ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья искусственный интеллект и машинное обучение

УДК 004.932.75'1, 004.89
10.25209/2079-3316-2025-16-4-173-216

Сравнительный анализ архитектур backbone для инстанс-сегментации объектов на аэрофотоснимках с использованием Mask R-CNN

Игорь Викторович **Винокуров**[1✉], Дарья Александровна **Фролова**[2], Андрей Иванович **Ильин**[3], Иван Романович **Кузнецов**[4]

[1–4] Финансовый Университет при Правительстве Российской Федерации, Москва, Россия

[1✉] *igvvinokurov@fa.ru*

Аннотация. В работе проведено сравнительное исследование моделей Mask R-CNN с различными предобученными backbone-архитектурами для реализации инстанс-сегментации объектов недвижимости на аэрофотоснимках. Модели дообучались на специализированном наборе данных ППК «Роскадастр».

Анализ точности детектирования ограничивающих рамок и масок сегментации объектов выявил предпочтительные архитектуры – трансформеры Swin (Swin-S и Swin-T) и свёрточная сеть ConvNeXt-T. Высокая точность этих моделей объясняется их способностью учитывать глобальные контекстные зависимости между элементами изображения.

Результаты исследования позволяют сформулировать следующие рекомендации по выбору архитектуры backbone: для систем мониторинга в реальном времени, где критична скорость работы, целесообразно применение легковесных моделей (EfficientNet-B3, ConvNeXt-T, Swin-T), для offline задач, требующих максимальной точности (таких как картирование объектов недвижимости), рекомендована крупномасштабная модель Swin-S.

(*Здесь только русскоязычная часть оригинальной двуязычной статьи*)

Ключевые слова и фразы: инстанс-сегментация, backbone, Mask R-CNN, ResNet, DenseNet, EfficientNet, ConvNeXt, Swin

Для цитирования: Винокуров И. В., Фролова Д. А., Ильин А. И., Кузнецов И. Р. *Сравнительный анализ архитектур backbone для инстанс-сегментации объектов на аэрофотоснимках с использованием Mask R-CNN* // Программные системы: теория и приложения. 2025. **Т. 16**. № 4(67). С. 173–216. https://psta.psiras.ru/read/psta2025_4_173-216.pdf

Введение

Автоматизация анализа аэрофотоснимков является одной из ключевых задач современной геоинформатики, урбанистики, экологического мониторинга и управления территориями. Среди множества задач в этой области, задача инстанс-сегментации, заключающаяся в обнаружении и точном попиксельном (*pixel-wise*) выделении каждого объекта на изображении, занимает центральное место. Её решение позволяет автоматически идентифицировать и картировать на аэрофотоснимках, полученных с помощью летательных аппаратов, такие объекты, как здания, элементы дорожной инфраструктуры, сельскохозяйственные объекты и т.п.

Актуальность данной работы обусловлена растущим объёмом данных аэрофотосъемки и острой потребностью в их оперативном анализе. Однако разработка автоматических методов сталкивается с рядом сложностей, присущих именно аэрофотоснимкам – высокий динамический диапазон, значительная вариативность масштабов и ракурсов съемки, сложные погодные условия, оптические искажения, а также часто ограниченный объём экспертно размеченных данных для обучения моделей машинного обучения. Эти факторы предъявляют особые требования к надёжности и обобщающей способности моделей сегментации.

В последние годы модель Mask R-CNN [1] зарекомендовала себя в качестве фактического стандарта для решения задач инстанс-сегментации, демонстрируя стабильно высокие результаты в одновременном детектировании объектов и построении их точных бинарных масок. Её архитектура, расширяющая Faster R-CNN [2] за счёт добавления параллельной ветви для предсказания масок, обеспечивает эффективное сочетание точности локализации и качества семантической сегментации. Эта особенность делает Mask R-CNN особенно востребованной в приложениях, требующих не только обнаружения, но и точного контурного описания объектов сложной формы.

Итоговая эффективность этой модели в значительной степени определяется выбором архитектуры экстрактора признаков или backbone – глубокой свёрточной или трансформерной сети, ответственной за извлечение иерархических признаков из изображения. Относительно долгое время доминировали backbone-архитектуры семейства ResNet [3].

Однако появление более современных и эффективных CNN, таких как DenseNet [4], EfficientNet [5] и ConvNeXt [6], а также революционный прорыв архитектур на основе трансформеров, в частности Swin [7], кардинально изменили положение дел и расширили выбор исследователей. В связи с этим проведение систематического сравнительного анализа этих архитектур в контексте специфической и значимой задачи сегментации на аэрофотоснимках представляет значительный практический и научный интерес.

В рамках настоящего исследования не рассматривались некоторые современные и перспективные архитектуры, такие как Vision Transformer [8] в её базовой конфигурации, а также новейшие эффективные модели типа MobileOne [9], EdgeNeXt [10] или рекуррентные сетевые структуры, ориентированные на последовательную обработку признаков.

Кроме того, остались за рамками работы гибридные подходы, сочетающие свёрточные слои с механизмами внимания в рамках одного блока, например, модели семейства CoAtNet [11]. Это связано с акцентом исследований на широко распространенных и практически апробированных архитектурах, а также с необходимостью обеспечения репрезентативности и сопоставимости результатов. Указанные направления представляют интерес для будущих исследований в контексте оптимизации точности и вычислительной эффективности для задач анализа аэрофотоснимков.

Исследования, проведённые в этой работе, продолжают и развивают подходы, сформулированные в [12] и [13], предлагая альтернативное решение задачи, рассмотренной в [13]. В [12] были систематизированы целевые классы объектов, подлежащих идентификации на материалах аэрофотосъёмки, а также разработана комплексная методика формирования репрезентативного набора данных с детализированной семантической разметкой.

В работе [13] был реализован подход к повышению точности генерации масок объектов на основе генеративно-состязательных сетей (GAN), с фокусом на задаче постобработки результатов сегментации и субпиксельного уточнения контурных характеристик объектов. В отличие от данного подхода, в настоящей работе предлагается альтернативная методология, основанная на сравнительном анализе различных backbone-архитектур для модели Mask R-CNN, направленная на повышение точности сегментации на этапе первичного прогнозирования, а не последующей постобработки результатов.

