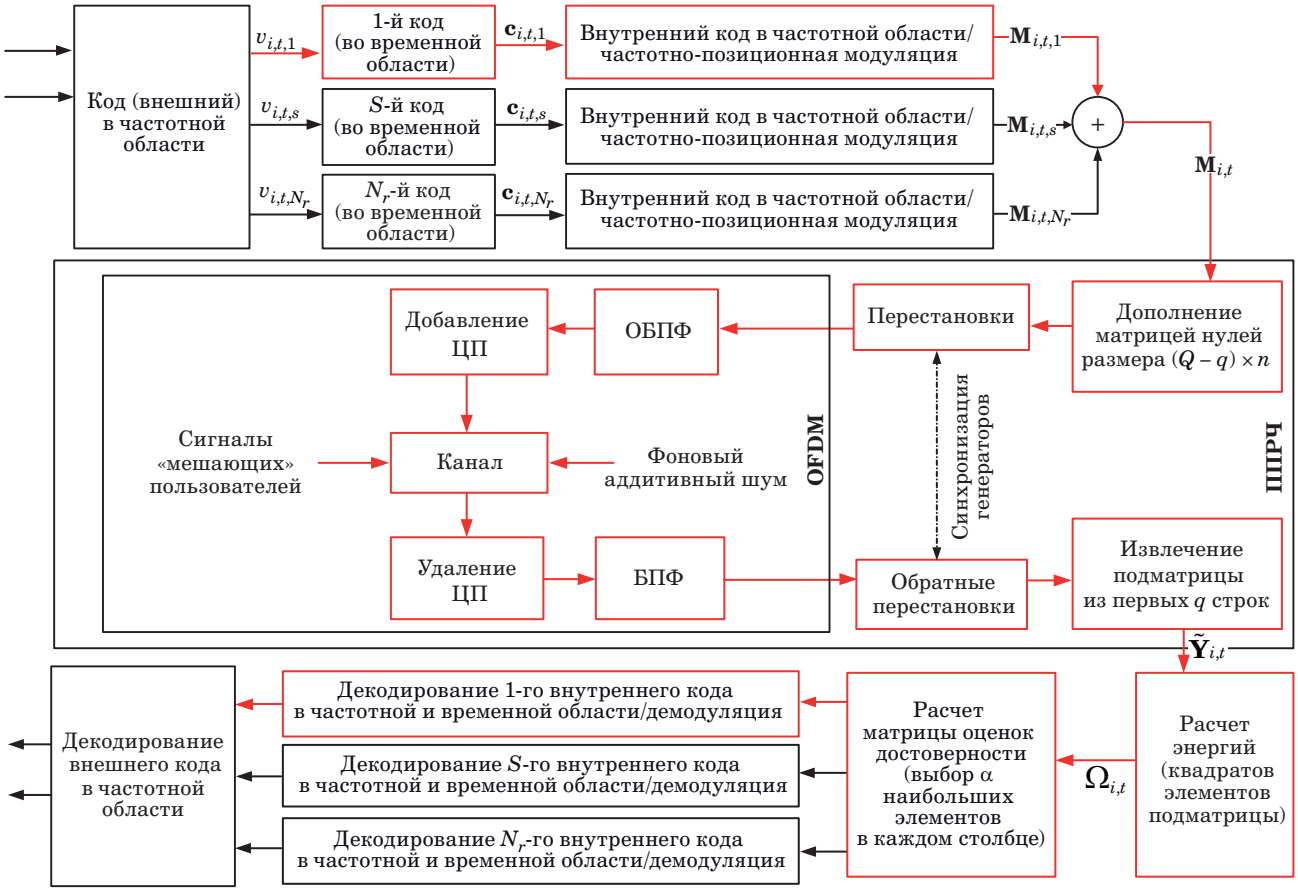
и возвращает вектор $\mathbf{f}_\rho = [f_\rho(1), f_\rho(2), \dots, f_\rho(q)]^{\mathrm{T}}$, где

$$f_\rho(m) = \begin{cases} 1 & S_{\rho,m}^{i,t} = \max\limits_{m}\left(S_{\rho,m}^{i,t}\right) \\ 0 & S_{\rho,m}^{i,t} \ne \max\limits_{m}\left(S_{\rho,m}^{i,t}\right) \end{cases}.$$

Описанная выше система, в которой символы кодовых слов различных внутренних кодов во временной области с одинаковым номером передаются на нескольких, вообще говоря, разных

поднесущих в одном и том же частотном поддиапазоне, является модификацией и обобщением одночастотной системы связи, описанной в [11], в которой в каждом OFDM-символе передавался символ только одного кодового слова внутреннего кода во временной области. Одночастотная система связи [11] в свою очередь представляла собой, по существу, модификацию системы связи на основе метода псевдослучайной перестройки рабочей частоты (ППРЧ, Frequency hopping) [12], которая отличалась от традиционной системы на базе ППРЧ динамическим выделением поддиапазонов, а также методом расчета оценок достоверности и способом декодирования (демодуляции) на основе вычисленных таким образом оценок. В частности, операция дополнения матрицы $\mathbf{M}_{i,t}$ матрицей $\mathbf{Z}$ размера $(Q-q)\times n$ и последующие перестановки эквивалентны отображению элементов на случайно выбранные поднесущие, т. е. динамическому формированию частотного поддиапазона, о котором говорилось выше.

На схеме приема и передачи в системе рассматриваемого типа (рис. 1) красным показаны блоки, сигналы и потоки данных, соответствующие одночастотной системе из работы [11]. Система рассматриваемого типа отличается по существу тем, что в ней каждый пользователь передает несколько символов, каждому из которых соответствует свой внутренний код во временной области. Именно по этой причине в предлагаемой многочастотной конструкции используется несколько различных внутренних кодов во временной области: помимо минимизации влияния многопользовательской помехи (эту функцию, как и в [11], выполняют совместно внутренний код во временной области, внутренний код в частотной области и псевдослучайные перестановки), выбранные описанным выше способом внутренние коды также позволяют восстановить символы, кодируемые различными внутренними кодами, с использованием одной и той же матрицы оценок достоверности. Внешний код в частотной



Используемые сокращения:

ЦП – циклический префикс; ОБПФ – обратное быстрое преобразование Фурье; БПФ – быстрое преобразование Фурье; ППРЧ – система связи, использующая метод псевдослучайной перестройки рабочей частоты; OFDM – Orthogonal Frequency Division Multiplexing (мультиплексирование с использованием ортогональных частот)

■ ***Рис. 1.*** Передача и прием в системе связи рассматриваемого типа

■ ***Fig. 1.*** Transmission and reception in the communication system under consideration

области необходим для исправления стираний и исправления или обнаружения ошибок, которые возникают в результате отказа от декодирования или ошибочного декодирования внутренних кодов во временной области соответственно.

В работе [9] использовалось декодирование внутренних кодов с «жестким» выходом, для каждого из векторов декодер внутреннего кода принимал решение, руководствуясь следующим правилом: если в векторе была единственная ненулевая координата, декодер принимал решение о том, что на позиции с номером стоял символ, соответствующий номеру этой ненулевой координаты, в противном случае принимал решение о стирании, после чего результаты декодирования передавались на вход декодера внешнего кода. В работе [10] обсуждались алгоритмы, которые можно интерпретировать как алгоритмы декодирования внешнего кода в частотной области с «мягким» входом. Такие алгоритмы используют оценки достоверности, которые хранятся в виде векторов $\mathbf{f}_\rho$ непосредственно, минуя фазу жесткого решения, и обеспечивают существенно лучшие вероятностные характеристики [10]. Здесь обратимся к следующему алгоритму декодирования.

Алгоритм 1. Алгоритм декодирования кодов в частотной области с «мягким» входом.

Декодер получает на вход матрицу $\mathbf{F} = \left[\mathbf{f}_1, \ldots, \mathbf{f}_{N_r}\right]$ из N_r двоичных векторов высоты q (где $\mathbf{f}_\rho = [f_\rho(1), f_\rho(2), \ldots, f_\rho(q)]^{\mathrm{T}}$), порождающую матрицу $\mathbf{G}_I$ используемого систематического кода и максимально допустимый размер списка L_M.

1.1. Декодер вычисляет число векторов, которые подлежат проверке: $L_o = \prod_{i=1}^{k_I} \sum_{t=1}^{q} f_i(t)$, и сравнивает это число с максимально допустимым размером списка.

В случае, если $L_o > L_M$, принимается решение об отказе от декодирования, в противном случае декодер переходит к шагу 1.2.

1.2. Декодер строит список информационных последовательностей, которым соответствуют ненулевые элементы в первых k_I столбцах, т. е. множество $\mathbf{U}$ таких информационных последовательностей, для которых выполняется

$$\forall\, \mathbf{u}_g = [u_{g,1}, \ldots, u_{g,k_I}] \in \mathbf{U},\ \forall s \in \{1, \ldots, k_I\},$$
$$f_s\left(\phi_I\left(u_{g,s}\right)\right) \equiv 1.$$

1.3. Декодер кодирует все последовательности из множества и проверяет выполнение условия

$$\forall \mathbf{u}_g \in \mathbf{U},\ \ \mathbf{v}_g = [v_{g,k_I+1}, \ldots, v_{g,N_r}] = \mathbf{u}_g \mathbf{G}_I,$$
$$\forall s \in \{k_I, \ldots, N_r + 1\},\ \ f_s\left(\phi_I\left(v_{g,s}\right)\right) \equiv 1,$$

формируя Λ_v^* — список кодовых последовательностей.

1.4. Если список Λ_v^* содержит только одно слово, декодер объявляет это слово результатом декодирования, в противном случае принимает решение об отказе.

Как уже отмечалось, использование алгоритмов с «мягким» входом существенно улучшает вероятности ошибки и отказа внешнего кода [10]. Однако многие реальные приложения требуют гораздо лучших вероятностных характеристик, чем те, которые обеспечивают внешние коды в частотной области, и, кроме того, используют более длинные сообщения, поэтому вполне естественной представляется идея дополнить описанную выше конструкцию внешним кодом во временной области. Здесь для решения этой задачи предлагается использовать подход на основе обобщенных каскадных кодов (ОКК), допускающих гибкий выбор комбинаций внешних и внутренних кодов и позволяющих аналитически подбирать параметры внешних кодов в зависимости от конкретных сценариев и требований. Ниже, в частности, будет предложена модификация алгоритма декодирования ОКК на случай мягкого декодирования внутренних кодов.

Обобщенные каскадные коды: кодирование и декодирование

Линейные обобщенные каскадные коды были предложены Э. Л. Блохом и В. В. Зябловым [13]. Здесь ограничимся представлением принципов кодирования и декодирования линейных ОКК в объемах, необходимых для описания конструкции, предлагаемой в настоящей работе, и методики расчета параметров кода и скоростей передачи, которых применение кода с такими параметрами позволяет достигать.

Для построения линейного ОКК используем систему заданных в систематической форме вложенных внутренних кодов $C_{I,l}$ ($l = 0 \ldots M-1$), для которых выполняется условие $C_{I,M-1} \subset C_{I,M-2} \subset \ldots C_{I,1} \subset C_{I,0}$ над полем $GF(q)$ с параметрами $C_q(N_I, k_{I,l})$, где

$$k_{I,l} = \begin{cases} N_I - r_0 & l = 0 \\ N_I - \sum_{l=0}^{l-1} \kappa_l - r_0 & l > 0 \end{cases},$$

и набор заданных в систематической форме внешних кодов, где каждый внешний код с параметрами $C_{q^{\kappa_l}}\left(N_O, K_{O,l}\right)$, соответствующий l-му слою, определен над расширением поля $GF(q)$ степени κ_l.

Слово ОКК представим в виде матрицы $\mathbf{c}$ из N_I строк и N_O столбцов над полем $GF(q)$.

Положим, что матрица **c** разбита на подматрицы, соответствующие проверкам и информационным символам слоев, которые пронумеруем последовательно, начиная с нижних строк. Для определенности будем считать, что информационные символы внешних кодов всегда располагаются в левых частях соответствующих подматриц, а позиции, соответствующие информационным символам внутренних кодов, — в верхней части матрицы. Рассмотрим частичные матрицы $\tilde{\mathbf{c}}_l$, каждая из которых соответствует процедуре кодирования l-го слоя, отдельно. С учетом принятых ранее соглашений и обозначений нижняя подматрица матрицы $\tilde{\mathbf{c}}_l$, соответствующая последним

$$r_l = \begin{cases} r_0 & l = 0 \\ r_0 + \sum_{m=0}^{l-1} \kappa_m & l > 0 \end{cases}$$

строкам, отводится под проверки внутреннего кода l-го слоя, следующие κ_l строк отводятся под символы внешнего кода l-го слоя, а оставшиеся $N_O - \kappa_l - r_l$ строк в верхней части матрицы $\tilde{\mathbf{c}}_l$ заполняются нулевыми символами из поля. Процесс кодирования для каждой частичной подматрицы, соответствующей l-му слою, выглядит следующим образом: сначала заполняются K_l столбцов в левой части подматрицы, отведенной под слово внешнего кода, соответствующие информационным символам внешнего кода, затем кодер внешнего кода вычисляет соответствующие этим информационным символам проверки кодового слова внешнего кода, заполняя оставшиеся в правой части подматрицы позиции, наконец, кодер внутреннего кода выполняет кодирование внутреннего кода по каждому столбцу, используя $k_{I,l} = N_I - \sum_{m=0}^{l-1} \kappa_m - r_0$ элементов в верхней части каждого столбца как информационные символы и заполняя $r_{I,l} = \sum_{m=0}^{l-1} \kappa_m + r_0$ символов в нижней части столбца, соответствующих проверкам внутреннего кода. Таким образом, каждая частичная матрица $\tilde{\mathbf{c}}_l$ представляет собой кодовое слово каскадного кода, в котором слову внешнего кода соответствует подматрица специального вида, описанная выше, а роль внутреннего кода играет код с параметрами $C_q(N_I, k_{I,l})$. Словом ОКК **c** будем называть сумму всех частичных матриц $\tilde{\mathbf{c}}_l$ над полем $GF(q)$: $\mathbf{c} = \sum_{l=0}^{M-1} \tilde{\mathbf{c}}_l$.

Рассмотрим теперь алгоритм декодирования. Классический алгоритм декодирования l-го слоя ОКК выглядит следующим образом.

Алгоритм 2. Алгоритм декодирования l-го слоя линейного ОКК с «жестким» входом.

Алгоритм получает на вход матрицу

$$\mathbf{y}_l = \mathbf{c}_l + \mathbf{e} = \sum_{m=l}^{M-1} \tilde{\mathbf{c}}_m + \mathbf{e}.$$

2.1. Декодер внутренних кодов выполняет жесткое декодирование внутренних кодов, вычисляя матрицы $\hat{\mathbf{d}}_l = \hat{\mathbf{c}}_l + \hat{\mathbf{e}}_l$ (матрицу, соответствующую значениям кодовых слов после декодирования внутренних кодов) и $\mathbf{\Phi}_l$ (матрицу, в которой столбцам, при декодировании которых было принято решение об отказе, соответствуют единичные столбцы, а столбцам, для которых декодирование завершилось успешно, — нулевые). Результат можно записать как

$$\mathbf{d}_l = \theta_x\left(\mathbf{\Phi}_l, \hat{\mathbf{c}}_l + \hat{\mathbf{e}}_l\right), \qquad (1)$$

где $\theta_x()$ — матричная функция, позволяющая учесть стирания, введенные в результате отказа внутренних кодов от декодирования. Структура кода позволяет выполнять эту операцию независимо для каждого внутреннего кода, каждого столбца, поэтому этот шаг можно распараллелить в N_O раз.

2.2. Декодер вычисляет матрицы, соответствующие h-му слою, кодируя верхние $k_{I,h} = N_O - \sum_{m=0}^{h-1} \kappa_m - r_0$ строк внутренним $C_q(N_I, k_{I,h})$-кодом, и вычитает их из текущей матрицы последовательно для h от $M-1$ до $l+1$:

$$\hat{\mathbf{d}}_l = \breve{\mathbf{d}}_{l,l+1} = \mathbf{d}_l - \sum_{h=l+1}^{M-1} \tilde{\mathbf{c}}_{l,h}. \qquad (2)$$

Операции кодирования внутренними кодами и вычитания также происходят независимо для всех столбцов матрицы и могут выполняться параллельно.

2.3. Декодер внешнего $C_{q^{\kappa_l}}(N_O, K_{O,l})$-кода выполняет декодирование внешнего кода, используя информацию, содержащуюся в соответствующей ему подматрице, и выполняя исправление ошибок и стираний. В случае, если декодер принимает решение об отказе от декодирования, принимается решение об отказе от декодирования всего слова ОКК в целом. В противном случае декодер переходит к следующему шагу.

2.4. Кодируя матрицу, состоящую из $\sum_{s=l+1}^{M-1} \kappa_s$ нулевых строк, соответствующих слоям с $(l+1)$-го до $(M-1)$-го, и κ_l строк, соответствующих кодовому слову внешнего кода l-го слоя, декодер получает матрицу $\hat{\mathbf{c}}_l$ — оценку матрицы $\tilde{\mathbf{c}}_l$, соответствующую l-му слою, вычитает ее из

матрицы $\mathbf{y}_l$ и передает полученную матрицу $\mathbf{y}_{l+1} = \sum_{m=l}^{M-1} \tilde{\mathbf{c}}_m + \mathbf{e} - \hat{\mathbf{c}}_l$ декодеру следующего слоя.

