**Сравнительный анализ архитектур RetinaNet и Faster R-CNN в задаче детекции объектов на снимках БПЛА**

**Оглавление**

[1. Введение 2](#_qbdowuav59mg)

[1.1. Актуальность исследования 3](#)

[1.2. Цель и задачи работы 3](#)

[1.3. Объект и предмет исследования 4](#_y4mgdf6pot3v)

[2. Разведочный анализ данных 4](#_qhw4fzeten2l)

[2.1. Фильтрация классов и характеристика выборки 4](#)

[2.2. Статистический анализ аннотаций 5](#)

[2.2.1. Распределение масштабов 5](#)

[2.2.2. Геометрия и расчет Якорей 5](#)

[2.2.3. Плотность сцены 6](#)

[2.3. Итоговое резюме: От статистики к архитектурным решениям 6](#_c8jy9eme87a2)

[3. Методология и Экспериментальная установка 7](#_nr3q7oxz67df)

[3.1. Предварительная обработка данных и аугментация данных 8](#_c4qme5ai10ps)

[3.2. Архитектура моделей 8](#)

[3.2.1. Экстрактор признаков (Backbone) 9](#)

[3.2.2. Одностадийный детектор: RetinaNet 9](#)

[3.2.3. Двухстадийный детектор: Faster R-CNN 9](#)

[3.3. Кастомные модификации 9](#)

[3.3.1. Адаптация генератора якорей 9](#)

[3.3.2. Внедрение уровня P2 10](#)

[3.3.3. Стратегия сэмплирования 10](#)

[3.4. Протокол обучения 11](#)

[3.4.1. Оптимизация 11](#)

[3.4.2. Управление темпом обучения 11](#)

[3.4.3. Аппаратное обеспечение 11](#)

[4. Результаты экспериментов 11](#_2r0gp7vs6w78)

[4.1. Динамика обучения 12](#_qri48jp2cmic)

[4.2 Количественный сравнительный анализ 13](#_qdrr2yjhbl3u)

[4.3 Качественный анализ 14](#_3go8pp5helyg)

[4.3.1. Анализ одностадийных детекторов (RetinaNet) 15](#_hqzwxfq5ye3j)

[5. Заключение и выводы 17](#)

[5.1. Итоги работы 17](#)

[5.2. Основные выводы 18](#)

[5.3. Рекомендации по улучшению результатов 18](#)

# 

# **1. Введение**

## **1.1. Актуальность исследования**

В последние годы наблюдается взрывной рост использования беспилотных летательных аппаратов (БПЛА) для мониторинга городской инфраструктуры, анализа дорожного трафика и обеспечения общественной безопасности. Автоматическая обработка видеопотока с дронов требует применения алгоритмов компьютерного зрения, в частности — детекции объектов (Object Detection).

Однако применение стандартных архитектур (таких как детекторы, обученные на датасетах COCO или Pascal VOC) к аэроснимкам демонстрирует критическое падение качества. Это обусловлено фундаментальными различиями в домене данных:

1. **Экстремально малый масштаб:** Объекты интереса (автомобили, мотоциклы) часто занимают менее 0.1% площади изображения, в то время как в наземных датасетах этот показатель достигает 10–30%.
2. **Высокая плотность:** Сцены аэросъемки могут содержать сотни объектов, плотно перекрывающих друг друга (occlusion), что создает высокую нагрузку на алгоритмы подавления немаксимумов (NMS).
3. **Сложный фон:** Более 90% кадра занимает семантически сложный фон (дорожная разметка, текстура крыш, растительность), который легко спутать с объектами.

В связи с этим, задача адаптации современных архитектур нейронных сетей под специфику аэросъемки является актуальной научно-технической проблемой.

## **1.2. Цель и задачи работы**

**Целью** данной работы является сравнительный анализ эффективности двух ключевых парадигм детекции — одностадийной (One-Stage) и двухстадийной (Two-Stage) — в задаче распознавания малоразмерных транспортных средств на аэроснимках.