Целью настоящего исследования является проведение сравнительного анализа точности и эффективности семи различных предобученных backbone-архитектур (ResNet-50, ResNet-101, DenseNet-121, EfficientNet-B3, ConvNeXt-T, Swin-T, Swin-S) в рамках модели Mask R-CNN для задачи инстанс-сегментации объектов на аэрофотоснимках.

Выбор этих архитектур backbone для сравнительного анализа обусловлен необходимостью охватить репрезентативный спектр современных подходов к проектированию глубоких нейронных сетей, обеспечив сопоставимость результатов и проверку ключевых исследовательских гипотез:

ResNet-50 и *ResNet-101* [3] реализуют концепцию остаточного обучения, которая решает проблему исчезающих градиентов посредством введения skip-connections, обеспечивающих стабильный поток градиентов в процессе обучения и позволяющих эффективно масштабировать глубину сети. Они включены как общепринятый базовый стандарт в компьютерном зрении, позволяющий оценить влияние глубины сети на качество сегментации.

DenseNet-121 [4] использует парадигму плотных соединений, в рамках которой каждый слой получает прямой доступ к картам признаков всех предшествующих слоёв. Она интересна как архитектура с плотными соединениями для интенсивного повторного использования признаков, снижения количества параметров и улучшения градиентного потока.

EfficientNet-B3 [5] применяет стратегию составного масштабирования (compound scaling), оптимизирующую глубину, ширину и разрешение входных данных и позволяющую достичь оптимального баланса между точностью предсказания и вычислительной сложностью модели при заданных ресурсных ограничениях.

ConvNeXt-T [6] представляет собой современную интерпретацию классической свёрточной архитектуры, объединяющую наиболее успешные решения из области трансформеров. Архитектура предполагает использование увеличенного размера ядра свёртки, усовершенствованных методов нормализации и активации, что в совокупности обеспечивает конкурентоспособную производительность.

Swin-T и *Swin-S* [7] относятся к классу иерархических трансформеров, использующих механизм самовнимания в рамках локальных окон с последующим их сдвигом. Такой подход позволяет эффективно моделировать глобальные контекстные зависимости при сохранении линейной вычислительной сложности относительно размера изображения, что особенно значимо для обработки данных аэрофотосъёмки высокого разрешения.

Такой отбор обеспечивает покрытие как классических, так и современных парадигм, позволяя системно оценить влияние архитектурных инноваций на точность и эффективность инстанс-сегментации в условиях аэрофотосъёмки.

1. Определение цели и задач исследований

Анализ современных публикаций выявил существование различных подходов к использованию данных архитектур в составе модели Mask R-CNN. В работе [14] демонстрируется преимущество трансформерных backbones для задач точной сегментации, что отражает тенденцию

перехода к архитектурам на основе механизма внимания, способным эффективно моделировать глобальные контекстные зависимости [15, 16].

В [17], наоборот, показывается эффективность современных CNN для решения многих практических задач; такой подход находит подтверждение в работах, посвященных оптимизации lightweight-архитектур для задач реального времени [18].

Сравнительный анализ в [19] выявляет компромисс между точностью и вычислительной сложностью различных архитектур, что соответствует методологии всесторонней оценки современных моделей, учитывающей не только метрики точности, но и скорость инференса, потребление памяти и вычислительную сложность [20].

При всём этом оценка эффективности использования различных backbone для задач анализа аэрофотоснимков, характеризующихся высокой вариативностью масштабов и необходимостью сегментации малых объектов, остаётся исследованной недостаточно.

Основные задачи работы:

(1) Адаптировать модель Mask R-CNN для работы с использованием различных backbone-архитектур.
(2) Реализовать дообучение (fine tuning) каждой из конфигураций модели на собственном наборе данных из аэрофотоснимков.
(3) Вычислить для каждой модели стандартные метрики `loss_bbox`, `loss_mask`, `bbox_mAP` и `segm_mAP`.
(4) Проанализировать полученные результаты, выявив закономерности между типом архитектуры и точностью детектирования.

2. Методология проведения эксперимента

Вычислительный эксперимент для проведения сравнительного анализа производительности различных архитектур backbone в модели Mask R-CNN был организован с использованием фреймворка MMDetection – открытой инструментальной библиотеки для детекции и сегментации объектов на основе PyTorch.

Сформированный для обучения и валидации моделей набор данных представлял собой совокупность из 435 полученных с помощью квадрокоптера аэрофотоснимков. В набор данных входит и такое же количество JSON-файлов в формате ПО LabelMe, содержащих массивы координат точек полигональной контуризации и имена (метки) объектов пяти типов – дачный домик (метка «building», 12470 экземпляров) экземпляров, теплица (метка «greenhouse», 6450 экземпляров), хозпостройка (метка «outbuilding», 2150 экземпляров), транспортное средство (метка «vehicle»,

1516 экземпляров), бассейн (метка «swimming», 490 экземпляров) [12]. Для обучения использовалось 75% элементов набора данных, для валидации оставшиеся 25%.

Обучение проходило на GPU A100 80GB.

Базовые конфигурационные файлы для каждого backbone были взяты из официального репозитория MMDetection (v3.3.0). Во всех конфигурационных файлах переопределялись (формировались) параметры `data_root` и `metainfo`. В первом из них переопределялось имя корневой папки набора данных, во втором формировалась информация о пяти целевых классах. Параметры `num_classes` в `bbox_head` и `mask_head` устанавливались по количеству целевых классов в 5.

Для корректного трансферного обучения модели инициализировались предобученными весами для набора данных COCO. Веса загружались из репозитория MMDetection (версии 2 или 3). Стратегия дообучения включала два этапа – на первых пяти эпохах производилась заморозка слоёв backbone с обучением только головных частей модели (`lr` = `1e-3`). Это позволило адаптировать классификационные слои к целевым классам.

После этого следовала полная настройка всех слоёв с пониженной скоростью обучения (`lr` = `1e` $-$ `4`) в течение 100 эпох. Критерием остановки служило отсутствие улучшения метрики `bbox_mAP` на валидационной выборке на протяжении 15 эпох.