Особенности рассматриваемой конструкции и модифицированный алгоритм декодирования линейных обобщенных каскадных кодов

В настоящей работе рассмотрена конструкция, в которой каждый вектор, поступающий на вход кодера внутреннего кода во времени, является столбцом матрицы, представляющей собой слово ОКК. Таким образом, каждый столбец матрицы (слова внешнего ОКК) является, по существу, кодом в частотной области.

Алгоритмы с «мягким» входом такие, как описанный здесь алгоритм 1, обеспечивают лучшие вероятностные характеристики, чем традиционные алгоритмы декодирования с «жестким» входом [10]. Классический алгоритм декодирования линейных ОКК, описанный выше, подразумевает именно жесткое декодирование внутренних кодов. Задача разработки алгоритма декодирования ОКК, в котором для декодирования внутренних кодов использовались методы с «мягким» входом, была изучена в работах [14–16], однако в этих работах рассматривалось декодирование для канала с аддитивным гауссовым шумом, в частности использование алгоритма Чейза для декодирования двоичных кодов или стек-алгоритма на синдромных решетках надкодов. В случае, который исследуется в настоящей работе, эти подходы неприменимы. По этой причине ниже будет предложена модифицированная версия алгоритма 1, адаптированная для декодирования линейных ОКК, а затем и модифицированный алгоритм декодирования линейных ОКК, использующий мягкое декодирование внутренних кодов.

Алгоритм декодирования кодов в частотной области, описанный выше, выдает жесткое решение, т. е., как и в случае с классическими алгоритмами декодирования, результатом его работы является либо единственное кодовое слово, либо решение об отказе, поэтому шаги 2.2 и 2.3 алгоритма 2 могут быть оставлены без изменений. Однако на вход декодера ОКК в рассматриваемом случае подаются оценки достоверности, а не значения символов, как в классическом случае, поэтому шаг 2.4 классического алгоритма не может быть реализован в том виде, в котором он описан в алгоритме 2. Ниже будет приведена модификация алгоритма декодирования внутреннего кода, которая позволяет объединить шаг 2.4 при декодировании l-го слоя и шаг 2.1 в процедуре декодирования $(l + 1)$-го слоя.

Для того чтобы модифицировать процедуру декодирования для интересующего нас случая, достаточно заметить, что по построению каждый вектор (он же t-й столбец матрицы-кодового слова ОКК), передаваемый по каналу способом, описанным в разделе «Система многотональной связи, использующая подкоды МДР-кодов, динамически выделяемый поддиапазон и декодирование на основе порядковых статистик: передача, прием и стратегии декодирования», является кодовым словом внутреннего кода $\mathbf{v}^*_{0,t} = \sum_{l=0}^{M-1} \mathbf{v}_{l,t}$.

На каждом шаге декодирования ОКК, соответствующем g-му слою, декодеру известны кодовые слова, соответствующие предыдущим $g-1$ слоям. Таким образом, декодер каждый раз ищет слово

$$\mathbf{v}^*_{0,t} = \sum_{l=0}^{g-1} \mathbf{v}_{l,t} + \sum_{l=g}^{M-1} \mathbf{v}_{l,t},$$

причем первое слагаемое известно декодеру. Второе слагаемое, т. е. вектор

$$\mathbf{v}^*_{g,t} = \sum_{l=g}^{M-1} \mathbf{v}_{l,t},$$

является кодовым словом внутреннего вложенного кода, используемого для кодирования g-го слоя. Алгоритм 3 учитывает структуру кодовых слов в линейном ОКК, используя информацию о результатах декодирования предшествующего слоя для декодирования внутреннего кода нового слоя.

Алгоритм 3. Алгоритм декодирования внутренних кодов l-го слоя линейного ОКК с «мягким» входом для оценок достоверности рассматриваемого типа.

Декодер получает на вход кодовое слово кода предыдущего уровня $\mathbf{v}_{l-1} = \left[v_{l-1,1}, \ldots, v_{l-1,N_r} \right]$, матрицу $\mathbf{F} = \left[\mathbf{f}_1, \ldots, \mathbf{f}_{N_r} \right]$ из N_r двоичных векторов высоты q (где $\mathbf{f}_s = [f_s(1), f_s(2), \ldots, f_s(q)]^{\mathrm{T}}$), порождающую матрицу $\mathbf{D}_{I,l}$ используемого систематического кода l-го уровня и максимально допустимый размер списка L_M.

3.1. Декодер вычисляет число векторов, которые подлежат проверке: $L_o = \prod_{i=1}^{k_{I,l}} \sum_{t=1}^{q} f_i(t)$, и сравнивает это число с максимально допустимым размером списка.

В случае, если $L_o > L_M$, принимается решение об отказе от декодирования, в противном случае декодер переходит к шагу 3.2.

3.2. Декодер строит список информационных последовательностей, которым соответствуют ненулевые элементы в первых $k_{I,l}$ столбцах, т. е.

множество $\mathbf{U}$ таких информационных последовательностей, для которых

$$\forall\, \mathbf{u}_g = [u_{g,1}, \ldots, u_{g,k_I}] \in \mathbf{U},\ \forall s \in \{1, \ldots, k_{I,l}\},$$
$$f_s\left(\phi_I\left(u_{g,s} + v_{l-1,s}\right)\right) \equiv 1.$$

3.3. Декодер кодирует все последовательности из множества и проверяет выполнение условия

$$\forall \mathbf{u}_g \in \mathbf{U},$$
$$\mathbf{v}_g = [v_{g,k_I+1}, \ldots, v_{g,N_r}] = \mathbf{u}_g \mathbf{G}_{I,l},$$
$$\forall s \in \{k_{I,l}, \ldots, N_r + 1\},\ f_s\left(\phi_I\left(v_{g,s} + v_{l-1,s}\right)\right) \equiv 1,$$

формируя Λ_v^* — список кодовых последовательностей.

3.4. Если список Λ_v^* содержит только одно слово, декодер объявляет это слово результатом декодирования, в противном случае принимает решение об отказе.

Теперь приведем формальное описание модифицированного алгоритма декодирования линейного ОКК.

Алгоритм 4. Модифицированный алгоритм декодирования l-го слоя ОКК.

Декодер получает на вход матрицу, столбцами которой являются кодовые слова внутреннего кода, полученные в результате декодирования предыдущего слоя $\left(\mathbf{v}_{l-1} = \left[v_{l-1,1}, \ldots, v_{l-1,N_r}\right]\right.$, в частности при декодировании 0-го слоя $\left.\mathbf{v}_{-1} = \left[\bar{\mathbf{0}}, \ldots, \bar{\mathbf{0}}\right]\right)$, трехмерную матрицу $\Theta = [\mathbf{F}_1, \ldots, \mathbf{F}_N]$, где $\mathbf{F}_s$ — матрица оценок достоверности для s-го столбца, порождающую матрицу $\mathbf{D}_{I,l}$ используемого на l-м слое систематического кода внутреннего кода и максимально допустимый размер списка L_M.

4.1. Выполняет декодирование для каждого (s-го) столбца, используя матрицу оценок достоверности $\mathbf{F}_s$, кодовое слово, соответствующее этому столбцу при декодировании предыдущего ((l – 1)-го) слоя $\mathbf{v}_{s,l-1}$, порождающую матрицу $\mathbf{D}_{I,l}$, максимально допустимый размер списка L_M и алгоритм 3, вычисляя матрицы $\hat{\mathbf{d}}_l = \hat{\mathbf{c}}_l + \hat{\mathbf{e}}_l$ (матрицу, соответствующую значениям кодовых слов после декодирования внутренних кодов) и $\mathbf{\Phi}_l$ (матрицу, в которой столбцам, при декодировании которых было принято решение об отказе, соответствуют единичные столбцы, а столбцам, для которых декодирование завершилось успешно, — нулевые). Результирующую матрицу можно записать аналогично (1).

Структура кода позволяет выполнять эту операцию независимо для каждого внутреннего кода, каждого столбца, поэтому этот шаг можно распараллелить в N_O раз.

4.2. Декодер вычисляет матрицы, соответствующие h-му слою, кодируя верхние $k_{I,h} = N_O - \sum_{m=0}^{h-1} \kappa_m - r_0$ строк внутренним $C_q(N_I,\ k_{I,h})$-кодом, и вычитает их из текущей матрицы последовательно для h от $M-1$ до $l+1$ по формуле (2).

Операции кодирования внутренними кодами и вычитания также происходят независимо для всех столбцов матрицы и могут выполняться параллельно.

4.3. Декодер внешнего $C_{q^{\kappa_l}}(N_O, K_{O,l})$-кода выполняет декодирование внешнего кода, используя информацию, содержащуюся в соответствующей ему подматрице, и выполняя исправление ошибок и стираний. В случае, если декодер принимает решение об отказе от декодирования, принимается решение об отказе от декодирования всего слова ОКК в целом. В противном случае декодер переходит к следующему шагу.

4.4. Кодируя матрицу, состоящую из $\sum_{s=l+1}^{M-1} \kappa_s$ нулевых строк, соответствующих слоям с (l + 1)-го до (M – 1)-го, и κ_l строк, соответствующих кодовому слову внешнего кода l-го уровня, внутренним кодом, декодер получает матрицу $\hat{\mathbf{c}}_l = \left[\mathbf{v}_{1,l}, \ldots, \mathbf{v}_{N,l}\right]$ — оценку матрицы $\tilde{\mathbf{c}}_l$, соответствующую l-му слою, и передает ее декодеру следующего слоя.

Особенности и параметры кодовых конструкций рассматриваемого типа и их исследование

Здесь и далее будем рассматривать конструкцию, в которой все внешние коды определяются над тем же полем $GF(q)$, что и внутренние, т. е. $\kappa_l = 1$. В качестве внутреннего кода, соответствующего 0-му слою, будем всегда использовать код с проверкой на четность длины $N_I = N_r$, а коды последующих уровней выбирались как подкоды кода каждого предыдущего уровня. Таким образом, внутренний код каждого (l-го) слоя является $C_q(N_r,\ N_r - l - 1,\ l + 2)$ МДР-кодом (кодом с максимально достижимым расстоянием). Для того чтобы вложенные коды были систематическими, использован следующий подход.

Пусть $\mathbf{G}_{I,l} = \begin{bmatrix} \mathbf{g}_{l,1}^{\mathrm{T}} \\ \vdots \\ \mathbf{g}_{l,N_r-l-1}^{\mathrm{T}} \end{bmatrix}$ — порождающая матрица $C_q(N_r, N_r - l - 1, l + 2)$-кода. Тогда $\mathbf{G}_{I,l+1} = \begin{bmatrix} \mathbf{g}_{l+1,1}^{\mathrm{T}} \\ \vdots \\ \mathbf{g}_{l+1,N_r-l-2}^{\mathrm{T}} \end{bmatrix}$ — порождающая матрица

внутреннего кода каждого ((l + 1)-го) слоя, где каждый вектор-строка определяется как $\mathbf{g}_{l+1,m}^{\mathrm{T}} = \mathbf{g}_{l,m}^{\mathrm{T}} + \beta^{2m+1} \cdot \mathbf{g}_{l,N_r-l-1}^{\mathrm{T}}$.

Последнему слою соответствует $C_q(N_r, 1, N_r)$ обобщенный код с повторением.

В качестве внешних кодов будем использовать $C_q(N_O, K_{O,l}, N_O - K_{O,l} + 1)$ МДР-коды, каждый из которых получается выкалыванием проверок из $C_q(q-1, K_{O,l}, q - K_{O,l})$-кодов Рида — Соломона. Для декодирования внешних кодов необходимо использовать алгоритмы декодирования, исправляющие ошибки и стирания, подобные тем, которые описаны в [17, 18].

Для расчета параметров кодовой конструкции в настоящей работе предлагается использовать оптимизационные процедуры, подобные описанным в [16]. Отличие состоит в том, что в работе [16] рассматривался сценарий, при котором в качестве внутренних использовались сравнительно длинные двоичные коды (в частности, коды БЧХ) и традиционное алгебраическое декодирование с «жестким» входом, т. е. фактически модель двоичного симметричного канала, что давало возможность аналитически оценить вероятности ошибочного декодирования и отказа от декодирования. В нашем случае, так как для декодирования внутренних кодов используется алгоритм декодирования с «мягким» входом, такая аналитическая оценка не представляется возможной, поэтому оценки вероятностей отказа и ошибочного декодирования внутренних кодов ОКК (кодов-столбцов в частотной области) были получены с помощью имитационного моделирования. Следует также отметить, что выбор оптимальных значений параметра α в такой конструкции также сопряжен с минимизацией вероятностей ошибки/отказа (в частности, в рассматриваемом в данной работе случае минимизации вероятности отказа от декодирования внутреннего кода 0-го слоя), и получаемые в результате такой оптимизации значения α существенно больше, чем значения α, полученные в результате оптимизации в [11], так как там рассматривалась одночастотная система, в которой в каждом кадре передавался только один символ одного кодового слова, и задача состояла в том, чтобы выбрать параметр α таким образом, чтобы максимизировать пропускную способность. При этом, разумеется, необходимое условие состояло в гарантии того, что единственный ненулевой элемент, который соответствовал символу, переданному рассматриваемым пользователем, был с высокой вероятностью «покрыт» вектором веса α на выходе канала. В рассматриваемом случае в каждом кадре каждым пользователем передаются символы сразу нескольких кодовых слов внутреннего кода во временной области, таким образом, в каждом кадре передается несколько ненулевых элементов, и задача состоит в том, чтобы выбрать вес вектора на выходе канала α так, чтобы вектор веса α с высокой вероятностью покрывал все ненулевые элементы, соответствующие переданным рассматриваемым пользователем сигналам, чем и объясняется необходимость использовать существенно большие, чем в [11], значения α.

Расчет вероятностей для внешнего кода производился аналитически, при этом, как и в [16], предполагалось, что отказ от декодирования или ошибочное декодирование происходят, если число стираний и ошибок превышает гарантированную корректирующую способность кода. Так как отказы при декодировании внутренних кодов приводят к стираниям соответствующих символов, а ошибочное декодирование — к появлению ошибок, распределение ошибок, стираний и безошибочных символов в кодовом слове описывается полиномиальным (мультиномиальным) или, точнее, триномиальным распределением, и вероятность того, что при декодировании внешнего кода l-го слоя произойдет ошибка или отказ, определяется выражением

$$FER_l\left(N_O, d_{O,l}, p_{e,l}\left(n,N_r,U,q\right), p_{x,l}\left(n,N_r,U,q\right)\right) = 1 - \sum_{\varepsilon=0}^{\left\lfloor \frac{d_{O,l}-1}{2} \right\rfloor} \sum_{\xi=0}^{d_{O,l}-2\varepsilon-1} \times \left(\frac{N_O!}{\varepsilon!\xi!(N_O-\varepsilon-\xi)!} \cdot p_{e,l}^{\varepsilon}\left(n,N_r,U,q\right) \times p_{x,l}^{\xi}\left(n,N_r,U,q\right) \times \left(1 - p_{e,l}\left(n,N_r,U,q\right) - p_{x,l}\left(n,N_r,U,q\right)\right)^{N-\varepsilon-\xi} \right),$$

где $p_{e,l}(n, N_r, U, q)$ и $p_{x,l}(n, N_r, U, q)$ — вероятности ошибочного декодирования и отказа от декодирования l-го слоя обобщенного каскадного кода соответственно, зависящие от длины внутреннего кода n, длины кода в частотной области (высоты столбца матрицы обобщенного каскадного кода, длины внутреннего кода обобщенного каскадного кода) N_r, числа «мешающих» пользователей U и мощности алфавита q.

Кроме того, для каждой сигнально-кодовой конструкции определялась эффективная скорость и длина кодовой комбинации во временной области. Так как для передачи каждого столбца кодового слова ОКК требуется передать n символов, общая длина кодовой комбинации во временной области

$$N_T = n \cdot N_O\left(p_{e,0}\left(n,N_r,U,q\right), \ldots, p_{e,M-1}\left(n,N_r,U,q\right), p_{x,0}\left(n,N_r,U,q\right), \ldots, p_{x,M-1}\left(n,N_r,U,q\right), n\right).$$

Эта величина также определяет количество OFDM-символов, которые соответствуют слову ОКК. По аналогии с [11], скорость передачи определялась в битах на однократное использование канала (сокращенно б/ОИК), но в данном случае эта скорость определяется при условии, что вероятность ошибочного декодирования/отказа не превышает порогового значения. Эта скорость равна

$$R_e\left(p_{e,0}(n,N_r,U,q),\ldots,p_{e,M-1}(n,N_r,U,q),\right.$$
$$\left.p_{x,0}(n,N_r,U,q),\ldots,p_{x,M-1}(n,N_r,U,q),n\right)=$$
$$=\frac{\log_2 q\sum_{l=0}^{M-1}K_{O,l}\left(p_{e,l}(n,N_r,U,q),p_{x,l}(n,N_r,U,q),n\right)}{n\cdot N_O\begin{pmatrix}p_{e,0}(n,N_r,U,q),\ldots,p_{e,M-1}(n,N_r,U,q),\\p_{x,0}(n,N_r,U,q),\ldots,p_{x,M-1}(n,N_r,U,q),n\end{pmatrix}}. \quad (3)$$

Рассмотрим два вида оптимизационных процедур. Введем следующие обозначения: вектор на выходе декодера на l-м шаге декодирования, соответствующий вектору-столбцу слова ОКК $\tilde{\mathbf{d}}_l$, слово внутреннего кода l-го слоя, соответствующее тому же столбцу $\tilde{\mathbf{x}}_l$, $\chi_l = 1$ — отказ от декодирования на l-м шаге (а наличие решения, напротив, $\chi_l = 0$).