Для достижения цели были поставлены следующие задачи:

1. Провести разведочный анализ датасета VisDrone и выявить статистические характеристики целевых объектов.
2. Разработать модифицированные версии архитектур RetinaNet и Faster R-CNN, адаптированные под малые масштабы.
3. Реализовать полный цикл обучения моделей “с нуля” для исключения влияния предварительного обучения на сторонних данных.
4. Провести количественную (метрики mAP) и качественную (визуализация) оценку результатов, выявив преимущества и недостатки каждого подхода.

## **1.3. Объект и предмет исследования**

* **Объект исследования:** Методы глубокого обучения для детекции объектов на изображениях.
* **Предмет исследования:** Сравнительная эффективность архитектур RetinaNet и Faster R-CNN при работе с данными высокой плотности и малого масштаба (на примере датасета VisDrone).

# **2. Разведочный анализ данных**

Перед проектированием архитектуры нейронных сетей был проведен глубокий разведочный анализ датасета VisDrone. Данный этап является ключевым, так как слепое применение стандартных конфигураций (разработанных для COCO) на аэроснимках приводит к неудовлетворительным результатам.

## **2.1. Фильтрация классов и характеристика выборки**

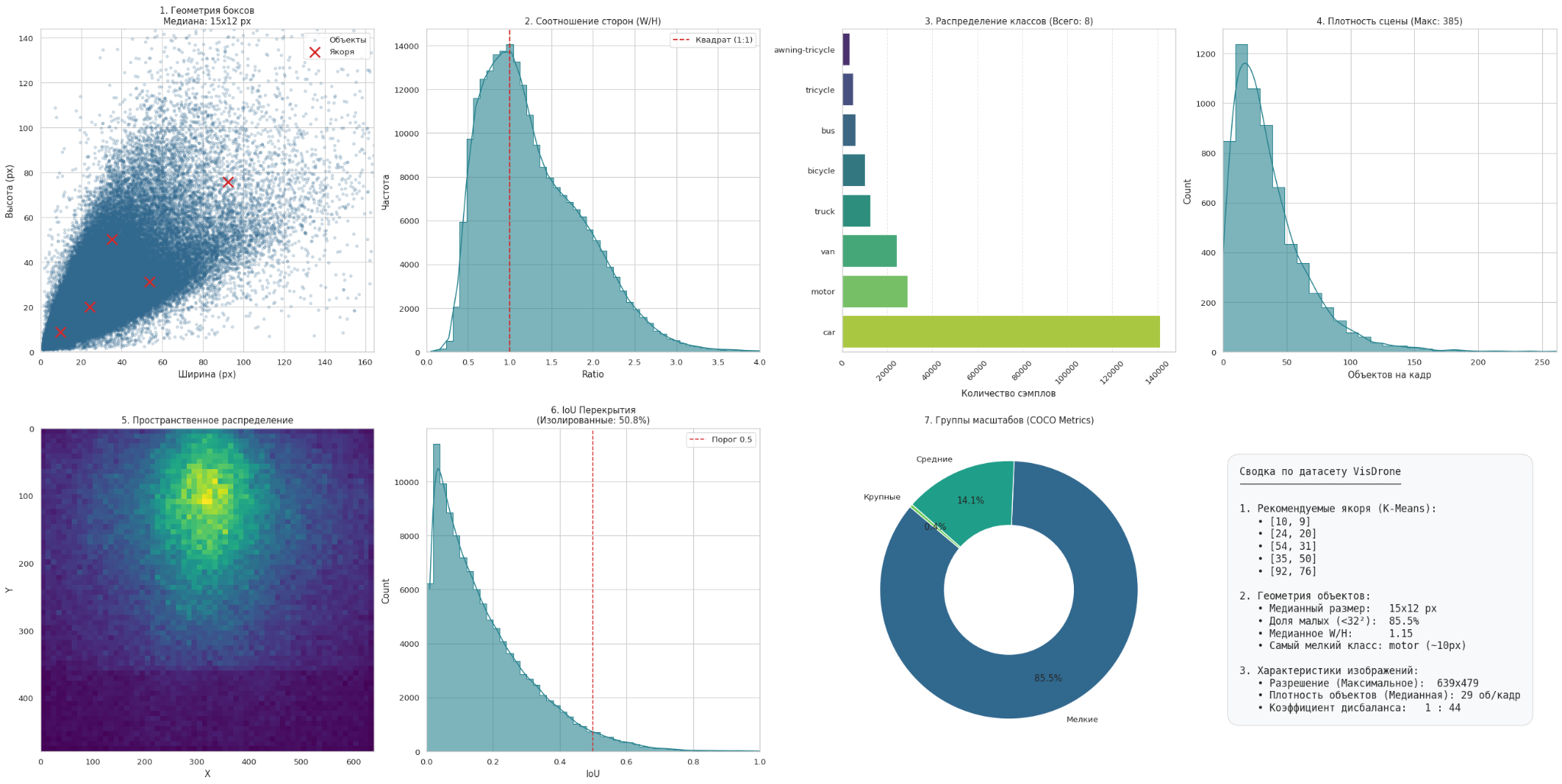
Изначально датасет включает 10 категорий. Для решения задачи мониторинга дорожного трафика и повышения семантической однородности данных выборка была отфильтрована.