Для оптимизации использовался AdamW с `weight_decay` = 0.05 (значение по умолчанию в MMDetection) и градиентным клиппингом с порогом 1.0 для стабилизации обучения. Валидация проводилась после каждой эпохи, лучшая модель сохранялась автоматически при улучшении метрики `bbox_mAP` на валидационном наборе. Все эксперименты проводились на единой аппаратной конфигурации с фиксированным `random_seed` = 42.

Оценка производительности осуществлялась на выделенном валидационном наборе с использованием метрик COCO `bbox_mAP`, `segm_mAP` для порогов IoU от 0.5 до 0.95 с шагом 0.05. Значения этих метрик рассчитывались для объектов больших (`mAP_l`), средних (`mAP_m`) и малых (`mAP_s`) размеров.

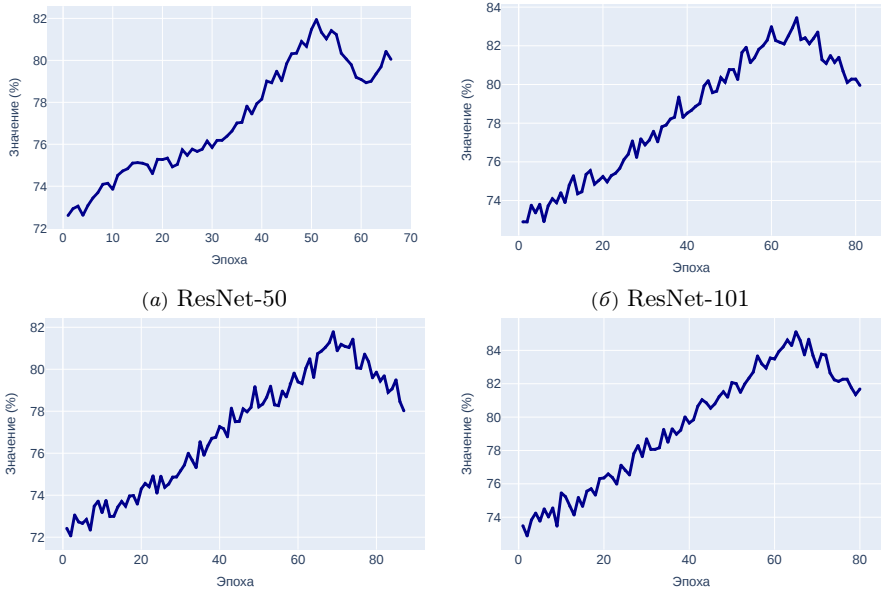
Площадь малых объектов составляет менее 32^2 пикселей ($<$1024 px^2), средних – от 32^2 до 96^2 пикселей, (1024–9216 px^2), больших – более 96^2 пикселей ($>$9216 px^2); `mAP` – среднее значение метрик для объектов детектирования всех размеров.

3. Результаты эксперимента

3.1. Анализ точности классификации ROI

Область интереса (Region of Interest, ROI) представляет собой выделенный регион на карте признаков (feature maps), которую генерирует backbone модели. В конвейерах детекции и сегментации объектов эти области в дальнейшем обрабатываются для классификации объектов (предсказания масок сегментации) и регрессии ограничивающих рамок.

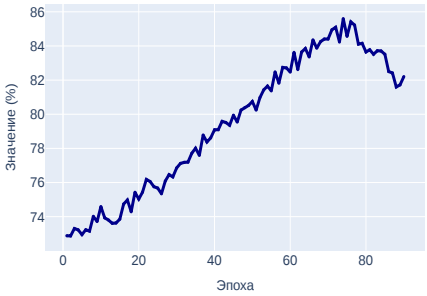
Проведенный анализ экспериментальных данных выявляет выраженную тенденцию к улучшению точности классификации ROI при переходе к современным трансформерным архитектурам (Swin-T, Swin-S) и усовершенствованным архитектурам CNN (ConvNeXt-T), что наглядно демонстрируется графиками динамики их обучения на рисунках 1 и 2.



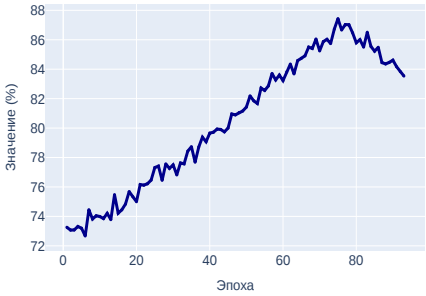
(*а*) ResNet-50 (*б*) ResNet-101

(*в*) DenseNet-121 (*г*) EfficientNet-B3

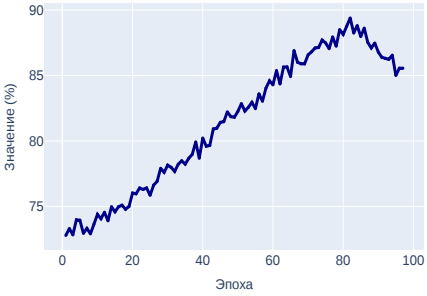
Рисунок 1. Точность классификации ROI для классических CNN архитектур



(*а*) ConvNeXt-T



(*б*) Swin-T



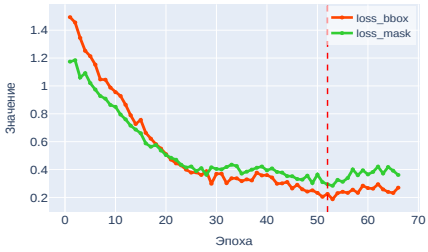
(*в*) Swin-S

Рисунок 2. Точность классификации ROI для трансформерных и усовершенствованной CNN архитектуры

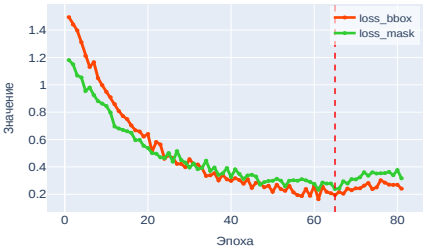
3.2. Анализ потерь при детектировании ограничивающих рамок и масок

Функции потерь для регрессии ограничивающих рамок (loss_bbox) и для детектирования масок (loss_mask) являются критически важными компонентами составной функции оптимизации в задачах инстанс-сегментации. Loss_bbox обеспечивает точное позиционирование объектов путём минимизации расхождения между предсказанными и эталонными координатами ограничивающих рамок. В свою очередь, loss_mask отвечает за точность попиксельной классификации внутри каждой ROI, гарантируя соответствие предсказанной маски контуру объекта.