А. Минимизация общей длины кодовой комбинации при заданной эффективной скорости R_e^* и выполнении условия

$$FER_l\left(N_O,d_{O,l},p_{e,l},p_{x,l}\right)\le FER^*,$$
$$\forall l=0\ldots M-1. \quad (4)$$

А.1. Для каждого значения n и N_r при помощи имитационного моделирования определяются вероятности ошибки и отказа для каждого из N_r слоев ОКК для различных значений α:

$$\left(p\left(\tilde{\mathbf{d}}_l\ne\tilde{\mathbf{x}}_l\mid\alpha,\chi_l=0,\tilde{\mathbf{d}}_{l-1}=\tilde{\mathbf{x}}_{l-1},\ldots,\tilde{\mathbf{d}}_0=\tilde{\mathbf{x}}_0,\right.\right.$$
$$\left.\chi_{l-1}=0,\ldots,\chi_0=0\right),$$

$$\left.\left.p\left(\chi_l=1\middle|\alpha,\tilde{\mathbf{d}}_{l-1}=\tilde{\mathbf{x}}_{l-1},\ldots,\tilde{\mathbf{d}}_0=\tilde{\mathbf{x}}_0,\chi_{l-1}=0,\ldots,\chi_0=0\right)\right).$$

А.2. Выбирается оптимальное значение параметра α

$$\alpha^*=\arg\min_{\alpha}p\left(\tilde{\mathbf{d}}_0\ne\tilde{\mathbf{x}}_0\middle|\alpha,\chi_0=0,\tilde{\mathbf{d}}_{-1}=\mathbf{0}\right),$$

т. е. выбирается значение α, минимизирующее вероятность ошибочного декодирования для 0-го слоя и соответствующие ему значения вероятностей ошибки и отказа для всех слоев.

А.3. Перебираются значения N_O (начиная с $N_O = q-1$). Для каждого значения N_O и для каждого (пусть для определенности l-го) слоя перебираются значения $d_{O,l}$ (в диапазоне от трех до N_O), вычисляется вероятность ошибки/отказа внешнего кода, проверяется условие (4) и рассчитывается скорость передачи по формуле (3). В результате для каждого l выбирается минимальное значение $d_{O,l}$, при котором выполняется условие ($R_e \ge R_e^*$). Процедура повторяется до тех пор, пока не найдено минимальное значение N_O (при заданных n и N_R), для которого выполняется (4).

А.4. Из различных пар значений n и N_r выбирается та, которой соответствует минимальное значение общей длины кодовой комбинации N_T.

Б. Максимизация эффективной скорости R_e при выполнении условия (4).

Б.1. Для каждого значения n и N_r при помощи имитационного моделирования определяются вероятности ошибки и отказа для каждого из N_r слоев ОКК для различных значений α:

$$p\left(\tilde{\mathbf{d}}_l\ne\tilde{\mathbf{x}}_l\mid\alpha,\chi_l=0,\tilde{\mathbf{d}}_{l-1}=\tilde{\mathbf{x}}_{l-1},\ldots,\tilde{\mathbf{d}}_0=\tilde{\mathbf{x}}_0,\right.$$
$$\left.\chi_{l-1}=0,\ldots,\chi_0=0\right),$$

$$p\left(\chi_l=1\middle|\alpha,\tilde{\mathbf{d}}_{l-1}=\tilde{\mathbf{x}}_{l-1},\ldots,\tilde{\mathbf{d}}_0=\tilde{\mathbf{x}}_0,\chi_{l-1}=0,\ldots,\chi_0=0\right).$$

Б.2. Выбирается оптимальное значение параметра α

$$\alpha^*=\arg\min_{\alpha}p\left(\chi_0=0\middle|\alpha,\tilde{\mathbf{d}}_{-1}=0\right),$$

т. е. выбирается значение α, минимизирующее вероятность отказа от декодирования для 0-го слоя, и соответствующие ему значения вероятностей ошибки и отказа для всех слоев.

Б.3. Для всех возможных значений и каждого слоя вычисляются минимальные значения, для которых выполняется (4) и эффективная скорость передачи максимальна.

Б.4. Выбирается пара n и N_r, которой соответствует максимальная скорость R_e^*.

Исследуемый сценарий, результаты численных экспериментов и примеры оптимизированных конструкций

Рассмотрим следующий сценарий: будем полагать, что, помимо рассматриваемого пользователя, передачу по каналу вверх одновременно ведут $U = 500$ пользователей, причем все пользователи используют одну и ту же сигнальную конструкцию, подразумевающую передачу N_r кодовых слов внутреннего кода во временной области и, соответственно, N_r сигналов каждым из пользователей. Таким образом, в каждый момент времени, помимо N_r сигналов от рассматриваемого пользователя, по каналу передаются UN_r «мешающих» сигналов. При этом для передачи

каждого из этих сигналов пользователи могут воспользоваться одной из $\tilde{Q} = \lfloor 0{,}8 \cdot Q \rfloor$ поднесущих, так как поднесущие на краях полосы отводятся под защитные полосы (далее в этом разделе всегда будем полагать, что $Q = 4096$ и, следовательно, $\tilde{Q} = 3276$). Как и в [11], будем считать, что потери мощности сигналов, переданных каждым из пользователей, описываются лог-нормальной моделью [19], а соотношение мощностей — «пессимистической» моделью [11]. В рамках этой модели исследуется сценарий, в котором рассматриваемый пользователь находится на расстоянии d^* от приемника, а расстояние от каждого из остальных («мешающих») пользователей выбирается случайно и при этом выполняется $d^* \geq d_i \; \forall \; i \in [1 \ldots K]$, т. е. расстояние от передатчика каждого из «мешающих» пользователей до приемника не больше, чем расстояние от рассматриваемого пользователя до приемника. В этом случае отношение мощности $P_{t,i}$ сигнала, переданного i-м «мешающим» пользователем, и мощности $P_{r,i}$ этого же сигнала на приемном конце описывается выражением [19]

$$L(d_i) = 10\log\left(\frac{P_{t,i}}{P_{r,i}}\right) = \bar{L}(d_i) + X_\sigma = $$

$$= L_{fs}(d_0) + 10\gamma_0 \log\left(\frac{d_i}{d_0}\right) + X_\sigma,$$

где d_i — расстояние между передатчиком i-го «мешающего» пользователя и приемником; X_σ — гауссова случайная величина с математическим ожиданием 0 и среднеквадратическим отклонением σ; L_{fs} — показатель потерь при распространении в свободном пространстве, задаваемый формулой Фрииса; d_0 — эталонное расстояние; γ_0 — коэффициент ослабления при распространении. Отношение мощности P_t^* сигнала, переданного рассматриваемым пользователем, к мощности P_r^* этого сигнала на приемном конце:

$$L(d^*) = 10\log\left(\frac{P_t^*}{P_r^*}\right) = \bar{L}(d^*) + X_\sigma = $$

$$= L_{fs}(d_0) + 10\gamma_0 \log\left(\frac{d^*}{d_0}\right) + X_\sigma,$$

где d^* — расстояние между передатчиком рассматриваемого пользователя и приемником. Таким образом, логарифм отношения мощности сигнала от рассматриваемого пользователя на приемном конце к мощности сигнала от каждого из «мешающих» пользователей на приемном конце также является гауссовой случайной величиной, зависящей от случайной величины $\mu_i = \log_{10}\frac{d^*}{d_i}$.

Как и в [11], будем считать, что величины μ_i выбираются случайно и равновероятно из диапазона значений от 0 до A_m с шагом ΔA_m. Ниже во всех численных экспериментах использовались параметры $\gamma_0 = 3{,}5$, $\sigma = 8$ дБ, $A_m = 2$, $\Delta A_m = 0{,}01$. То, что $\mu_i \geq 0$, означает, в частности, что средняя мощность сигнала от рассматриваемого пользователя не больше, чем средняя мощность сигнала от любого из «мешающих» пользователей.

Выше сказано, что фазы всех сигналов моделировались как случайные величины, равномерно распределенные на окружности от 0 до 2π. Фоновый аддитивный шум моделировался как двумерный случайный гауссов процесс с нулевым математическим ожиданием, описываемый отношением сигнал/шум, которое задавалось выражением

$$SNR = 10 \cdot \log_{10}\left(\frac{E_s}{\log_2(q) \cdot N_r \cdot E_N}\right),$$

где E_s — энергия сигнала, переданного пользователем; E_N — энергия шума (во всей полосе, предоставленной активным пользователям). Во всех численных экспериментах, которые будут обсуждаться здесь, использовалось значение отношения сигнал/шум SNR = –20 дБ.

Ниже представлены результаты численных экспериментов для двух описанных оптимизационных процедур. Для удобства анализа и сравнения оптимальные значения для различных величин сведены в табл. 1 и 2, показаны только значения, близкие к выбранному в результате оптимальному набору параметров.

■ ***Таблица 1.*** Результаты оптимизации параметров кодовых конструкций для $FER^* = 10^{-4}$ и $R_e^* = 2$ б/ОИК (минимальная длина кодовой комбинации, $L_M = 64$)

■ ***Table 1.*** Optimized code construction parameters for $FER^* = 10^{-4}$ and $R_e^* = 2$ (minimum overall code length, $L_M = 64$)

N_r	n	α^*	Параметры внешних кодов	N_T	R_e, б/ОИК
3	6	117	(20,12)	120	2
			(20,18)		
3	7	120	(28,23)	196	2
			(28,26)		
4	8	`143	(13,4)	104	2
			(13,11)		
			(13,11)		
4	9	143	(14,8)	126	2,0317
			(14,12)		
			(14,12)		
4	10	142	(16,12)	160	2
			(16,14)		
			(16,14)		

■ ***Таблица 2.*** Результаты оптимизации параметров кодовых конструкций для $FER^* = 10^{-6}$ (максимальная эффективная скорость, $L_M = 64$)

■ ***Table 2.*** Optimized code construction parameters for $FER^* = 10^{-6}$ (maximum effective rate, $L_M = 64$)

N_r	n	α^*	Параметры внешних кодов	N_T	R_e, б/ОИК
4	9	143	(249,215)	2241	2,4775
			(249,239)		
			(249,240)		
4	8	141	(255,159)	2040	2,5412
			(255,244)		
			(255,245)		
5	12	163	(253,218)	3036	2,4242
			(253,234)		
			(253,234)		
			(253,234)		
5	11	165	(248,145)	2728	2,5044
			(248,233)		
			(248,238)		
			(248,238)		

Как видно из табл. 1, минимальное значение общей длины кодовой комбинации $N_T = 104$ достигается при $N_r = 4$, $n = 8$ и $N_O = 13$. Несмотря на то, что минимизация задержки не была заявлена в качестве результата процедуры, полученные параметры позволяют решить и эту задачу. Так как параметры кодов 1-го и 2-го уровня одинаковы, можно сократить число уровней до двух, использовав в качестве внешнего кода 1-го уровня два независимых кодовых слова одного и того же МДР-кода (13,11,3) над $GF(2^8)$, полученного выкалыванием проверок из (255,11,245)-кода Рида — Соломона (частичная матрица, соответствующая 1-му слою, в этом случае представляет собой слово кода-произведения). Такой выбор позволяет сократить число шагов декодирования и использовать только два декодера для внешних кодов вместо трех, причем кодовые слова, использованные в качестве внешних кодов 1-го слоя, могут декодироваться независимо и параллельно.

По приведенным в табл. 2 данным видно, что в рассматриваемом случае оптимальной оказалась конструкция с $N_r = 4$, $n = 8$ и оптимальным значением $\alpha^* = 141$, однако для максимизации скорости потребовалась существенно большая длина внешнего кода. В качестве внешних используются коды Рида — Соломона (255,159,97), (255,244,12) и (255,245,11) над $GF(2^8)$. Как и в прошлом примере, параметры кодов очень близки (что объясняется тем, что вероятности отказа для 1-го и 2-го слоев для этого случая отличаются незначительно, а вероятности ошибок очень малы), и можно уменьшить аппаратную сложность и задержку, сократив число слоев до двух и использовав два слова одного и того же кода в качестве внешнего кода для 1-го слоя, что несколько уменьшит эффективную скорость (до 2,5373 б/ОИК, т. е. примерно на 1,54 %). Отметим также, что в работе [10] использование внешнего кода с общей проверкой на четность в частотной области и декодирования с «мягким» входом позволяло получать вероятности отказа порядка 10^{-3} и более на блок при скорости 2 б/ОИК. Дополнив кодирование в частотной области внешним кодом и используя свойства обобщенных каскадных кодов и алгоритм декодирования обобщенных каскадных кодов с мягким декодированием, нам удалось построить примеры конструкций со скоростями до 2,5 б/ОИК и выше и вероятностями отказа/ошибки в диапазоне от 10^{-4} до 10^{-6} на блок. Разумеется, необходимо отметить, что этот результат достигнут ценой увеличения как аппаратной сложности (за счет использования дополнительных кодеров внутренних и внешних кодов и декодеров внешних кодов), так и задержки (в силу необходимости в последовательном декодировании слоев и дополнительных процедур декодирования внешних кодов и увеличения длины кодовой комбинации).

Заключение

В работе рассмотрена задача выбора внешнего кода в системе связи, использующей частотно-позиционное кодирование в динамически выделяемом частотном поддиапазоне, прием на основе порядковых статистик и подкоды МДР-кодов. Для решения поставленной задачи применялся аппарат ОКК. В рамках этого подхода предложена кодовая конструкция, в которой в качестве внутренних кодов были выбраны короткие МДР-коды в частотной области, а в качестве внешних — МДР-коды во временной области. Для этой конструкции разработана модификация алгоритма декодирования с «мягким» входом, адаптированная к оценкам достоверности, вычисляемым в приемнике описанного выше типа, и модифицированный алгоритм декодирования ОКК с таким декодированием внутренних кодов. Предложенный подход и полуаналитический метод расчета позволяют эффективно рассчитывать и сравнивать характеристики кодовых конструкций и выбирать оптимальные значения параметров с учетом конкретных сценариев и требований к проектируемым системам. Все построенные с помощью предлагаемого подхода конструкции допускают эффективную

аппаратную реализацию: кодеры и декодеры внутренних кодов имеют небольшую сложность и могут работать параллельно, внешние коды принадлежат к известному классу кодов, для которого существуют многочисленные примеры эффективной аппаратной реализации.

Финансовая поддержка

Работа выполнена в Институте проблем передачи информации им. А. А. Харкевича РАН в рамках работ по госзаданию № FFNU-2025-0028.

Литература

1. **Шашин А. Э., Белогаев А. А., Красилов А. Н., Хоров Е. М.** Алгоритм выбора параметров передачи для апериодического URLLC-трафика в восходящем канале. *Информационные процессы*, 2022, т. 22, № 2, с. 29–41. doi:10.53921/18195822_2022_22_2_29
2. **Бакулин М. Г., Бен Режеб Т. Б. К., Крейнделин В. Б., Панкратов Д. Ю., Смирнов А. Э.** Технология NOMA с кодовым разделением в 3GPP: 5G или 6G? *T-Comm: Телекоммуникации и транспорт*, 2022, т. 16, № 1, с. 4–14. doi:10.36724/2072-8735-2022-16-1-4-14, EDN: ZBEDVU
3. **Liu X., Wang X., Zhao X., Du F., Zhang Y., Geng S.** Coexistence of energy-minimizing URLLC and eMBB in power IoT via NOMA-based collaborative MEC heterogeneous network. *IEEE Transactions on Vehicular Technology*, 2024, vol. 73, no. 7, pp. 10316–10332. doi:10.1109/TVT.2024.3376524
4. **Gao P., Liu Z., Xiao P., Foh C. H., Zhang J.** Low-complexity channel estimation and multi-user detection for uplink grant-free NOMA systems. *IEEE Wireless Communications Letters*, 2022, vol. 11, no. 2, pp. 263–267. doi:10.1109/LWC.2021.3125453
5. **Zhang X., Zhou Z., Ding Z., Ma Z., Hao L.** Joint sparse channel estimation and multiuser detection using spike and slab prior-based gibbs sampling for uplink grant-free NOMA. *IEEE Transactions on Vehicular Technology*, Dec. 2024, vol. 73, no. 12, pp. 19338–19349. doi:10.1109/TVT.2024.3451450
6. **Крещук А. А., Потапов В. Г.** Некоторые статистические демодуляторы для частотно-позиционного кодирования с быстрой перестройкой частот. *Автоматика и телемеханика*, 2013, № 10, с. 109–118. EDN: PMZXXS
7. **Kreshchuk A., Potapov V.** Better goodness-of-fit statistics for coded FSK decoding. *Electronic Notes in Discrete Mathematics*, March 2017, vol. 57, pp. 139–145. https://doi.org/10.1016/j.endm.2017.02.024
8. **Kreshchuk A.** Soft-decision statistical decoder for coded DHA FH OFDMA. *2018 Engineering and Telecommunication (EnT-MIPT)*, Moscow, 15–16 Nov 2018, pp. 8–11. doi:10.1109/EnT-MIPT.2018.00009
9. **Osipov D.** A multi-code multi-tone DHA FH OFDMA system with nonparametric reception. *Moscow Workshop on Electronic and Networking Technologies (MWENT)*, Moscow, 11–13 March, 2020, pp. 1–5. doi:10.1109/MWENT47943.2020.9067339
10. **Osipov D.** Concatenated coset coding in a multi-tone DHA FH OFDMA system with order statistics-based reception. *IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, Purwokerto, Indonesia, 17–18 July 2021, pp. 123–127.
11. **Osipov D. S.** Signal detection amid noise using order statistics: detector sensitivity analysis and parameter choice. *Информационно-управляющие системы*, 2023, № 1, с. 61–70. doi:10.31799/1684-8853-2023-1-61-70, EDN: PCFDIS
12. **Molisch A. F.** *Wireless Communications: From Fundamentals to Beyond 5G*. (3 ed.) Chichester, John Wiley & Sons, Inc., 2022. 1008 p.
13. **Блох Э. Л., Зяблов В. В.** *Обобщенные каскадные коды*. М., Связь, 1976. 240 с.
14. **Spinner J., Freudenberger J.** Soft input decoding of generalized concatenated codes using a stack decoding algorithm. *2nd BW-CAR Symposium on Information and Communication Systems (SInCom)*, 2015, pp. 1–5. https://www.researchgate.net/publication/307964232_Soft_Input_Decoding_of_Generalized_Concatenated_Codes_Using_a_Stack_Decoding_Algorithm (дата обращения: 25.12.2025).
15. **Freudenberger J., Rajab M., Shavgulidze S.** A soft-input bit-flipping decoder for generalized concatenated codes. *2018 IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, 2018, pp. 1301–1305.
16. **Spinner J.** *Channel Coding for Flash Memories*. Doct. thes. Doct. of Eng. Sc., Konstanz, Univ. Konstanz, 2019. 124 p.
17. **Wu Y.** New scalable decoder architectures for Reed – Solomon codes. *IEEE Transactions on Communications*, 2015, vol. 63, no. 8, pp. 2741–2761. doi:10.1109/TCOMM.2015.2445759
18. **Tang N., Chen C., Han Y. S.** Fast error and erasure decoding algorithm for Reed – Solomon codes. *IEEE Communications Letters*, April 2024, vol. 28, no. 4, pp. 759–762. doi:10.1109/LCOMM.2024.3358055
19. **Pätzold M.** *Mobile Radio Channels*. 2nd ed. Chichester, John Wiley Sons, 2011. 583 p.

UDC 621.391
doi:10.31799/1684-8853-2026-3-49-62
EDN: ZJAEPG

Generalized concatenated code-based constructions for communication systems with order-statistics-based reception

D. S. Osipov[a,b], Dr. Sc., Tech., Senior Researcher, orcid.org/0000-0003-0400-7181, d_osipov@iitp.ru
[a]Kharkevich Institute for Information Transmission Problems, 19, bld. 1, Bolshoy Karetny Per., 127051, Moscow, Russian Federation
[b]HSE University, 34, Tallinskaya St., 123458, Moscow, Russian Federation

Introduction: In many communication systems under construction and those to be created power control and channel estimation techniques developed for the previous generation communication systems fail to provide desired precision. One way to solve this problem is to use order-statistics-based reception techniques that do not need channel estimation or power control. To ensure the desired probability characteristics coded modulation schemes using order-statistics-based reception techniques should be complemented with outer codes. **Purpose:** To propose a practical outer code providing both desired probability characteristics and a relatively high transmission rate for the communication system under consideration. **Results:** We develop a general approach to outer code design based on the theory of generalized concatenated codes. Within the scope of this approach codes in frequency domain treated as inner codes are complemented with outer codes in time domain that can be optimized according to the requirements corresponding to the specific requirements of the communication system to be designed. We devise soft-input decoding algorithms for inner code decoding and, thus, for the generalized concatenated codes decoding. We give examples of practical outer codes developed with the proposed approach combining simulation-based estimation of inner code error and decoding failure probabilities with analytical evaluation for the outer code choice. **Practical relevance:** Herein it is demonstrated that the proposed approach results in the elaboration of outer codes suitable to be optimized in several ways and to be applicable for hardware implementation in real-life applications.

Keywords — machine-to-machine communication, frequency hopping, an order-statistics-based reception method, generalized concatenated codes, inner codes, soft input decoding.

For citation: Osipov D. S. Generalized concatenated code-based constructions for communication systems with order-statistics-based reception. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 49–62 (In Russian). doi:10.31799/1684-8853-2026-3-49-62, EDN: ZJAEPG

Financial support
The research was carried out in IITP RAS (Kharkevich Institute) within the state assignment of Ministry of Science and Higher Education of the Russian Federation No. FFNU-2025-0028.

References

1. Shashin A. E., Belogaev A. A., Krasilov A. N., Khorov E. M. Algorithm for transmission parameters selection for sporadic URLLC traffic in uplink. *Information Processes*, 2022, vol. 22, no. 2, pp. 29–41 (In Russian). doi:10.53921/18195822_2022_22_2_29
2. Bakulin M. G., Ben Rejeb T. B. C., Kreyndelin V. B., Pankratov D. Y., Smirnov A. E. Code domain NOMA in 3GPP specifications: 5G or 6G? *T-Comm*, 2022, vol. 16, no. 1, pp. 4–14 (In Russian). doi:10.36724/2072-8735-2022-16-1-4-14, EDN: ZBEDVU
3. Liu X., Wang X., Zhao X., Du F., Zhang Y., Geng S. Coexistence of energy-minimizing URLLC and eMBB in power IoT via NOMA-based collaborative MEC heterogeneous network. *IEEE Transactions on Vehicular Technology*, 2024, vol. 73, no. 7, pp. 10316–10332. doi:10.1109/TVT.2024.3376524
4. Gao P., Liu Z., Xiao P., Foh C. H., Zhang J. Low-complexity channel estimation and multi-user detection for uplink grant-free NOMA systems. *IEEE Wireless Communications Letters*, 2022, vol. 11, no. 2, pp. 263–267. doi:10.1109/LWC.2021.3125453
5. Zhang X., Zhou Z., Ding Z., Ma Z., Hao L. Joint sparse channel estimation and multiuser detection using spike and slab prior-based gibbs sampling for uplink grant-free NOMA. *IEEE Transactions on Vehicular Technology*, Dec. 2024, vol. 73, no. 12, pp. 19338–19349. doi:10.1109/TVT.2024.345145
6. Kreshchuk A. A., Potapov V. G. Statistical demodulators for frequency shift keying with fast frequency hopping. *Automation and Remote Control*, 2013, vol. 74, no. 10, pp. 1688–1695. doi:10.1134/S0005117913100093, EDN: SKWQWV
7. Kreshchuk A., Potapov V. Better goodness-of-fit statistics for coded FSK decoding. *Electronic Notes in Discrete Mathematics*, March 2017, vol. 57, pp. 139–145. https://doi.org/10.1016/j.endm.2017.02.024
8. Kreshchuk A. Soft-decision statistical decoder for coded DHA FH OFDMA. *2018 Engineering and Telecommunication (EnT-MIPT)*, 2018, pp. 8–11. doi:10.1109/EnT-MIPT.2018.00009
9. Osipov D. A multi-code multi-tone DHA FH OFDMA system with nonparametric reception. *Moscow Workshop on Electronic and Networking Technologies* (*MWENT*), Moscow, pp. 1–5. doi:10.1109/MWENT47943.2020.9067339
10. Osipov D. Concatenated coset coding in a multi-tone DHA FH OFDMA system with order statistics-based reception. *IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, Purwokerto, 2021, pp. 123–127.
11. Osipov D. S. Signal detection amid noise using order statistics: detector sensitivity analysis and parameter choice. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2023, no. 1, pp. 61–70. doi:10.31799/1684-8853-2023-1-61-70, EDN: PCFDIS
12. Molisch A. F. *Wireless Communications: From Fundamentals to Beyond 5G*. (3 ed.) Chichester, John Wiley & Sons, Inc., 2022. 1008 p.
13. Blokh E. L., Zyablov V. V. *Obobschennye kaskadnye kody* [Generalized concatenated codes]. Moscow, Svyaz Publ., 1976, 240 p. (In Russian).
14. Spinner J., Freudenberger J. Soft input decoding of generalized concatenated codes using a stack decoding algorithm. *2nd BW-CAR Symposium on Information and Communication Systems (SInCom)*, 2015, pp. 1–5. Available at: https://www.researchgate.net/publication/ 307964232_Soft_Input_Decoding_of_Generalized_Concatenated_Codes_Using_a_Stack_Decoding_Algorithm (accessed 25 December 2025).
15. Freudenberger J., Rajab M., Shavgulidze S. A soft-input bit-flipping decoder for generalized concatenated codes. *2018 IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, 2018, pp. 1301–1305.
16. Spinner J. *Channel Coding for Flash Memories*. Doct. thes. Doct. of Eng. Sc., Konstanz, Univ. Konstanz, 2019. 124 p.
17. Wu Y. New scalable decoder architectures for Reed – Solomon codes. *IEEE Transactions on Communications*, 2015, vol. 63, no. 8, pp. 2741–2761. doi:10.1109/TCOMM.2015.2445759
18. Tang N., Chen C., Han Y. S. Fast error and erasure decoding algorithm for Reed – Solomon codes. *IEEE Communications Letters*, April 2024, vol. 28, no. 4, pp. 759–762. doi:10.1109/LCOMM.2024.3358055
19. Pätzold M. *Mobile Radio Channels*. 2nd ed. Chichester, John Wiley Sons, 2011. 583 p.

УДК 621.396
doi:10.31799/1684-8853-2026-3-63-74
EDN: WJHXOQ

Научные статьи
Articles

Оптимизация размещения беспилотных летательных аппаратов для покрытия территории в транкинговых системах связи

В. С. Иванов[a], *канд. техн. наук, доцент, orcid.org/0000-0001-9827-1690, ivanovmirea1@yandex.ru*
[a]*МИРЭА – Российский технологический университет, Вернадского пр., 78, Москва, 119454, РФ*

Введение: *значимость размещения БПЛА для беспроводного покрытия территории возрастает в условиях сложного рельефа, влияющего на распространение сигнала.* ***Цель:*** *разработать и провести сравнительный анализ методов минимизации числа БПЛА, гарантирующих покрытие дискретной области с учетом рельефа, на основе комбинирования жадных эвристик и стратегий декомпозиции пространства.* ***Результаты:*** *для каждой клетки сетки высот предварительно вычислен максимальный радиус действия БПЛА по модели Окумуры – Хата. Предложены три жадных алгоритма, отличиями которых являются критерии выбора позиции: максимизация числа покрытых клеток внутри области с минимизацией внешнего покрытия; приоритетная минимизация внешнего покрытия; максимизация радиуса без учета границ. На их основе реализованы три варианта декомпозиции рабочей зоны: вертикальное, горизонтальное и квадрантное разбиение с перекрытием, – при которых каждая подобласть обрабатывается независимо с последующим объединением решений. Эксперименты на случайных картах высот 20 × 20 показали, что базовый алгоритм с приоритетом покрытия внутренних клеток использует в среднем 42 БПЛА и создает 85,2 внешней точки. Алгоритм, минимизирующий внешнюю площадь, снижает ее до 70,5, но требует 56 БПЛА. Максимизация радиуса дает 42 БПЛА при наибольшем внешнем покрытии (94,4). Квадрантное разбиение показало 42 БПЛА и 82,7 внешней точки, что на 2,5 точки меньше базового при сохранении числа аппаратов. Вертикальное разбиение снизило число БПЛА до 41, но не улучшило внешнюю площадь, горизонтальное оказалось наименее эффективным.* ***Практическая значимость:*** *декомпозиция области с перекрытием и выбор оптимальной стратегии для каждой подобласти повышают качество размещения. Квадрантное разбиение демонстрирует лучший компромисс между числом БПЛА и нежелательным излучением, а также наименьший разброс показателей. Предложенные методы могут служить основой для формирования начальных популяций в генетических алгоритмах в целях поиска глобально оптимальных конфигураций сетей на базе БПЛА.*

Ключевые слова *– БПЛА, эвристические алгоритмы, беспроводная связь, транкинговые системы связи, оптимизация размещения, воздушная базовая станция, генетический алгоритм.*

Для цитирования: Иванов В. С. Оптимизация размещения беспилотных летательных аппаратов для покрытия территории в транкинговых системах связи. *Информационно-управляющие системы*, 2026, № 3, с. 63–74. doi:10.31799/1684-8853-2026-3-63-74, EDN: WJHXOQ
For citation: Ivanov V. S. Optimization of the placement of unmanned aerial vehicles to cover the territory in trunking communication systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 63–74 (In Russian). doi:10.31799/1684-8853-2026-3-63-74, EDN: WJHXOQ

Введение

Стремительное развитие беспилотных летательных аппаратов (БПЛА) открыло новые горизонты в области беспроводной связи. Благодаря своей мобильности, гибкости развертывания и возможности обеспечения каналов с высокой вероятностью прямой видимости БПЛА рассматриваются как перспективное решение для временного или постоянного усиления инфраструктуры наземных сетей [1]. Одной из центральных задач при проектировании сетей на базе БПЛА является определение минимального количества летательных аппаратов и их оптимального пространственного расположения, обеспечивающих гарантированное покрытие заданной территории с учетом рельефа местности и требований к качеству сигнала.

Проведено исследование зарубежных работ, посвященных применению БПЛА в качестве носителя приемопередающего оборудования, а также обзорных работ с анализом технологий беспроводной связи. Работы [2–4] посвящены всестороннему изучению использования БПЛА для создания и улучшения беспроводных сетей, исследуется оптимизация развертывания БПЛА с помощью генетического алгоритма. Подробно рассмотрены три ключевых аспекта: применение БПЛА, существующие проблемы и направления будущих исследований. Описаны основные сценарии применения БПЛА в сетях, такие как расширение покрытия сети в труднодоступных или временных местах; обеспечение связи в чрезвычайных ситуациях (после стихийных бедствий); сбор данных с устройств интернета вещей; улучшение производительности наземных сотовых сетей за счет оптимального размещения. Проанализированы технические трудности, включая моделирование и оптимизацию каналов связи с учетом трехмерной мобильности БПЛА; проблемы энергоэффективности и ограниченного времени полета; вопросы безопасности и конфиденциальности

при функционировании беспилотников; необходимость эффективно управлять помехами как от БПЛА, так и для них. Обозначены направления для будущих исследований: разработка алгоритмов для автономной и совместной работы группы БПЛА; интеграция БПЛА в сотовые сети 5G и далее; создание экономически эффективных и надежных архитектур сетей с участием БПЛА.

Работы [5, 6] представляют собой обзорный анализ технологий беспроводной связи, применяемых в сфере общественной безопасности. Основное внимание уделяется системам связи, которые используются экстренными службами (полицией, пожарными, скорой помощью) для координации действий и обмена информацией в кризисных ситуациях, включая стихийные бедствия и техногенные катастрофы. В рамках исследований авторы рассматривают эволюцию технологий — от традиционных профессиональных мобильных радиосистем, таких как TETRA и P25, до современных широкополосных решений на базе стандартов LTE и Wi-Fi. Подробно анализируются ключевые требования, предъявляемые к сетям общественной безопасности: надежность, отказоустойчивость, приоритезация трафика, защищенная передача данных и возможность взаимодействия между разными службами и даже странами. Особое место занимает анализ проблем. Среди них выделяются: фрагментация спектра и отсутствие единых глобальных стандартов; сложность обеспечения совместимости унаследованных систем с новыми технологиями; недостаточная пропускная способность сетей для передачи видео высокого качества с места событий.

Статьи [7–9] посвящены анализу возможностей и проблем, связанных с использованием БПЛА в системах беспроводной связи. Работы фокусируются на принципиально новой роли БПЛА не как пассивных объектов наблюдения, а как активных элементов сети — воздушных базовых станций или ретрансляторов. Авторы подробно рассматривают ключевые преимущества, которые открывает применение БПЛА. Главным из них является высокая вероятность наличия канала прямой видимости между БПЛА и наземными пользователями благодаря возможности оптимального трехмерного размещения. Это позволяет значительно улучшить качество связи и увеличить зону покрытия по сравнению с наземными вышками, особенно в сложных условиях городской застройки или труднодоступной местности. Кроме того, подчеркиваются мобильность и гибкость БПЛА, позволяющие развертывать сеть по требованию, например для временного усиления покрытия на массовых мероприятиях или для восстановления связи в зонах бедствий.

Работы [10, 11] посвящены решению задачи оптимального трехмерного размещения базовых станций на БПЛА в сотовых сетях. Основная цель исследований — разработка метода, позволяющего определить минимально необходимое количество и оптимальное положение БПЛА в пространстве для обслуживания всех пользователей в заданной области с учетом их неравномерного распределения. Авторы предлагают подход, использующий эвристический алгоритм (в частности, алгоритм роя частиц), который позволяет эффективно находить позиции БПЛА в трех измерениях. Ключевая задача заключается в том, чтобы, варьируя не только координаты в горизонтальной плоскости, но и высоту размещения БПЛА, обеспечить требуемое качество обслуживания для всех абонентов при минимальном количестве задействованных летательных аппаратов. Результаты моделирования показывают, что оптимальная высота размещения БПЛА напрямую зависит от плотности абонентов. В областях с низкой плотностью БПЛА могут работать на большей высоте, обеспечивая широкую зону покрытия и минимизируя взаимные помехи. В зонах с высокой концентрацией пользователей для обеспечения необходимой пропускной способности требуется снижение высоты, что позволяет эффективнее использовать частотный ресурс за счет пространственного разделения.