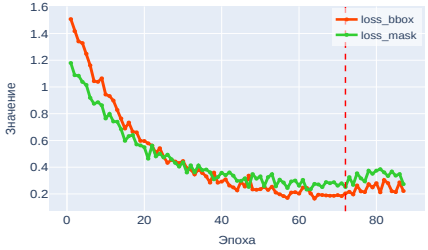
Из графиков рисунка 3 видно, что современные архитектуры, в частности трансформеры Swin-T и Swin-S, демонстрируют высокую эффективность в оптимизации функций потерь. Архитектура Swin-S достигает минимальных значений по обеим метрикам, что свидетельствует об её явном преимуществе благодаря способности эффективно моделировать



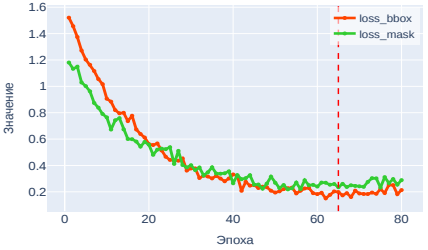
(*а*) ResNet-50



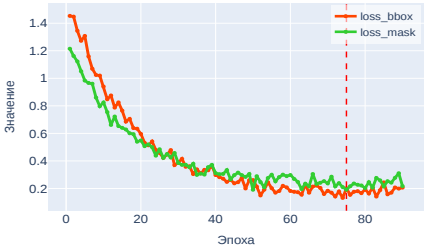
(*б*) ResNet-101



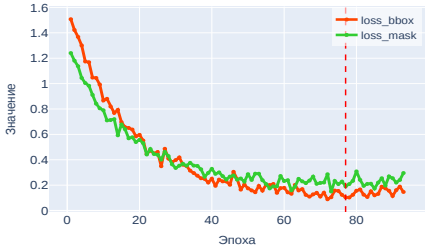
(*в*) DenseNet-121



(*г*) EfficientNet-B3



(*д*) ConvNeXt-T



(*е*) Swin-T

(*ж*) Swin-S

Рисунок 3. Потери регресии рамок и детектирования масок объектов для различных CNN архитектур

глобальные контекстные зависимости и извлекать иерархические признаки с высокой дискриминативной способностью, реализуя тем самым более точное позиционирование рамок и детализацию масок сегментации.

Наилучшие показатели Swin-S подтверждают перспективность использования трансформерных подходов для задач инстанс-сегментации, где критически важны как точность локализации, так и качество пиксельной классификации.

При этом таблица 1 показывает сопоставимую с DenseNet-121 эффективность моделей Swin-T и ConvNeXt-T, входящих в число наиболее сбалансированных решений по оптимизации обеих метрик.

Таблица 1. Значения функций потерь на валидационном наборе данных для различных архитектур backbone

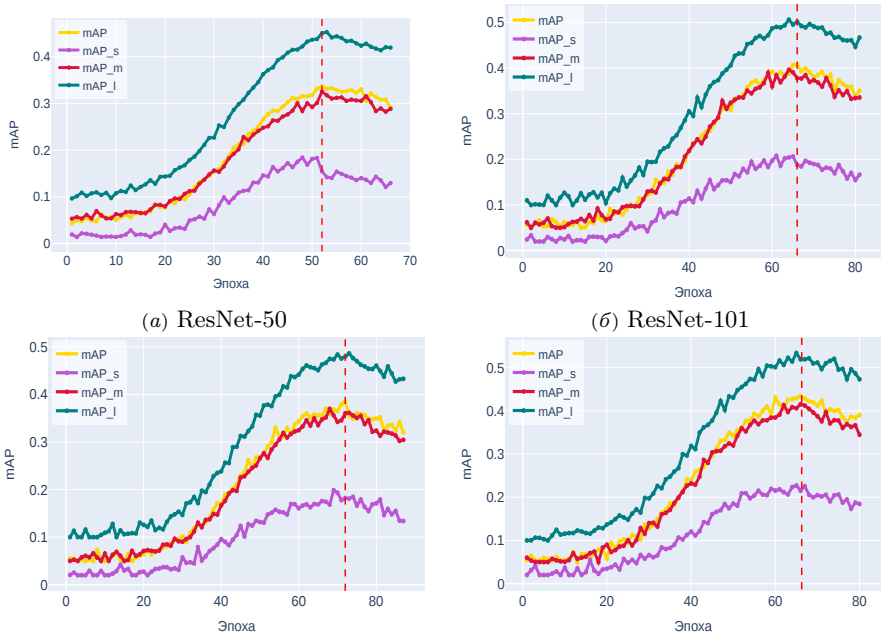
Backbone	loss_bbox	loss_mask
ResNet-50	0.2185	0.2891
ResNet-101	0.2137	0.3211
DenseNet-121	0.1854	0.2244
EfficientNet-B3	0.2136	0.2666
ConvNeXt-T	0.1793	0.2603
Swin-T	0.1813	0.2148
Swin-S	0.1020	0.1141

3.3. Анализ точности детектирования ограничивающих рамок

На рисунках 4 и 5 представлены графики значений метрик точности детектирования ограничивающих рамок bbox_mAP, полученные в результате экспериментальных исследований для объектов различных размеров – mAP_s, mAP_m и mAP_l, см. раздел 2.

Анализ экспериментальных данных выявляет последовательное улучшение метрики mAP в исследуемых архитектурах, показанное на таблице 2. В последнем столбце указана эпоха, на которой формируется архитектура backbone с наилучшими весами для детектирования ограничивающих рамок.

ResNet-50 демонстрирует базовый уровень точности ($mAP \approx 0.3236$), в то время как DenseNet-121, благодаря механизму плотных соединений, показывает заметный прирост (mAP ≈0.3800). ResNet-101, несмотря на увеличенную глубину, демонстрирует лишь умеренное улучшение (mAP ≈0.4080) по сравнению с DenseNet-121, что подчеркивает ограничения парадигмы простого наращивания глубины.