Настоящая работа посвящена разработке и сравнительному анализу жадных алгоритмов и методов декомпозиции для минимизации числа БПЛА, необходимых для полного покрытия рабочей области, представленной дискретной сеткой высот. В отличие от известных подходов, в данной статье зона действия каждого БПЛА не определяется фиксированным радиусом, а рассчитывается индивидуально для каждой потенциальной точки базирования на основе эмпирической модели Окумуры — Хата, адаптированной для учета разности высот и условий прямой видимости [12, 13]. Такой подход позволяет более реалистично моделировать распространение сигнала в условиях пересеченной местности за более короткий промежуток времени. Применение детерминированных моделей влечет за собой временные задержки [14], а статические методы не позволяют оценить рельеф местности [15].

Задача размещения БПЛА с ретрансляционным оборудованием является классической оптимизационной проблемой: требуется найти пространственную точку, максимизирующую покрытие абонентов при ограничениях на дальность прямой видимости, энергопотребление и рельеф местности. В условиях динамичной обстановки необходимо применять эффективные вычислительные методы. Среди известных подходов выделяется жадный алгоритм, последовательно принимающий локально оптимальные решения [16].

Более сложные метаэвристики вдохновлены природными процессами: генетический алгоритм имитирует естественный отбор, скрещивая и мутируя популяцию решений [17, 18]; пчелиный алгоритм моделирует поиск нектара разведчиками с последующей детальной разведкой перспективных зон [19, 20]; муравьиный алгоритм использует феромонную маркировку для постепенного усиления наилучших путей [21].

Основная цель настоящего исследования заключается в создании детерминированных алгоритмов размещения, которые гарантированно покрывают все клетки рабочей зоны при стремлении к минимизации двух ключевых показателей: общего количества задействованных БПЛА и площади нежелательного покрытия за границами области (в дальнейшем будет использоваться понятие «внешние точки»).

В рамках решения задачи минимизации количества БПЛА, необходимого для полного покрытия рабочей зоны, представленной сеткой $N \times N$, разработано три жадных алгоритма и три варианта декомпозиции. Полученные с их помощью решения будут использованы в качестве исходных популяций для генетического алгоритма, где посредством скрещивания планируется найти оптимальный вариант. Зона действия каждого БПЛА определяется эмпирической зависимостью от высоты точки базирования и окружающего рельефа (потенциальных позиций абонентов). На вход алгоритмов подаются матрица высот **H** (размером $N \times N$), рабочая частота f (в мегагерцах) и предельно допустимый уровень потерь сигнала $L_{\text{доп}}$ (в децибелах). Работа жадной эвристики строится итеративно: на каждом шаге выбирается наилучшая из еще не покрытых клеток для установки аппарата, после чего все клетки, вошедшие в зону его покрытия, исключаются из дальнейшего рассмотрения.

Предварительный расчет максимальных зон покрытия

На начальном этапе для каждой клетки (x, y) сетки размером $N \times N$ осуществляется предварительный расчет максимального радиуса действия $R_{\max}(x, y)$ БПЛА, помещенного в данную клетку. Значение $R_{\max}(x, y)$ определяется как наибольший радиус круга с центром в (x, y), внутри которого аппарат может поддерживать устойчивую связь со всеми абонентами. Процедура нахождения $R_{\max}(x, y)$ заключается в последовательном увеличении пробного целочисленного радиуса d от 0 до максимально возможного расстояния между двумя клетками сетки:

$$d_{\max} = \sqrt{(n-1)^2 + (n-1)^2}.$$

Для текущего d проверяются все клетки (i, j), для которых выполняется условие

$$d \geq \sqrt{(i-x)^2} + (j-y)^2.$$

Для каждой такой клетки с использованием эмпирической модели Окумуры — Хата, модифицированной для учета разности высот местности и условий прямой видимости, вычисляется требуемый радиус связи R_t:

$$L_U = 69{,}55 + 26{,}16 \times \log_{10}(f) + 13{,}82 \times \\ \times \log_{10}(h_{\text{б.с}} \pm \Delta h) + (0{,}8 + (1{,}1 \times \log_{10}(f) - 0{,}7) \times \\ \times (h_{\text{п.с}} \pm \Delta h)) + (44{,}9 - 6{,}55 \times \log_{10}(h_{\text{б.с}} \pm \Delta h)) \times \log_{10} R_t,$$

откуда

$$\log_{10} R_t = L_U - (69{,}55 + 26{,}16 \times \log_{10}(f) + 13{,}82 \times \\ \times \log_{10}(h_{\text{б.с}} \pm \Delta h) + (0{,}8 + (1{,}1 \times \log_{10}(f) - 0{,}7) \times \\ \times (h_{\text{п.с}} \pm \Delta h))) / (44{,}9 - 6{,}55 \times \log_{10}(h_{\text{б.с}} \pm \Delta h))),$$

где L_U — уровень потерь на трассе распространения радиосигнала, дБ; f — частота передачи сигнала, МГц; $h_{\text{б.с}}$ — высота подвеса антенны базовой станции, м; $h_{\text{п.с}}$ — высота подвеса антенны портативной станции, м; Δh — разница в высотах над уровнем моря в местах нахождения базовой и портативной станций, м; R_t — расстояние между объектами, км.

Если хотя бы для одной клетки окажется, что R_t меньше фактического расстояния до нее, то радиус d признается недостижимым, и максимальным принимается предыдущее значение $d - 1 : R_{\max}(x, y) = d - 1$.

В противном случае радиус увеличивается на единицу, и процесс повторяется до тех пор, пока не будет достигнут предельный радиус либо не возникнет первое нарушение условия. Если все проверки вплоть до максимального d завершаются успешно, то $R_{\max}(x, y)$ полагается равным этому максимальному значению: $R_{\max}(x, y) = d$.

После завершения описанных вычислений для всех клеток рабочей зоны формируется матрица радиусов radius_map, каждый элемент которой radius_map[x][y] содержит значение максимального радиуса, обеспечивающего устойчивую связь БПЛА, размещенного в клетке (x, y), со всеми абонентами внутри соответствующего круга.

Для дальнейшего размещения БПЛА подготавливаются следующие структуры данных. Вводится битовая матрица покрытия **C** размером $N \times N$, где **C**[i][j] = 0 означает, что клетка (i, j) еще не обслуживается ни одним БПЛА, а **C**[i][j] = 1 — что она уже покрыта. Изначально все элементы **C** устанавливаются в 0 (или False). Также создается пустой список **D**, в который последовательно добавляются записи о каждом размещенном БПЛА в формате $(x, y, R_{\max})$ — координаты центра и радиус покрытия. Для количественной

оценки нежелательного излучения за пределами рабочей зоны вводится переменная P_{sum}, аккумулирующая общее количество так называемых «внешних точек». Начальное значение этой переменной равно нулю.

Основной цикл размещения

Второй этап, специфика которого варьируется для каждого из рассматриваемых алгоритмов, заключается в итерационном процессе размещения БПЛА. Основной цикл повторяется до тех пор, пока не будет достигнуто полное покрытие всех клеток рабочей зоны:

$$\exists(i;j):C[i][j]==0.$$

На каждой итерации выполняется поиск наилучшего кандидата среди еще не покрытых клеток; клетки, уже отмеченные как покрытые, исключаются из рассмотрения.

Для каждого кандидата (x, y) с радиусом R = radius_map[x][y] вычисляются две величины:

K — количество непокрытых клеток, попадающих внутрь круга радиуса R с центром (x, y);

P — количество целочисленных точек вне сетки (координаты которых меньше 0 или больше $N - 1$), также принадлежащих этому кругу.

Перед началом перебора переменные инициализируются следующим образом: лучшее найденное значение K_{best}, соответствующее минимальное значение P_{best}, а координаты лучшего кандидата (x_{best}, y_{best}). Кандидат сравнивается с текущим лучшим по правилу: если $K > K_{best}$ (или $K = K_{best}$ и $P < P_{best}$), то данный кандидат признается новым лучшим:

$$K[x][y]=K_{\max}, P[x][y]=P_{\min}, [x][y]=best.$$

Таким образом, приоритет отдается максимизации числа вновь покрытых клеток внутри зоны, а при равной эффективности — минимизации нежелательного покрытия за ее пределами:

$$\forall(i;j):C[i][j]==1.$$

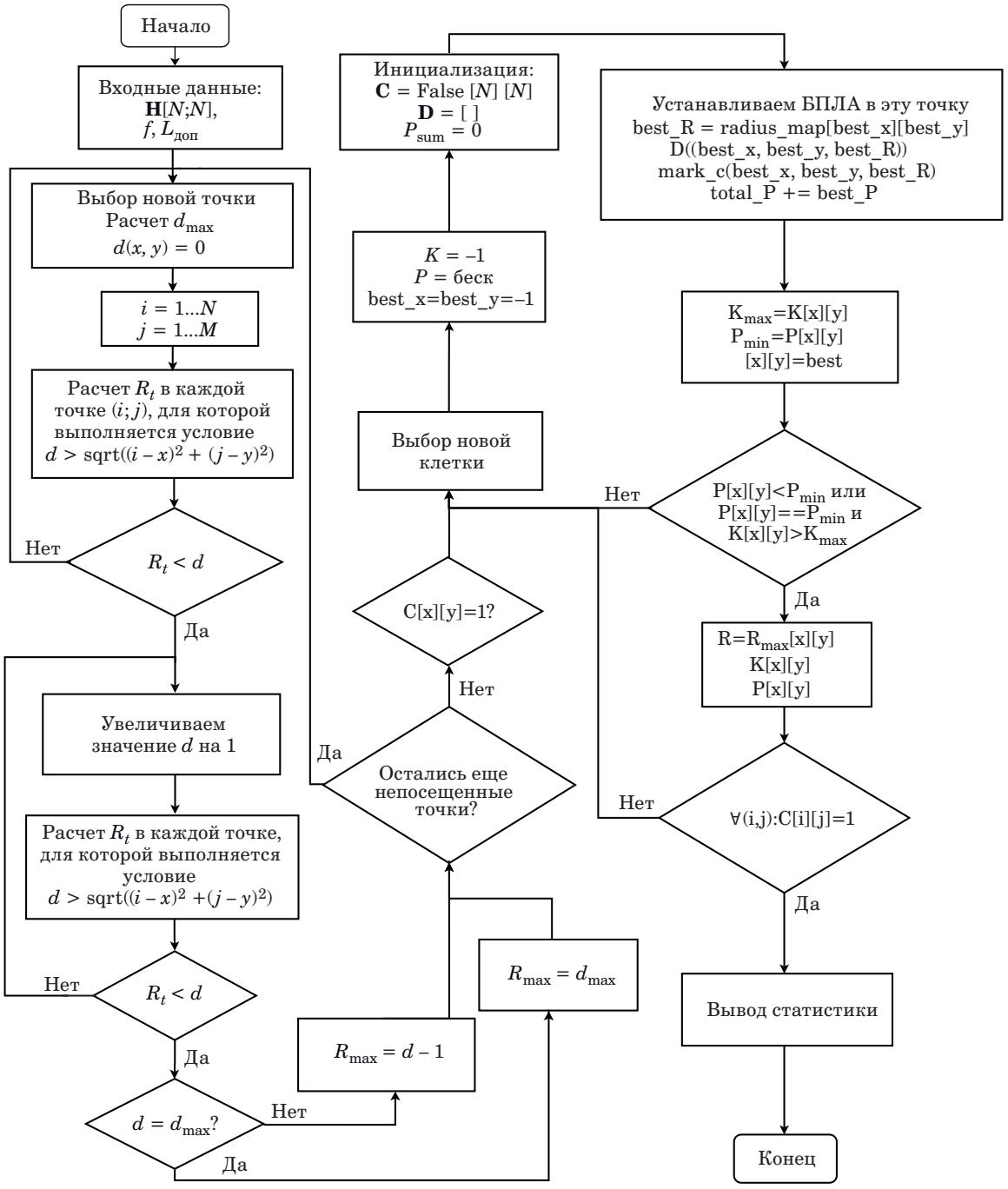
После завершения перебора выбранная позиция (x_{best}, y_{best}) с радиусом R_{best} добавляется в список D. С помощью функции mark_covered все клетки, попавшие в круг данного радиуса, помечаются как покрытые, а значение P_{best} прибавляется к суммарному счетчику внешних точек P_{sum}. Процесс завершается, когда все клетки становятся покрытыми. Если на каком-либо шаге не найдено ни одного кандидата с положительным значением K (что может свидетельствовать о недостижимости оставшихся клеток из-за особенностей рельефа), алгоритм также прекращает работу, однако в рамках постановки задачи предполагается, что полное покрытие достижимо.

Во втором варианте алгоритма (рис. 1) при сохранении вычисления тех же двух показателей, K и P, критерий выбора лучшего кандидата изменяется: предпочтение отдается клетке с наименьшим значением P, а при равенстве этого показателя — с наибольшим K. Формально это выражается условием: если $P < P_{best}$ (или $P = P_{best}$ и $K > K_{best}$), то данный кандидат становится новым лучшим. Таким образом, приоритет смещается в сторону минимизации нежелательного покрытия вне рабочей зоны, тогда как максимизация числа покрываемых клеток внутри зоны выступает в качестве вторичного критерия.

Третий вариант алгоритма полностью исключает из рассмотрения внешние точки, не учитывая показатель P при выборе позиции. Основным критерием здесь выступает максимально возможный радиус R БПЛА для данной клетки. Среди непокрытых клеток выбирается та, у которой значение R наибольшее; в случае равенства радиусов предпочтение отдается кандидату с бóльшим K. Поскольку величина P не оказывает влияния на решение, итоговое размещение может характеризоваться значительной внешней площадью покрытия.

Также разработано три варианта декомпозиции, построенных на трех вариантах жадного алгоритма. Первый вариант декомпозии (Split V) основан на разделении рабочей зоны на левую и правую подобласти с перекрытием. Сначала для левой части независимо применяются все три базовых жадных алгоритма и из полученных решений выбирается наилучшее по критерию минимизации числа БПЛА и внешней площади. Затем, с учетом покрытия созданного БПЛА из левой части, аналогичная процедура выполняется для правой подобласти. Итоговое решение формируется объединением БПЛА из обеих частей, при этом выбор комбинации алгоритмов для правой части также подчиняется указанному критерию. Такой подход позволяет локально оптимизировать размещение в каждой половине, сохраняя глобальную согласованность покрытия.

Во втором варианте (Split H) область делится на верхнюю и нижнюю части с перекрытием. Процедура аналогична Split V: сначала решается верхняя подобласть с перебором трех алгоритмов и выбором лучшего решения, затем с учетом его покрытия оптимизируется нижняя часть. Комбинации решений оцениваются по тому же правилу: в первую очередь минимизируется число БПЛА, затем внешняя площадь. Горизонтальное разбиение ориентировано на учет особенностей рельефа, изменяющихся преимущественно по вертикали.



■ ***Рис. 1.*** Блок-схема жадного алгоритма 2
■ ***Fig. 1.*** Block diagram of greedy algorithm 2

Третий вариант (Split Q) — наиболее детализированный метод декомпозиции, при котором рабочая зона разбивается на четыре квадранта с перекрытием. Решение строится последовательно: для верхнего левого квадранта выбирается лучший алгоритм (по критерию минимизации числа БПЛА и внешней площади), затем с учетом полученного покрытия аналогично обрабатывается верхний правый, нижний левый и, наконец, нижний правый квадранты. На каждом этапе выбор алгоритма для текущего квадранта производится так, чтобы итоговое объединенное решение было оптимальным по указанным метрикам. Такой подход позволяет более гибко адаптироваться к локальным изменениям рельефа, что в ряде случаев дает лучшее сочетание числа БПЛА и внешней площади по сравнению с двухчастным разбиением.

Визуализация расчетов

Для целей вычислительного моделирования и визуализации создано специализированное программное обеспечение на языке Python. Расчет радиусов покрытия, реализация трех жадных эвристик, а также построение карт высот и зон действия БПЛА выполнены с использованием библиотек NumPy и Matplotlib. Функционал программного обеспечения способствует последовательному запуску всех трех алгоритмов на идентичных наборах данных для корректного сравнения их эффективности. В качестве тестового полигона использовалась матрица высот размерностью 20×20, сгенерированная случайным образом с фиксированным начальным значением генератора псевдослучайных чисел для обеспечения воспроизводимости. Предельно допустимый уровень потерь сигнала был установлен равным 120 дБ.

В процессе выполнения алгоритмов в консоль выводилась детальная информация о каждом шаге: координаты выбранной позиции, рассчитанный радиус, количество вновь покрытых клеток внутри зоны, а также число точек, покрытых за ее пределами (рис. 2).

По завершении вычислений фиксировались следующие итоговые показатели:

— суммарное количество задействованных БПЛА;

— общее количество покрытий клеток вне рабочей зоны (с учетом перекрытий зон действия);

— количество уникальных внешних клеток, попавших под покрытие;

— доля уникального внешнего покрытия по отношению к общей покрытой площади.

Для обеспечения наглядности при сравнительном анализе результатов каждым из алгоритмов выполнялась визуализация карты высот с наложенными зонами покрытия БПЛА (рис. 3). Это позволило провести качественную оценку характера распределения устройств. Детерминированность реализованных алгоритмов гарантирует полную идентичность результатов при многократных запусках на одних и тех же исходных данных, что является необходимым условием корректности сравнения.