(*а*) ResNet-50 (*б*) ResNet-101

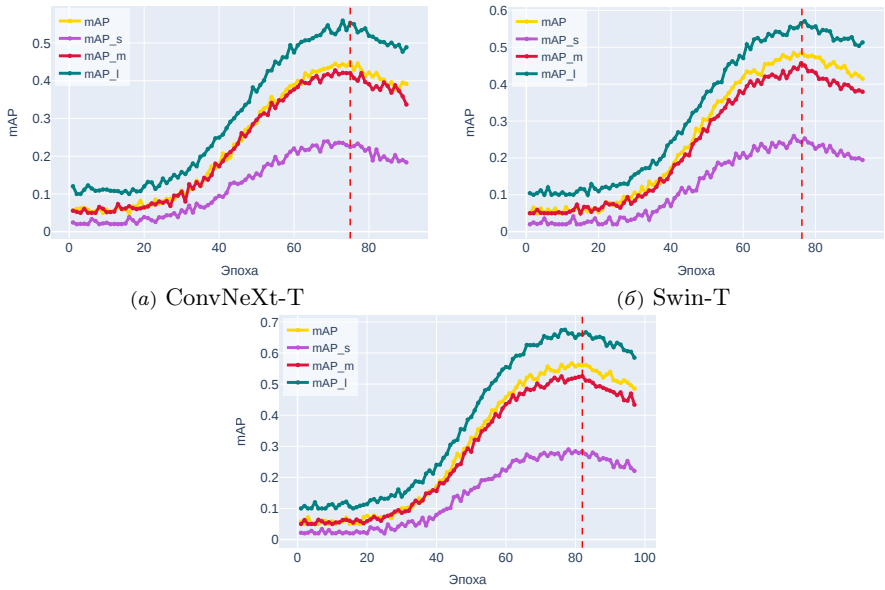
(*в*) DenseNet-121 (*г*) EfficientNet-B3

Рисунок 4. Точность детектирования ограничивающих рамок на валидационном наборе данных для классических CNN архитектур

Таблица 2. Метрики точности детектирования ограничивающих рамок для различных архитектур backbone

Backbone	mAP	mAP_s	mAP_m	mAP_l	Эпоха
ResNet-50	0.3236	0.1500	0.3200	0.4500	51
ResNet-101	0.4080	0.2000	0.3900	0.5000	66
DenseNet-121	0.3800	0.1800	0.3600	0.4800	71
EfficientNet-B3	0.4340	0.2200	0.4100	0.5300	64
ConvNeXt-T	0.4300	0.2300	0.4200	0.5400	76
Swin-T	0.4702	0.2400	0.4300	0.5500	77
Swin-S	0.5606	0.2800	0.5200	0.6500	81

Заметный скачок производительности у EfficientNet-B3 (mAP $\approx$0.4340) демонстрирует эффективность стратегии составного масштабирования. Увеличение точности у архитектуры ConvNeXt-T (mAP $\approx$0.4407) свидетель-



(*а*) ConvNeXt-T (*б*) Swin-T

(*в*) Swin-S

Рисунок 5. Точность детектирования ограничичвающих рамок на валидационном наборе данных для трансформерных и усовершенствованной CNN архитектуры

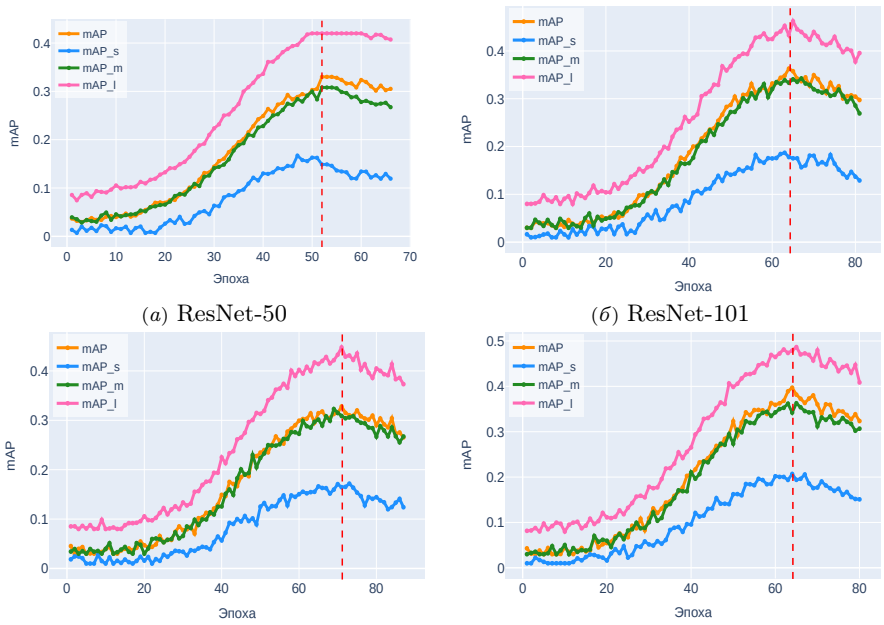
ствует о потенциале конволюционных сетей с улучшенной способностью к моделированию сложных пространственно-контекстных зависимостей.

Дальнейшее улучшение показывает Swin-T (mAP ≈0.4702), использующий механизм оконного внимания для эффективного моделирования глобальных контекстных взаимодействий при сохранении линейной вычислительной сложности. Наилучший результат демонстрирует Swin-S (mAP ≈0.5606), что наглядно иллюстрирует преимущества трансформерных архитектур и их исключительную эффективность в точной локализации объектов всех размерных категорий, особенно малых объектов, где наблюдается наиболее значимый прирост производительности.

3.4. Анализ точности детектирования масок сегментации

На рисунках 6 и 7, представлены графики значений метрик точности детектирования масок сегментации segm_mAP для объектов различных размеров, см. раздел 2.