Проведенное сравнение позволило выявить особенности каждого подхода:

— алгоритм, максимизирующий покрытие внутри зоны, показал 40 БПЛА;

— алгоритм с приоритетом минимизации выхода за границы привел к увеличению количества БПЛА, но не всегда к уменьшению уникальной внешней площади;

— алгоритм, ориентированный на максимальный радиус, показал 37 БПЛА;

— вариант декомпозиции с вертикальным разделением рабочей зоны показал 38 БПЛА;

Drone Placement Report

Precomputing radii...

Radii computed.

Step 1: drone at (3, 6) with radius 3, covered 29 new cells, outside points 0
Step 2: drone at (6, 4) with radius 3, covered 21 new cells, outside points 0
Step 3: drone at (2, 11) with radius 2, covered 13 new cells, outside points 0
Step 4: drone at (2, 17) with radius 2, covered 13 new cells, outside points 0
Step 5: drone at (4, 14) with radius 2, covered 13 new cells, outside points 0
Step 6: drone at (6, 10) with radius 2, covered 13 new cells, outside points 0
Step 7: drone at (7, 16) with radius 2, covered 13 new cells, outside points 0
Step 8: drone at (9, 8) with radius 2, covered 13 new cells, outside points 0
Step 9: drone at (9, 13) with radius 2, covered 13 new cells, outside points 0
Step 10: drone at (11, 2) with radius 2, covered 13 new cells, outside points 0
Step 11: drone at (11, 17) with radius 2, covered 13 new cells, outside points 0
Step 12: drone at (12, 6) with radius 2, covered 13 new cells, outside points 0
Step 13: drone at (12, 11) with radius 2, covered 13 new cells, outside points 0
Step 14: drone at (14, 14) with radius 2, covered 13 new cells, outside points 0
Step 15: drone at (15, 3) with radius 2, covered 13 new cells, outside points 0
Step 16: drone at (15, 8) with radius 2, covered 13 new cells, outside points 0
Step 17: drone at (16, 17) with radius 2, covered 13 new cells, outside points 0
Step 18: drone at (17, 11) with radius 2, covered 13 new cells, outside points 0
Step 19: drone at (2, 2) with radius 2, covered 10 new cells, outside points 0
Step 20: drone at (17, 5) with radius 2, covered 10 new cells, outside points 0
Step 21: drone at (17, 2) with radius 2, covered 8 new cells, outside points 0
Step 22: drone at (17, 14) with radius 2, covered 7 new cells, outside points 0
Step 23: drone at (5, 17) with radius 2, covered 5 new cells, outside points 0
Step 24: drone at (7, 12) with radius 2, covered 4 new cells, outside points 0
Step 25: drone at (9, 3) with radius 2, covered 4 new cells, outside points 0
Step 26: drone at (12, 15) with radius 2, covered 4 new cells, outside points 0
Step 27: drone at (11, 9) with radius 2, covered 3 new cells, outside points 0
Step 28: drone at (13, 4) with radius 2, covered 2 new cells, outside points 0

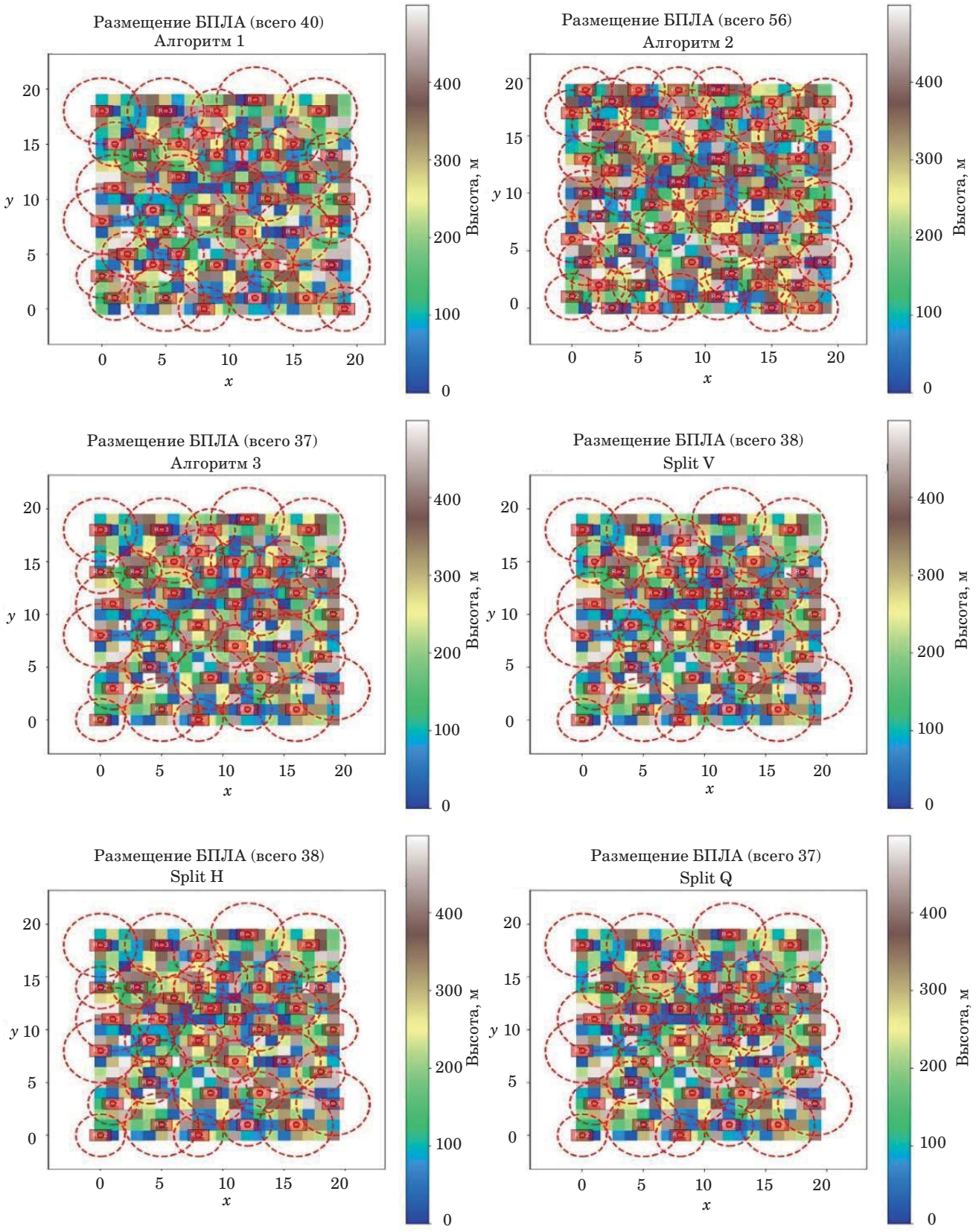
■ ***Рис. 2.*** Результат размещения БПЛА по жадному алгоритму 2

■ ***Fig. 2.*** Result of UAV placement using greedy algorithm 2

— варианты декомпозиции с горизонтальным и квадрантным разделением рабочих зон показали 37 БПЛА с размещением в разных точках.

Сравнительный анализ работы алгоритмов на различных картах

Для получения количественных оценок эффективности жадных алгоритмов была проведена серия из 10 экспериментов на различных случайно сгенерированных картах высот размером 20×20 клеток. В каждом испытании фиксировались два ключевых показателя: общее количество размещенных БПЛА и площадь уникального покрытия за пределами рабочей зоны (рис. 4). Анализ полученных данных, представленный на рис. 5, позволил выявить устойчивые закономерности в работе алгоритмов и сформулировать обоснованные выводы относительно их сравнительной эффективности.



■ ***Рис. 3.*** Визуализация результатов размещения БПЛА шестью алгоритмами
■ ***Fig. 3.*** Visualization of UAV deployment results using six algorithms

```
--- Test 1 (seed=0) ---
Alg1: drones=43, outside=88, time=981.02 ms
Alg2: drones=56, outside=75, time=1018.86 ms
Alg3: drones=44, outside=100, time=938.00 ms
SplitV: drones=44, outside=88, time=9017.87 ms
SplitH: drones=46, outside=91, time=9022.86 ms
SplitQ: drones=45, outside=84, time=9033.47 ms

--- Test 2 (seed=1) ---
Alg1: drones=39, outside=68, time=965.72 ms
Alg2: drones=55, outside=76, time=1032.48 ms
Alg3: drones=37, outside=86, time=1039.72 ms
SplitV: drones=39, outside=81, time=9239.00 ms
SplitH: drones=44, outside=84, time=9271.83 ms
SplitQ: drones=42, outside=94, time=9219.17 ms

--- Test 3 (seed=2) ---
Alg1: drones=40, outside=96, time=971.99 ms
Alg2: drones=56, outside=72, time=1084.93 ms
Alg3: drones=37, outside=99, time=1031.50 ms
SplitV: drones=38, outside=96, time=9408.06 ms
SplitH: drones=37, outside=98, time=9313.48 ms
SplitQ: drones=37, outside=95, time=9204.09 ms

--- Test 4 (seed=3) ---
Alg1: drones=43, outside=87, time=969.03 ms
Alg2: drones=54, outside=70, time=1021.90 ms
Alg3: drones=42, outside=93, time=937.07 ms
SplitV: drones=43, outside=93, time=9208.02 ms
SplitH: drones=45, outside=91, time=9189.17 ms
SplitQ: drones=44, outside=91, time=8795.56 ms

--- Test 5 (seed=4) ---
Alg1: drones=43, outside=74, time=951.61 ms
Alg2: drones=59, outside=64, time=1060.56 ms
Alg3: drones=43, outside=87, time=915.52 ms
SplitV: drones=44, outside=83, time=9188.55 ms
SplitH: drones=46, outside=69, time=9084.27 ms
SplitQ: drones=43, outside=74, time=8675.27 ms

--- Test 6 (seed=5) ---
Alg1: drones=40, outside=100, time=926.11 ms
Alg2: drones=57, outside=83, time=1007.62 ms
Alg3: drones=40, outside=112, time=899.76 ms
SplitV: drones=39, outside=97, time=9019.92 ms
SplitH: drones=40, outside=108, time=9078.94 ms
SplitQ: drones=40, outside=93, time=8883.90 ms

--- Test 7 (seed=6) ---
Alg1: drones=44, outside=81, time=936.36 ms
Alg2: drones=55, outside=67, time=1022.22 ms
Alg3: drones=45, outside=89, time=925.99 ms
SplitV: drones=43, outside=78, time=9107.69 ms
SplitH: drones=41, outside=75, time=9146.60 ms
SplitQ: drones=41, outside=72, time=8844.91 ms

--- Test 8 (seed=7) ---
Alg1: drones=41, outside=114, time=937.52 ms
Alg2: drones=56, outside=75, time=1019.19 ms
Alg3: drones=40, outside=108, time=920.86 ms
SplitV: drones=38, outside=96, time=9136.25 ms
SplitH: drones=42, outside=100, time=9142.26 ms
SplitQ: drones=40, outside=86, time=8777.77 ms

--- Test 9 (seed=8) ---
Alg1: drones=43, outside=72, time=924.07 ms
Alg2: drones=57, outside=62, time=1013.97 ms
Alg3: drones=44, outside=71, time=899.90 ms
SplitV: drones=40, outside=64, time=8859.75 ms
SplitH: drones=40, outside=66, time=9053.68 ms
SplitQ: drones=42, outside=69, time=8589.84 ms

--- Test 10 (seed=9) ---
Alg1: drones=39, outside=72, time=904.49 ms
Alg2: drones=54, outside=61, time=1004.88 ms
Alg3: drones=40, outside=99, time=896.28 ms
SplitV: drones=39, outside=72, time=8978.67 ms
SplitH: drones=39, outside=71, time=9092.77 ms
SplitQ: drones=39, outside=69, time=8727.40 ms

============================================================
FINAL STATISTICS (10 tests, N=20, Lu=120)
============================================================
Algorithm 1:
```

■ ***Рис. 4.*** Результаты расчетов алгоритмов на основе 10 тестов

■ ***Fig. 4.*** Results of algorithm calculations based on 10 tests

Алгоритм 1 (максимизация K и минимизация P) демонстрирует сбалансированные показатели: в среднем 42 БПЛА и внешняя площадь 85,2. Он служит точкой отсчета для сравнения.

Алгоритм 2 (минимизация P и максимизация K) обеспечивает наименьшую внешнюю площадь (70,5), но ценой значительного увеличения числа БПЛА — в среднем 56, что на 34 % больше, чем у алгоритма 1. Это делает его применимым только при жестких ограничениях на внешнее покрытие.

Алгоритм 3 (максимизация радиуса) дает примерно такое же число БПЛА (42), как и алгоритм 1, но при этом имеет наибольшую внешнюю площадь (94,4), что подтверждает нецелесообразность игнорирования выхода за границы.

Вертикальное разделение позволило снизить среднее число БПЛА до 41 (минимальное среди всех) при внешней площади 84,8, что немного лучше алгоритма 1. Он показывает стабильность и выигрыш в числе БПЛА.

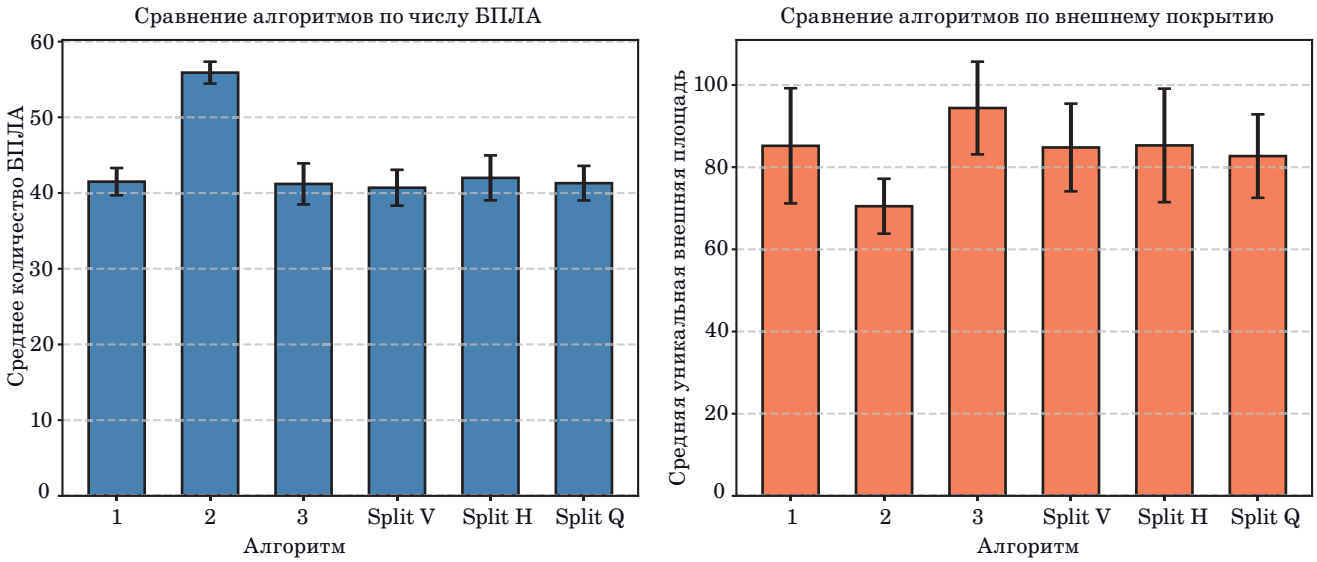
Горизонтальное разделение оказалось наименее эффективным среди методов декомпозиции: число БПЛА 42 и внешняя площадь 85,3.

Квадрантное разделение продемонстрировало среднее число БПЛА — 42 (близко к алгоритмам 1 и 3) и минимальную среди Split-методов внешнюю площадь — 82,7.

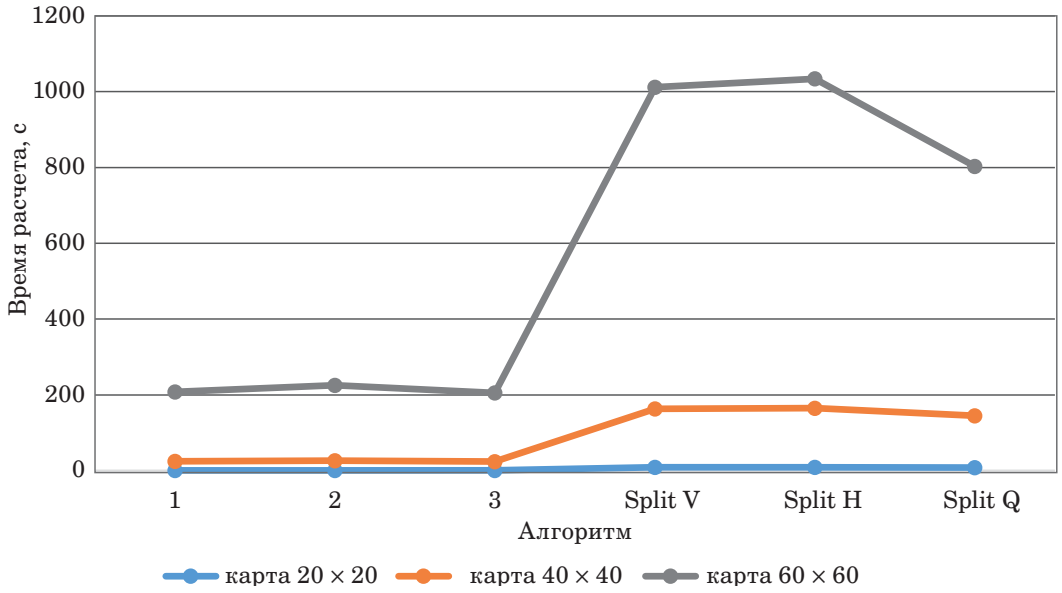
Вычисление проводилось на компьютере HP ProDesk с процессором Intel Core i5-10500T и 16 ГБ ОЗУ. Вычисление каждого алгоритма на карте 20×20 распределилось следующим образом: алгоритм 1 — 0,928 с; алгоритм 2 — 1,017 с; алгоритм 3 — 0,906 с; Split V — 8,870 с; Split H — 8,870 с; Split Q — 8,557 с.