Начальный уровень точности детектирования демонстрирует ResNet-50 (mAP ≈0.3000), DenseNet-121 позволяет достичь определённого улучшения



(*а*) ResNet-50 (*б*) ResNet-101

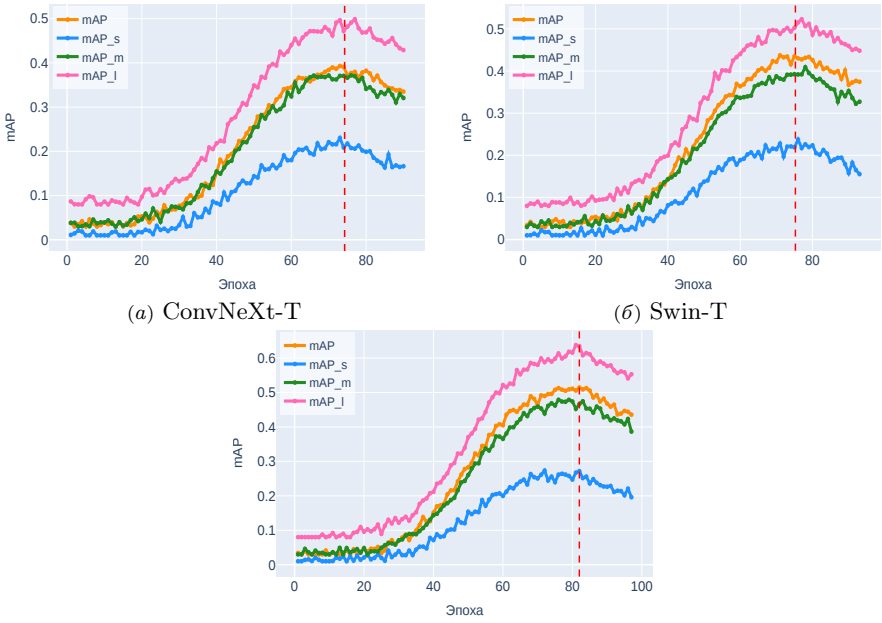
(*в*) DenseNet-121 (*г*) EfficientNet-B3

Рисунок 6. Точность детектирования масок сегментации на валидационном наборе данных для классических CNN архитектур

показателей (mAP ≈0.3121) за счёт эффективного переиспользования признаков через плотные соединения. Увеличение глубины сети, реализованное в ResNet-101, обеспечивает лишь незначительный прирост качества сегментации (mAP ≈0.3561), что свидетельствует об ограничении данного подхода. Применение составного масштабирования в EfficientNet-B3 позволяет стабилизировать показатели на уровне mAP ≈0.3795, однако настоящий качественный скачок наблюдается при переходе к современным архитектурным решениям.

Усовершенствованная CNN архитектура ConvNeXt-T демонстрирует существенное улучшение (mAP ≈0.3948), превосходя традиционные подходы за счёт оптимизации процесса извлечения пространственных признаков. Трансформерная архитектура Swin-T с механизмом оконного внимания показывает дальнейшее улучшение (mAP ≈0.4332), эффективно комбинируя локальные и глобальные контекстные зависимости.

Наивысшую эффективность в задаче инстанс-сегментации реализует трансформерная архитектура Swin-S (mAP ≈0.5030), устанавливающая



(*а*) ConvNeXt-T (*б*) Swin-T

(*в*) Swin-S

Рисунок 7. Точность детектирования масок сегментации на валидационном наборе данных для трансформерных и усовершенствованной CNN архитектуры

новый стандарт точности для объектов всех размерных категорий. Существенно значимые результаты наблюдаются для сегментации малых объектов, где данный подход демонстрирует наиболее значительное преимущество над другими архитектурами, см. таблицу 3. В последнем столбце указана эпоха, на которой формируется архитектура backbone с наилучшими весами для детектирования масок сегментации.

3.5. Анализ метрик mAP для различных архитектур

Результаты сравнительного анализа метрик точности детектирования ограничивающих рамок и масок сегментации объектов на аэрофотоснимках выявляют значимые различия производительности моделей, рисунок 8. Классические свёрточные сети (ResNet-50, ResNet-101, DenseNet-121) демонстрируют наименьшие значения метрик, в то время как современные архитектуры (ConvNeXt-T, Swin-T) обеспечивают существенный прирост точности. Архитектура Swin-S превосходит все рассмотренные модели по обоим показателям.

Таблица 3. Метрики точности детектирования масок сегментации для различных архитектур backbone

Backbone	mAP	mAP_s	mAP_m	mAP_l	Эпоха
ResNet-50	0.3000	0.1200	0.2800	0.4200	52
ResNet-101	0.3561	0.1900	0.3600	0.5100	67
DenseNet-121	0.3121	0.1600	0.3300	0.4800	73
EfficientNet-B3	0.3795	0.2000	0.3700	0.5200	65
ConvNeXt-T	0.3948	0.2300	0.4000	0.5500	75
Swin-T	0.4332	0.2400	0.4100	0.5600	78
Swin-S	0.5030	0.2800	0.4700	0.6300	82

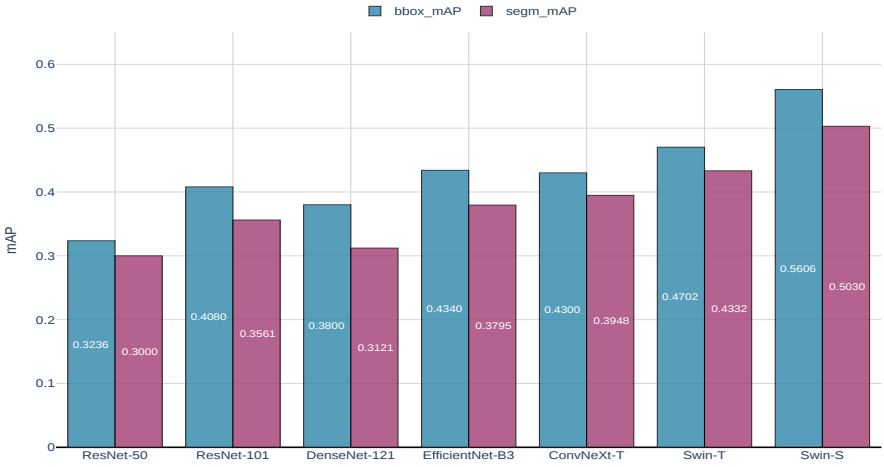


Рисунок 8. Сравнение метрик bbox_mAP и segm_mAP для различных архитектур backbone модели Mask R-CNN

Для всех архитектур характерно устойчивое превышение значения bbox_mAP над segm_mAP, что согласуется с теоретическими ожиданиями — задача точного пиксельного выделения объектов является более сложной по сравнению с детектированием ограничивающих рамок.