Для оценки масштабируемости предложенных алгоритмов были проведены дополнительные эксперименты на картах размером 40×40 и 60×60 клеток. Измерялось среднее время выполнения каждого алгоритма на основе 10 тестов. График зависимости времени расчета от размера карт представлен на рис. 6.

Все алгоритмы демонстрируют резкий рост времени выполнения при увеличении размера сетки. Переход от 20×20 к 40×40 (увеличение числа клеток в 4 раза) приводит к росту времени в 20–27 раз для базовых жадных алгоритмов (алгоритмы 1–3) и в 16–18 раз для методов декомпозиции (Split V, H, Q). При дальнейшем увеличении до 60×60 (в 9 раз больше клеток, чем 20×20) время возрастает уже в 224–228 раз для базовых алгоритмов и в 114–117 раз для Split. Для ускорения работы алгоритмов на больших картах в дальнейшем предполагается распараллелить предварительный расчет радиусов и реализовать кеширование результатов расчета радиусов.



■ ***Рис. 5.*** Сравнительный анализ результатов расчетов алгоритмов
■ ***Fig. 5.*** Comparative analysis of algorithm calculation results



■ ***Рис. 6.*** Сравнительный анализ результатов расчетов на картах различных размеров
■ ***Fig. 6.*** Comparative analysis of calculation results on maps of different sizes

Заключение

В ходе исследования разработаны и реализованы три жадных алгоритма размещения БПЛА, а также три варианта декомпозиции рабочей области, направленные на минимизацию числа аппаратов при обеспечении полного покрытия с учетом рельефа местности и эмпирической модели потерь сигнала. Проведенные вычислительные эксперименты на случайных картах высот размером 20 × 20 клеток позволили выполнить сравнительный анализ предложенных методов по двум ключевым критериям: общему количеству задействованных БПЛА и площади уникального покрытия вне рабочей зоны.

Установлено, что базовый жадный алгоритм, максимизирующий число вновь покрытых клеток с последующей минимизацией внешних точек, демонстрирует сбалансированные показатели и может служить опорным решением. Алгоритм, ориентированный исключительно на минимизацию внешней площади, приводит

к существенному росту числа БПЛА, что ограничивает его применение лишь задачами с жесткими требованиями к излучению за пределами зоны. Игнорирование внешних точек в пользу максимального радиуса ухудшает результат по площади нежелательного покрытия без выигрыша в количестве БПЛА.

Применение декомпозиции с последовательным решением подобластей и комбинированием различных жадных стратегий позволило улучшить качество размещения. Среди методов декомпозиции вертикальное разбиение позволяет минимизировать количество БПЛА (41), а квадрантное — минимизировать внешнюю площадь (82,7). Выбор конкретного метода зависит от приоритетов проектирования: если задача заключается в сокращении числа БПЛА, то предпочтительно вертикальное разбиение; если же важнее ограничить нежелательное излучение за пределами зоны, то квадрантное. При этом квадрантное разбиение также характеризуется наименьшим разбросом показателей, что может быть полезно для обеспечения предсказуемости решения.

Измерение вычислительных затрат на картах 20×20, 40×40 и 60×60 показало, что базовые жадные алгоритмы работают за 3–4 минуты на карте 60×60, что приемлемо для задач однократного планирования. Методы декомпозиции, особенно квадрантный, обеспечивают лучшее качество покрытия (меньшее число БПЛА и внешних точек), но требуют значительно большего времени. Таким образом, выбор алгоритма зависит от приоритетов: при жестких ограничениях по времени следует использовать базовые алгоритмы, при требовании минимального числа БПЛА или внешнего покрытия — методы декомпозиции.

Полученные решения могут быть использованы в качестве начальных популяций для генетического алгоритма, что открывает перспективу дальнейшей оптимизации путем скрещивания и поиска глобально лучших конфигураций. Предложенные методы обеспечивают эффективное автоматическое размещение БПЛА с учетом рельефа и могут быть адаптированы для различных практических сценариев планирования беспроводных сетей.

Литература

1. **Чан Т. З., Кучерявый А. Е.** Оптимизация использования ресурсов воздушных базовых станций на основе методов искусственного интеллекта. *Труды учебных заведений связи*, 2025, т. 11, № 1, с. 62–68. doi:10.31854/1813-324X-2025-11-1-62-68, EDN: RVENVC
2. **Fujio S., Ozaki K.** Radio propagation environment prediction method for UAV-assisted cellular networks with millimeter-wave beamforming. *2024 International Conference on Computing, Networking and Communications (ICNC)*, 2024, pp. 957–961. doi:10.1109/ICNC59896.2024.10556093
3. **Mozaffari M., Saad W., Bennis M., Debbah M.** A tutorial on UAVs for wireless networks: Applications, challenges, and open problems. *IEEE Communications Surveys & Tutorials*, 2019, vol. 21, no. 3, pp. 2334–2360. doi:10.1109/COMST.2019.2902862
4. **Pandey A., Gunjan V.** Genetic algorithm-based optimization of UAV placement for fixed and dynamic target scenarios. *2025 International Conference on Electronics, AI and Computing (EAIC)*, India, 2025, pp. 1–8. doi:10.1109/EAIC66483.2025.11101363
5. **Florin R., Grozea C., Enache M., Nelega R., Kovacs G., Puschita R.** Mission-critical services in 4G/5G and beyond: Standardization, key challenges, and future perspectives. *Sensors*, 2025, vol. 25, no. 16, pp. 1–48. doi:10.3390/s25165156
6. **Kumbhar A., Koohifar F., Güvenç İ., Mueller B.** A survey on legacy and emerging technologies for public safety communications. *IEEE Communications Surveys & Tutorials*, 2017, vol. 19, no. 1, pp. 97–124. doi:10.1109/COMST.2016.2612223
7. **Zeng Y., Zhang R., Lim T. J.** Wireless communications with unmanned aerial vehicles: Opportunities and challenges. *IEEE Communications Magazine*, 2016, vol. 54, no. 5, pp. 36–42. doi:10.1109/MCOM.2016.7470933
8. **Khuwaja A. A., Chen Y., Zhao N., Alouini M. S., Dobbins P.** A survey of channel modeling for UAV communications. *IEEE Communications Surveys & Tutorials*, 2018, vol. 20, no. 4, pp. 2804–2821. doi:10.1109/comst.2018.2856587
9. **Mahbub M., Saym M. M., Jahan S., Paul A. K., Niyato D.** UAV-assisted wireless communications in the 6G-and-beyond era: An extensive survey on characteristics, standardization and regulations, enabling technologies, challenges, and future directions. *Vehicular Communications*, 2025, vol. 15, no. 3, pp. 45–67. doi:10.1016/j.vehcom.2025.100977
10. **Kalantari E., Yanikomeroglu H., Yongacoglu A.** On the number and 3D placement of drone base stations in wireless cellular networks. *Proc. of IEEE Vehicular Technology Conference (VTC2016-Fall)*, Montreal, QC, Canada, 2016, pp. 1–6. doi:10.1109/VTCFall.2016.7881122
11. **Savkin A. V., Huang H.** Proactive deployment of aerial drones for coverage over very uneven terrains: A version of the 3D art gallery problem. *Sensors*, 2019, vol. 19, no. 6, Article 1438. doi:10.3390/s19061438
12. **Ruz-Nieto A., Egea-López E., Molina-García-Pardo J.-M., Santa J.** A 3D simulation framework with ray-tracing propagation for LoRaWAN communication. *Internet of Things*, 2023, vol. 24, Article 100964. doi:10.1016/j.iot.2023.100964
13. **Иванов В. С., Увайсов С. У., Иванов И. А.** Алгоритм автоматического размещения базовых станций

транкинговых систем связи. *Труды учебных заведений связи*, 2023, т. 9, № 5, с. 25–34. doi:10.31854/1813-324X-2023-9-5-25-34, EDN: JMSIAV

14. **Ali N., Al-Behadili H. A. H.** A comprehensive review of radio signal propagation prediction for terrestrial wireless communication systems. *Misan Journal of Engineering Sciences*, 2024, vol. 3, no. 1, pp. 100–120. doi:10.61263/mjes.v3i1.80
15. **Мамченко М. В., Зорин В. А., Романова М. А.** Эмпирическая модель расчета затухания сигнала с учетом коэффициента застройки местности для беспилотных транспортных средств. *Известия Кабардино-Балкарского научного центра РАН*, 2022, № 1 (105), с. 59–72. doi:10.35330/1991-6639-2022-1-105-59-73, EDN: FQPPBV
16. **Jia Y., Zhou S., Zeng Q., Li C., Chen D., Zhang K., Liu L., Chen Z.** The UAV path coverage algorithm based on the greedy strategy and ant colony optimization. *Electronics*, 2022, vol. 11, no. 17, pp. 2667. doi:10.3390/electronics11172667
17. **Файзуллин Р. Ф.** Потенциал генетических алгоритмов в задачах покрытия территории группой БЛА. *Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика»*, 2024, № 1, с. 36–50. doi:10.28995/2686-679X- 2024-1-36-50, EDN: DPOXPG
18. **Wen X., Ruan Y., Li Y., Xia H., Zhang R., Wang C., Liu W., Jiang X.** Improved genetic algorithm based 3-D deployment of UAVs. *Journal of Communications and Networks*, 2022, vol. 24, no. 2, pp. 223–231. doi:10.23919/JCN.2022.000014
19. **Перепелкин Д. А., Нгуен В. Т.** Интеллектуальная многопутевая маршрутизация в программно-конфигурируемых сетях на основе алгоритма искусственной пчелиной колонии. *Информационные технологии*, 2022, т, 28, № 8, с. 395–404. doi:10.17587/it.28.395-404, EDN: BONWWZ
20. **Егорова К. В.** Имитационная модель управления полетом группы беспилотных летательных аппаратов на основе алгоритма пчелиной колонии. *Вестник Воронежского государственного технического университета*, 2023, т. 19, № 2, с. 68–71. doi:10.36622/VSTU.2023.19.2.010, EDN: INIXQJ
21. **Chen Y., Li N., Zhong X., Xie W.** Joint trajectory and scheduling optimization for the mobile UAV aerial base station: A fairness version. *Applied Sciences*, 2019, vol. 9, no. 15, pp. 3101. doi:10.3390/app9153101

UDC 621.396
doi:10.31799/1684-8853-2026-3-63-74
EDN: WJHXOQ

Optimization of the placement of unmanned aerial vehicles to cover the territory in trunking communication systems

V. S. Ivanov[a], PhD, Tech., Associate Professor, orcid.org/0000-0001-9827-1690, ivanovmirea1@yandex.ru
[a]MIREA – Russian Technological University, 78, Vernadsky Ave., 119454, Moscow, Russian Federation

Introduction: The relevance of deploying unmanned aerial vehicles (UAVs) for wireless coverage of the territory increases in conditions of difficult terrain affecting signal propagation. **Purpose:** To develop and comparatively analyze methods for minimizing the number of UAVs that guarantee coverage of a discrete area, taking into account the terrain, based on a combination of greedy heuristics and spatial decomposition strategies. **Results:** For each cell of the height grid, the maximum range of the UAV has been calculated using the Okumura – Hata model. We propose three greedy algorithms which differ in the criteria for choosing a position: maximizing the number of covered cells inside the area while minimizing the external coverage; external coverage priority minimizing; maximizing the radius without taking into account the boundaries. Based on these criteria, three variants of the decomposition of the working area (vertical, horizontal and quadrant partitioning with overlap) are implemented, in which each subdomain is processed independently and then solutions are combined. Experiments on 20×20 random height maps have shown that the basic algorithm with priority for covering internal cells uses an average of 41.5 UAVs and creates 85.2 external points. The algorithm that minimizes the external area reduces it to 70.5, but requires 55.9 UAVs. Maximizing the radius yields 41.2 UAVs with the highest external coverage (94.4). The quadrant division provides the best balance: 41.3 UAVs and 82.7 external points, which is 2.5 points less than the base while maintaining the number of vehicles. The vertical division reduces the number of UAVs to 40.7, but does not improve the external area, the horizontal one turns out to be the least effective. **Practical relevance:** The decomposition of an area with overlap and the choice of an optimal strategy for each subdistrict improve the quality of placement. The quadrant division demonstrates the best compromise between the number of UAVs and unwanted radiation, as well as the smallest variation in indicators. The proposed methods can serve as a basis for the formation of initial populations in genetic algorithms in order to find globally optimal configurations of UAV-based networks.

Keywords – UAVs, heuristic algorithms, wireless communication, trunking communication systems, placement optimization, air base station, genetic algorithm.

For citation: Ivanov V. S. Optimization of the placement of unmanned aerial vehicles to cover the territory in trunking communication systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 63–74 (In Russian). doi:10.31799/1684-8853-2026-3-63-74, EDN: WJHXOQ

References

1. Tran T. D., Koucheryavy A. E. Resource optimization of airborne base stations using artificial intelligence methods. *Proceedings of Telecommunication Universities*, 2025, vol. 11, no. 1. pp. 62–68 (In Russian). doi:10.31854/1813-324X-2025 -11-1-62-68
2. Fujio S., Ozaki K. Radio propagation environment prediction method for UAV-assisted cellular networks with millimeter-wave beamforming. *2024 International Conference on Computing, Networking and Communications (ICNC)*, 2024, pp. 957–961. doi:10.1109/ICNC59896.2024.10556093

3. Mozaffari M., Saad W., Bennis M., Debbah M. A tutorial on UAVs for wireless networks: Applications, challenges, and open problems. *IEEE Communications Surveys & Tutorials*, 2019, vol. 21, no. 3, pp. 2334–2360. doi:10.1109/COMST.2019.2902862
4. Pandey A., Gunjan V. Genetic algorithm-based optimization of UAV placement for fixed and dynamic target scenarios. *2025 International Conference on Electronics, AI and Computing (EAIC)*, India, 2025, pp. 1–8. doi:10.1109/EAIC66483.2025.11101363
5. Florin R., Grozea C., Enache M., Nelega R., Kovacs G., Puschita R. Mission-critical services in 4G/5G and beyond: Standardization, key challenges, and future perspectives. *Sensors*, 2025, vol. 25, no. 16, pp. 1–48. doi:10.3390/s25165156
6. Kumbhar A., Koohifar F., Güvenç İ., Mueller B. A survey on legacy and emerging technologies for public safety communications. *IEEE Communications Surveys & Tutorials*, 2017, vol. 19, no. 1, pp. 97–124. doi:10.1109/COMST.2016.2612223
7. Zeng Y., Zhang R., Lim T. J. Wireless communications with unmanned aerial vehicles: Opportunities and challenges. *IEEE Communications Magazine*, 2016, vol. 54, no. 5, pp. 36–42. doi:10.1109/MCOM.2016.7470933
8. Khuwaja A. A., Chen Y., Zhao N., Alouini M. S., Dobbins P. A survey of channel modeling for UAV communications. *IEEE Communications Surveys & Tutorials*, 2018, vol. 20, no. 4, pp. 2804–2821. doi:10.1109/comst.2018.2856587
9. Mahbub M., Saym M. M., Jahan S., Paul A. K., Niyato D. UAV-assisted wireless communications in the 6G-and-beyond era: An extensive survey on characteristics, standardization and regulations, enabling technologies, challenges, and future directions. *Vehicular Communications*, 2025, vol. 15, no. 3, pp. 45–67. doi:10.1016/j.vehcom.2025.100977
10. Kalantari E., Yanikomeroglu H., Yongacoglu A. On the number and 3D placement of drone base stations in wireless cellular networks. *Proc. of IEEE Vehicular Technology Conference (VTC2016-Fall)*, Montreal, QC, Canada, 2016, pp. 1–6. doi:10.1109/VTCFall.2016.7881122
11. Savkin A. V., Huang H. Proactive deployment of aerial drones for coverage over very uneven terrains: A version of the 3D art gallery problem. *Sensors*, 2019, vol. 19, no. 6, Article 1438. doi:10.3390/s19061438
12. Ruz-Nieto A., Egea-López E., Molina-García-Pardo J.-M., Santa J. A 3D simulation framework with ray-tracing propagation for LoRaWAN communication. *Internet of Things*, 2023, vol. 24, Article 100964. doi:10.1016/j.iot.2023.100964
13. Ivanov V. S., Uvaisov S. U., Ivanov I. A. Automatic placement algorithm of base stations trunking communication systems. *Proceedings of Telecommunication Universities*, 2023, vol. 9, no. 5, pp. 25–34 (In Russian). doi:10.31854/1813-324X-2023-9-5-25-34, EDN: JMSIAV
14. Ali N., Al-Behadili H. A. H. A comprehensive review of radio signal propagation prediction for terrestrial wireless communication systems. *Misan Journal of Engineering Sciences*, 2024, vol. 3, no. 1, pp. 100–120. doi:10.61263/mjes.v3i1.80
15. Mamchenko M. V., Zorin V. A., Romanova M. A. Empirical model for propagation loss using floor space index for unmanned vehicles. *News of the Kabardino-Balkarian Scientific Center of RAS*, 2022, no. 1 (105), pp. 59–72 (In Russian). doi:10.35330/1991-6639-2022-1-105-59-73, EDN: FQPPBV
16. Jia Y., Zhou S., Zeng Q., Li C., Chen D., Zhang K., Liu L., Chen Z. The UAV path coverage algorithm based on the greedy strategy and ant colony optimization. *Electronics*, 2022, vol. 11, no. 17, pp. 2667. doi:10.3390/electronics11172667
17. Faizullin R. F. Potential of genetic algorithms in tasks of the territory coverage by a group of UAVs. *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, 2024, no. 1, pp. 36–50 (In Russian). doi:10.28995/2686-679X-2024-1-36-50, EDN: DPOXPG
18. Wen X., Ruan Y., Li Y., Xia H., Zhang R., Wang C., Liu W., Jiang X. Improved genetic algorithm based 3-D deployment of UAVs. *Journal of Communications and Networks*, 2022, vol. 24, no. 2, pp. 223–231. doi:10.23919/JCN.2022.000014
19. Perepelkin D. A., Nguyen V. T. Intelligent multipath routing in software-configurable networks based on an artificial bee colony algorithm. *Information Technologies*, 2022, vol. 28, no. 8, pp. 395–404 (In Russian). doi:10.17587/it.28.395-404, EDN: BONWWZ
20. Egorova K. V. Simulation model of flight control of a group of unmanned aerial vehicles based on the bee colony algorithm. *Bulletin of Voronezh State Technical University*, 2023, vol. 19, no. 2, pp. 68–71 (In Russian). doi:10.36622/VSTU.2023.19.2.010, EDN: INIXQJ
21. Chen Y., Li N., Zhong X., Xie W. Joint trajectory and scheduling optimization for the mobile UAV aerial base station: A fairness version. *Applied Sciences*, 2019, vol. 9, no. 15, pp. 3101. doi:10.3390/app9153101

ПАМЯТКА ДЛЯ АВТОРОВ

Поступающие в редакцию статьи проходят обязательное рецензирование.

При наличии положительной рецензии статья рассматривается редакционной коллегией. Принятая в печать статья направляется автору для согласования редакторских правок. После согласования автор представляет в редакцию окончательный вариант текста статьи.

Процедуры согласования текста статьи могут осуществляться как непосредственно в редакции, так и по e-mail (ius.spb@gmail.com).

При отклонении статьи редакция представляет автору мотивированное заключение и рецензию, при необходимости доработать статью — рецензию.

Редакция журнала напоминает, что ответственность за достоверность и точность рекламных материалов несут рекламодатели.

БЕЛОВ
Андрей
Александрович

Ведущий инженер Высшей школы прикладной физики и космических технологий Института электроники и телекоммуникаций Санкт-Петербургского политехнического университета Петра Великого.
В 1989 году окончил Ленинградский политехнический институт им. М. И. Калинина по специальности «Радиофизика и электроника».
Является автором 30 научных публикаций и пяти патентов на изобретения.
Область научных интересов — обработка сигналов и изображений, СВЧ-электроника, радиолокация, дефектоскопия.
Эл. адрес: belov@spbstu.ru

ВОРОБЬЕВ
Александр
Александрович

Старший преподаватель кафедры автоматики, телемеханики, связи и вычислительной техники Донецкого института железнодорожного транспорта.
В 2017 году окончил Донецкий институт железнодорожного транспорта по специальности «Системы обеспечения движения поездов».
Является автором пяти научных публикаций.
Область научных интересов — искусственный интеллект, нейронные сети.
Эл. адрес:
vorobyov94@gmail.com

ИВАНОВ
Вячеслав
Сергеевич

Доцент кафедры конструирования и производства радиоэлектронных средств МИРЭА — Российского технологического университета, Москва.
В 2020 году окончил МИРЭА — Российский технологический университет по специальности «Радиоэлектронные системы и комплексы».
В 2024 году защитил диссертацию на соискание ученой степени кандидата технических наук.
Является автором 40 научных публикаций и пяти свидетельств интеллектуальной собственности.
Область научных интересов — беспроводные системы связи, печатные платы, аддитивные технологии.
Эл. адрес: ivanovmirea1@yandex.ru

ВОЛВЕНКО
Сергей
Валентинович

Старший научный сотрудник, заведующий научной лабораторией «СТЦ-Политех» Санкт-Петербургского государственного политехнического университета Петра Великого.
В 1995 году окончил Томский государственный университет им. В. В. Куйбышева по специальности «Радиофизика и электроника».
Является автором 97 научных публикаций и 14 патентов на изобретения.
Область научных интересов — спектрально-эффективные сигналы, цифровая обработка сигналов, метеорные системы связи, сверхширокополосные сигналы.
Эл. адрес: svolvenko@spbstu.ru

ГУ
Линюнь

Инженер-исследователь Санкт-Петербургского политехнического университета Петра Великого.
В 2018 году окончила факультет автоматизации Пекинского технологического университета по специальности «Автоматическое управление и распознавание образов», в 2021 — Школу аэронавтики и астронавтики Университета Цинхуа по специальности «Проектирование летательных аппаратов», Китай.
В 2026 году защитила диссертацию на соискание ученой степени кандидата технических наук.
Является автором 13 научных публикаций.
Область научных интересов — обработка дистанционных изображений, обнаружение объектов, дистилляция знаний, оптимизация нейронных сетей и др.
Эл. адрес: gulingyun7@gmail.com

КАШЕВНИК
Алексей
Михайлович

Доцент, старший научный сотрудник Санкт-Петербургского Федерального исследовательского центра РАН.
В 2005 году окончил Санкт-Петербургский политехнический университет по специальности «Автоматизированные системы обработки информации и управления».
В 2008 году защитил диссертацию на соискание ученой степени кандидата технических наук.
Является соавтором более 200 научных публикаций и двух патентов на изобретения.
Область научных интересов — искусственный интеллект, мониторинг человека, компьютерное зрение, интеллектуальный анализ данных.
Эл. адрес:
alexey.kashevnik@iias.spb.su

КУЧКОВ Алексей Васильевич

Аспирант факультета информационных технологий и программирования Университета ИТМО, Санкт-Петербург.
В 2024 году окончил магистратуру Балтийского государственного технического университета по направлению «Лазерная техника».
Является автором двух научных публикаций.
Область научных интересов — искусственный интеллект, мультимодальные модели, компьютерное зрение.
Эл. адрес: kuchkovaleksei@gmail.com

ПАВЛОВ Виталий Александрович

Доцент Высшей школы прикладной физики и космических технологий Института электроники и телекоммуникаций Санкт-Петербургского политехнического университета Петра Великого.
В 2012 году окончил Санкт-Петербургский государственный политехнический университет по специальности «Защищенные системы связи».
В 2020 году защитил диссертацию на соискание ученой степени кандидата технических наук.
Является автором 50 научных публикаций.
Область научных интересов — обработка сигналов, обработка изображений, компьютерное зрение, машинное обучение, глубокое обучение.
Эл. адрес: pavlov_va@spbstu.ru

РАСКОПИНА Анастасия Сергеевна

Аспирант Института информационных технологий и программирования Санкт-Петербургского государственного университета аэрокосмического приборостроения.
В 2024 году окончила магистратуру Санкт-Петербургского государственного университета аэрокосмического приборостроения по специальности «Прикладная информатика».
Является автором 15 научных публикаций.
Область научных интересов — нейронные сети, искусственный интеллект, оптимизация глубокого обучения.
Эл. адрес: raskopina.anastasia@yandex.ru

ОСИПОВ Дмитрий Сергеевич

Старший научный сотрудник лаборатории информационных технологий передачи, анализа и защиты данных Института проблем передачи информации им. А. А. Харкевича РАН, Москва.
В 2003 году окончил Московский государственный технический университет по специальности «Системы автоматического управления».
В 2023 году защитил диссертацию на соискание ученой степени доктора технических наук.
Является автором 35 научных публикаций.
Область научных интересов — теория передачи информации, разработка и исследование моделей систем множественного доступа, технологии защиты данных, передаваемых по беспроводным каналам связи и др.
Эл. адрес: d_osipov@iitp.ru

РАДКОВСКИЙ Сергей Александрович

Доцент, заведующий кафедрой автоматики, телемеханики, связи и вычислительной техники Донецкого института железнодорожного транспорта.
В 1999 году окончил Харьковскую государственную академию железнодорожного транспорта по специальности «Автоматика и телемеханика на железнодорожном транспорте».
В 2004 году защитил диссертацию на соискание ученой степени кандидата технических наук.
Является автором 34 научных публикаций.
Область научных интересов — автоматизированные системы управления, теория искусственных нейронных сетей, сетевое и системное администрирование.
Эл. адрес: serj_rsa@mail.ru

САЦЮК Александр Владимирович

Доцент кафедры автоматики, телемеханики и вычислительной техники, руководитель лаборатории искусственного интеллекта Донецкого института железнодорожного транспорта.
В 2009 году окончил Донецкий институт железнодорожного транспорта Украинской государственной академии железнодорожного транспорта по специальности «Автоматика и автоматизация на транспорте».
В 2019 году защитил диссертацию на соискание ученой степени кандидата технических наук.
Является автором 43 научных публикаций и трех патентов на изобретения.
Область научных интересов — интеллектуальные системы управления, автономные транспортные системы, глубокое обучение и нейронные сети в компьютерном зрении и др.
Эл. адрес: alexandrsatsuk@gmail.com

СОНИНА Светлана Дмитриевна

Старший преподаватель кафедры автоматики, телемеханики, связи и вычислительной техники Донецкого института железнодорожного транспорта.
В 2011 году окончила Донецкий государственный технический университет по специальности «Специализированные компьютерные системы».
Является автором 21 научной публикации.
Область научных интересов — теория искусственных нейронных сетей, методы оптимизации и адаптации параметров обучения нейросетевых моделей, нейросетевое моделирование цифровых устройств.
Эл. адрес: soninadonigt@yandex.com

ШАРИАТИ Фаридоддин

Старший преподаватель Высшей школы прикладной физики и космических технологий Института электроники и телекоммуникаций Санкт-Петербургского политехнического университета Петра Великого.
В 2021 году окончил Санкт-Петербургский политехнический университет Петра Великого по специальности «Инфокоммуникационные технологии и системы связи».
Является автором 35 научных публикаций.
Область научных интересов — искусственный интеллект, обработка изображений, автоматизированные системы диагностики.
Эл. адрес: shariati_f@spbstu.ru

ТАТАРНИКОВА Татьяна Михайловна

Профессор, директор Института информационных технологий и программирования Санкт-Петербургского государственного университета аэрокосмического приборостроения.
В 1993 году окончила Восточно-Сибирский технологический институт по специальности «Электронно-вычислительные машины, комплексы, системы и сети».
В 2007 году защитила диссертацию на соискание ученой степени доктора технических наук.
Является автором более 300 научных публикаций.
Область научных интересов — инфокоммуникации, взаимодействие неоднородных сетей, анализ данных.
Эл. адрес: tm-tatarn@yandex.ru

ШЕХОВЦОВ Алексей Игоревич

Декан факультета управления на железнодорожном транспорте, доцент кафедры организации перевозок и управления на железнодорожном транспорте Донецкого института железнодорожного транспорта.
В 2003 году окончил Донецкий институт железнодорожного транспорта Украинской государственной академии железнодорожного транспорта по специальности «Организация перевозок и управление на транспорте».
В 2020 году защитил диссертацию на соискание ученой степени кандидата технических наук.
Является автором 108 научных публикаций и одного патента на полезную модель.
Область научных интересов — совершенствование работы транспортных систем, автоматизация технологических процессов.
Эл. адрес: oleksa.i@mail.ru

Уважаемые авторы!

При подготовке рукописей статей необходимо руководствоваться следующими рекомендациями.

Статьи должны содержать изложение новых научных результатов. Название статьи должно быть кратким, но информативным. В названии недопустимо использование сокращений, кроме самых общепринятых (РАН, РФ, САПР и т. п.).

Текст рукописи должен быть оригинальным, а цитирование и самоцитирование корректно оформлено.

Объем статьи (текст, таблицы, иллюстрации и библиография) не должен превышать эквивалента в 20 страниц, напечатанных на бумаге формата А4 на одной стороне через 1,5 интервала Word шрифтом Times New Roman размером 13, поля не менее двух сантиметров.

Обязательными элементами оформления статьи являются: индекс УДК, заглавие, инициалы и фамилия автора (авторов), ученая степень, звание (при отсутствии — должность), полное название организации, аннотация и ключевые слова на русском и английском языках, ORCID и электронный адрес одного из авторов. При написании аннотации не используйте аббревиатур и не делайте ссылок на источники в списке литературы. Предоставляйте подрисуночные подписи и названия таблиц на русском и английском языках.

Статьи авторов, не имеющих ученой степени, рекомендуется публиковать в соавторстве с научным руководителем, наличие подписи научного руководителя на рукописи обязательно; в случае самостоятельной публикации обязательно предоставляйте заверенную по месту работы рекомендацию научного руководителя с указанием его фамилии, имени, отчества, места работы, должности, ученого звания, ученой степени.

Простые **формулы** набирайте в Word, сложные с помощью редактора Mathtype или Equation. Для набора одной формулы не используйте два редактора; при наборе формул в формульном редакторе знаки препинания, ограничивающие формулу, набирайте вместе с формулой; для установки размера шрифта в Mathtype никогда не пользуйтесь вкладкой Other, Smaller, Larger, используйте заводские установки редактора, не подгоняйте размер символов в формулах под размер шрифта в тексте статьи, не растягивайте и не сжимайте мышью формулы, вставленные в текст; пробелы в формуле ставьте только после запятых при перечислении с помощью Ctrl+Shift+Space (пробел); не отделяйте пробелами знаки: + = – ×, а также пространство внутри скобок; для выделения греческих символов в Mathtype полужирным начертанием используйте Style → Other → bold.

Для набора формул в Word никогда не используйте вкладки: «Уравнение», «Конструктор», «Формула» (на верхней панели: «Вставка» — «Уравнение»), так как этот ресурс предназначен только для внутреннего использования в Word и не поддерживается программами, предназначенными для изготовления оригинал-макета журнала.

При наборе символов в тексте помните, что символы, обозначаемые латинскими буквами, набираются светлым курсивом, русскими и греческими — светлым прямым, векторы и матрицы — прямым полужирным шрифтом.

Подробнее см. http://i-us.ru/index.php/ius/author-guide

Иллюстрации:

— рисунки, графики, диаграммы, блок-схемы предоставляйте в виде отдельных исходных файлов, поддающихся редактированию, используя векторные программы: Visio (*.vsd, *.vsdx); Adobe Illustrator (*.ai); Coreldraw (*.cdr, версия не выше 15); Excel (*.xls); Word (*.docx); AutoCad, Matlab (экспорт в PDF, EPS, SVG, WMF, EMF); Компас (экспорт в PDF); веб-портал DRAW.IO (экспорт в PDF); Inkscape (экспорт в PDF);

— фото и растровые — в формате *.tif, *.png с максимальным разрешением (не менее 300 pixels/inch).

Наличие подрисуночных подписей и названий таблиц на русском и английском языках обязательно (желательно не повторяющих дословно комментарии к рисункам в тексте статьи).

В редакцию предоставляются:

— сведения об авторе (фамилия, имя, отчество, место работы, должность, ученое звание, учебное заведение и год его окончания, ученая степень и год защиты диссертации, область научных интересов, количество научных публикаций, домашний и служебный адреса и телефоны, e-mail), фото авторов: анфас, в темной одежде на белом фоне, должны быть видны плечи и грудь, высокая степень четкости изображения без теней и отблесков на лице, фото можно представить в электронном виде в формате *.tif, *.png, *.jpg с максимальным разрешением — не менее 300 pixels/inch при минимальном размере фото 40×55 мм;

— экспертное заключение;

— экспортное заключение.

Список литературы составляется по порядку ссылок в тексте и оформляется следующим образом:

— для книг и сборников — фамилия и инициалы авторов, полное название книги (сборника), город, издательство, год, общее количество страниц, doi;

— для журнальных статей — фамилия и инициалы авторов, полное название статьи, название журнала, год издания, номер журнала, номера страниц, doi;

— ссылки на иностранную литературу следует давать на языке оригинала без сокращений;

— при использовании web-материалов указывайте адрес сайта и дату обращения.

Список литературы оформляйте двумя отдельными блоками по образцам lit.dot на сайте журнала (http://i-us.ru/paperrules): Литература и References.

Более подробно правила подготовки текста с образцами изложены на нашем сайте в разделе «Руководство для авторов» – http://i-us.ru/index.php/ius/author-guide.

Контакты

Куда: 190000, г. Санкт-Петербург, ул. Большая Морская, д. 67, лит. А, ГУАП, РИЦ
Кому: Редакция журнала «Информационно-управляющие системы»
Тел.: (812) 494-70-02
Эл. почта: ius.spb@gmail.com
Сайт: www.i-us.ru