Величина разрыва между метриками варьируется в зависимости от архитектуры. Наименьшее относительное отклонение наблюдается у Swin-S (≈ 0.0576), что свидетельствует о её сбалансированной эффективности при решении обеих задач. Напротив, наибольший разрыв характерен для ResNet-50 (0.0236) и DenseNet-121 (0.0679), что указывает на их менее оптимальную адаптацию к задаче инстанс-сегментации.

Таким образом, полученные данные подтверждают перспективность

использования трансформерных и современных сверточных архитектур для обработки аэрофотоснимков, где требуется одновременное достижение высокой точности детектирования и сегментации объектов.

3.6. Сравнительная визуализация контуров масок сегментации объектов

На рисунках 9–15 и приведены результаты детектирования объектов на фрагменте изображения аэрофотоснимка с использованием модели Mask R-CNN, имеющей разные backbone. Практическая значимость

Рисунок 9. ResNet-50, погрешность контуров масок ≈15%

Рисунок 10. ResNet-101, погрешность контуров масок ≈10–11%

эффективного решения этой задачи описана в [13].

Анализ погрешности детектирования контуров масок сегментации показал прямую зависимость между архитектурой backbone и точностью сегментации. Классические архитектуры CNN демонстрируют наибольшую погрешность: ResNet-50 (≈15%), ResNet-101 (≈10–11%) и DenseNet-121 (≈12%). EfficientNet-B3 показывает результат (≈10–11%), сопоставимый с ResNet-101 и DenseNet-121, что, возможно, свидетельствует о достижении

Рисунок 11. DenseNet-121, погрешность контуров масок ≈12%

Рисунок 12. EfficientNet-B3, погрешность контуров масок ≈10–11%

Рисунок 13. ConvNeXt-T, погрешность контуров масок ≈7%

предела эффективности традиционных свёрточных архитектур. Более низкие значения погрешности получены для современных архитектур ConvNeXt-T (≈7%) и Swin-T (≈3—5%). Наименьшая погрешность определения контуров масок сегментации зафиксирована у трансформерной

Рисунок 14. Swin-T, погрешность контуров масок ≈3–5%

Рисунок 15. Swin-S (погрешность контуров масок ≈1–2%)

архитектуры Swin-S (≈1–2%), рисунок 15.

Полученные результаты свидетельствуют, что использование современных архитектур, в частности трансформеров, позволяет существенно повысить точность позиционирования границ объектов по сравнению с классическими CNN-подходами. Это особенно значимо для решения прикладных задач, требующих высокой точности сегментации, таких как картографирование или кадастровый учёт [**12**, **13**]. Повышение точности открывает возможности для автоматизации процессов обработки аэрофотоснимков с минимальным участием человека.

3.7. Оценка времени обучения и инференса моделей

Экспериментальные исследования моделей Mask R-CNN с различными типами архитектур backbone осуществлялись с использованием GPU A100 80GB. Время обучения моделей и время их инференса для изображений, средний размер которых составляет 1010 × 750 px, приведено в таблице 4.

Таблица 4. Время обучения и инференса моделей

Backbone	Время обучения	Время инференса
ResNet-50	12–18 мин.	0.05–0.1 сек.
DenseNet-121	20–30 мин.	0.08–0.15 сек.
ResNet-101	18–25 мин.	0.07–0.12 сек.
EfficientNet-B3	22–35 мин.	0.09–0.16 сек.
ConvNeXt-T	15–22 мин.	0.06–0.1 сек.
Swin-T	30–45 мин.	0.1–0.2 сек.
Swin-S	1.5–2 час.	0.2–0.4 сек.

Заключение

В работе проведён сравнительный анализ эффективности семи архитектур backbone в составе модели Mask R-CNN для задачи инстанс-сегментации объектов на аэрофотоснимках. Результаты эксперимента демонстрируют преимущество архитектур на основе механизма внимания (Swin-T и Swin-S) и усовершенствованной CNN (ConvNeXt-T) над классическими свёрточными сетями.

Установлено, что способность модели к захвату глобальных контекстных зависимостей является значимым фактором для достижения высокой точности сегментации в условиях аэрофотосъёмки. Архитектура Swin-S показала наивысшие значения метрик $bbox\_mAP$ ($= 0.5606$) и $segm\_mAP$ ($= 0.5030$), особенно для сегментации объектов малого размера. При этом достижение максимальной точности сопряжено со значительными вычислительными затратами, что ограничивает применение Swin-S задачами offline обработки. Для сценариев реального времени более целесообразно использование моделей Swin-T и ConvNeXt-T, обеспечивающих приемлемый компромисс между точностью и производительностью.

Исследование предоставляет эмпирические данные для выбора архитектуры backbone в зависимости от требований конкретного приложения. Результаты исследований в настоящее время используются в информационной системе картографирования ППК «Роскадастр». Перспективным направлением дальнейших работ является разработка методов оптимизации соотношения точности и производительности, включая создание гибридных архитектур.

Список использованных источников

[1] He K., Gkioxari G., P. Dollár, Girshick R. B. *Mask R-CNN // 2017 IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy, 22–29 October 2017).– IEEE.– 2017.– ISBN 9781538610336.– Pp. 2980–2988. doi arXiv 1703.06870 ↑196

[2] Ren S., He K., Girshick R., Sun J. *Faster R-CNN: towards real-time object detection with region proposal networks // Advances in Neural Information Processing Systems 28*.– V. 1, NIPS 2015 (Montreal, Canada, 7–12 December 2015), Cambridge: MIT Press.– 2015.– ISBN 9781510825024.– Pp. 91–99. doi URL ↑196

[3] He K., Zhang X., Ren S., Sun J. *Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA, 27–30 June 2016).– IEEE.– 2016.– ISBN 9781467388528.– Pp. 770–778. doi arXiv 1512.03385 ↑196, 198

[4] Huang G., Liu Z., van der Maaten L., Weinberger K. Q. *Densely connected convolutional networks // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, USA, 21–26 July 2017).– 2017.– ISBN 9781538604588.– Pp. 2261–2269. doi arXiv 1608.06993 ↑196, 198

[5] Tan M., Le Q. V. *EfficientNet: rethinking model scaling for convolutional neural networks*, International Conference on Machine Learning (Long Beach, California, USA, 9–15 June 2019), Proceedings of Machine Learning Research.– vol. **97**.– 2019.– ISBN 9781510886988.– Pp. 6105–6114. doi URL arXiv 1905.11946 ↑196, 198

[6] Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S. *A ConvNet for the 2020s // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 18–24 June 2022).– IEEE.– 2022.– ISBN 9781665469470.– Pp. 11976–11986. arXiv 2201.03545 doi ↑196, 198

[7] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. *Swin Transformer: hierarchical vision transformer using shifted windows // 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC, Canada, 10–17 October 2021).– 2021.– ISBN 978-1-6654-2812-5.– Pp. 9992–10002. doi arXiv 2103.14030 ↑196, 198

[8] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. *An image is worth 16x16 words: transformers for image recognition at scale // International Conference on Learning Representations (ICLR)* (Vienna, Austria, 4 May 2021).– 2021.– ISBN 9798331321949.– 21 pp. URL arXiv 2010.11929 ↑197

[9] Vasu P. K. A., Gabriel J., Zhu J., Tuzel O., Ranjan A. *MobileOne: an improved one millisecond mobile backbone // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC, Canada, 18–22 June 2023).– IEEE.– 2023.– ISBN 9798350301304.– Pp. 7907–7917. doi arXiv 2206.04040 ↑197

[10] Maaz M., Shaker A., Cholakkal H., Khan S., Zamir S. W., Anwer R. M., Khan F. S. *EdgeNeXt: efficiently amalgamated CNN-transformer architecture for mobile vision applications*, Computer Vision – ECCV 2022 Workshops (ECCV 2022) (Tel Aviv, Israel, 23–27 October 2022), Lecture Notes in Computer Science.– vol. **13807**, Cham: Springer.– 2022.– ISBN 978-3-031-25082-8.– Pp. 3–20. doi arXiv 2206.10589 ↑197

[11] Dai Z., Liu H., Le Q. V., Tan M. *CoAtNet: marrying convolution and attention for all data sizes // NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* (6–14 December 2021), Red Hook: Curran Associates Inc..– 2021.– ISBN 978-1-7138-4539-3.– Pp. 3965–3977.– id. 303. doi arXiv 2106.04803 ↑197

[12] Винокуров И. В. *Использование модели Mask R-CNN для сегментации объектов недвижимости на аэрофотоснимках* // Программные системы: теория и приложения.– 2025.– Т. **16**.– № 1(64).– С. 3–44 (Англ., Рус.). doi URL ↑197, 200, 212

[13] Винокуров И. В. *Повышение точности сегментирования объектов с использованием генеративно-состязательной сети* // Программные системы: теория и приложения.– 2025.– Т. **16**.– № 2(65).– С. 111–152 (Англ., Рус.). doi URL ↑197, 210, 212

[14] Xie E., Wang W., Yu Z., Anandkumar A., Alvarez J. M., Luo P. *SegFormer: simple and efficient design for semantic segmentation with transformers // NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* (6–14 December 2021), Red Hook: Curran Associates Inc..– 2021.– ISBN 978-1-7138-4539-3.– Pp. 12077–12090. doi arXiv 2105.15203 ↑198

[15] Cheng B., Misra I., Schwing A. G., Kirillov A., Girdhar R. *Masked-attention mask transformer for universal image segmentation // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 18–24 June 2022).– IEEE.– 2022.– ISBN 9781665469470.– Pp. 1290-1299. doi arXiv 2112.01527 ↑199

[16] Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. *End-to-end object detection with transformers*, Computer Vision - ECCV 2020 (ECCV 2020), Lecture Notes in Computer Science.– vol. **12346**, Cham: Springer.– 2020.– ISBN 978-3-030-58452-8.– Pp. 213–229. doi arXiv 2005.12872 ↑199

[17] Tan M., Pang R., Le Q. V. *EfficientDet: scalable and efficient object detection // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 14–19 June 2020).– IEEE.– 2020.– ISBN 9781728171692.– Pp. 10778–10787. doi arXiv 1911.09070 ↑199

[18] Peng J., Liu Y., Tang S., Hao Y., Chu L., Chen G., Wu Z., Chen Z., Lai B. *PP-LiteSeg: a superior real-time semantic segmentation model.*– 2022.– 8 pp. doi arXiv 2204.02681 ↑199

[19] Hafiz A. M., Bhat G. M. *A survey on instance segmentation: state of the art* // International Journal of Multimedia Information Retrieval.– 2020.– Vol. **9**.– Pp. 171–189. doi ↑199

[20] Canziani A., Paszke A., Culurciello E. *An Analysis of Deep Neural Network models for practical applications.*– 2016. doi arXiv 1605.07678 ↑199

Поступила в редакцию 22.09.2025;
одобрена после рецензирования 27.09.2025;
принята к публикации 12.10.2025;
опубликована онлайн 19.10.2025.

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Информация об авторах:

Игорь Викторович Винокуров

Кандидат технических наук (PhD), ассоциированный профессор в Финансовом Университете при Правительстве Российской Федерации. Область научных интересов: информационные системы, информационные технологии, технологии обработки данных

0000-0001-8697-1032

e-mail: *igvvinokurov@fa.ru*

Дарья Александровна Фролова

Студент третьего курса бакалавриата Финансового Университета при Правительстве Российской Федерации. Область научных интересов: информационные технологии, автоматизация процессов, анализ данных

e-mail: *dari15frolowa@gmail.com*

Андрей Иванович Ильин

Студент третьего курса бакалавриата Финансового Университета при Правительстве Российской Федерации. Область научных интересов: информационные технологии, программирование, анализ данных

e-mail: *andrey08937@yandex.ru*

Иван Романович Кузнецов

Студент третьего курса бакалавриата Финансового Университета при Правительстве Российской Федерации. Область научных интересов: информационные технологии и технологии обработки данных, программирование

e-mail: *ivan.kuznetsov0709@mail.ru*

Вклад авторов: *И.В. Винокуров* – 70% (разработка методики проведения экспериментов, формирование конфигурационных файлов, реализация обучения и исследования моделей, интеграция результатов в информационные системы ППК «Роскадастр»); *Д.А. Фролова* – 10% (аннотирование изображений, формирование набора данных); *А.И. Ильин* – 10% (оптимизация метрик точности моделей); *И.Р. Кузнецов* – 10% (визуализация инференса моделей).

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*