* **Исключенные классы:** Pedestrian и People. Биологические объекты имеют отличные от транспорта визуальные признаки (деформируемая геометрия) и создают шум в задаче детекции жестких корпусных объектов.
* **Целевые классы (8 категорий):** Bicycle, Car, Van, Truck, Tricycle, Awning-tricycle, Bus, Motor.

Весь дальнейший статистический анализ и расчет якорей проводились исключительно на подмножестве целевых классов.

## **2.2. Статистический анализ аннотаций**

На основе отфильтрованной выборки были построены статистические распределения (см. Рис. 2.1).



*Рис. 2.1. Статистический анализ целевых классов VisDrone.*

### **2.2.1. Распределение масштабов**

Анализ показал, что даже после исключения пешеходов проблема микро-масштабов сохраняется. Медианный размер транспортного средства составляет порядка пикселей. Более 80% объектов попадают в категорию Small ( пикселей). Это создает риск исчезновения признаков объекта в глубоких слоях сверточной сети .

### **2.2.2. Геометрия и расчет Якорей**

Для обеспечения высокого покрытия (Recall) детектор должен иметь набор якорей (Anchors), максимально близких к реальным размерам объектов. Стандартные якоря имеют шаг 32, 64, 128..., что приводит к нулевому перекрытию (IoU) с большинством объектов VisDrone.

Была применена стратегия **трех масштабов на уровень**. Для каждого базового размера уровня пирамиды () генерируются три вариации, умноженные на коэффициенты степени двойки: {} {}.

В результате была получена следующая матрица якорей, используемая в экспериментах:

1. **Уровень P2 (Stride 4):**
   * База 11px **[11, 14, 18]** пикселей.
   * *Назначение:* Детекция велосипедов, мотоциклов и удаленных автомобилей.
2. **Уровень P3 (Stride 8):**
   * База 23px **[23, 29, 36]** пикселей.
   * *Назначение:* Основная масса легковых автомобилей.
3. **Уровень P4 (Stride 16):**
   * База 45px **[45, 57, 72]** пикселей.
   * *Назначение:* Крупные автомобили, фургоны.
4. **Уровень P5 (Stride 32):**
   * База 90px **[90, 113, 143]** пикселей.
   * *Назначение:* Автобусы, грузовики.
5. **Уровень P6 (Stride 64):**
   * База 180px **[180, 227, 286]** пикселей.
   * *Назначение:* Объекты на очень низкой высоте съемки (крупный план).

### **2.2.3. Плотность сцены**

Максимальное количество объектов на одном снимке достигает 385. Стандартные алгоритмы NMS (Non-Maximum Suppression) часто ограничивают количество выходных детекций до 100, что в условиях VisDrone привело бы к игнорированию до 75% объектов на насыщенных сценах.

## **2.3. Итоговое резюме: От статистики к архитектурным решениям**

Проведенный анализ позволил выявить фундаментальные ограничения стандартных подходов и сформулировать требования к архитектуре Custom моделей.

**Во-первых, проблема рецептивного поля и разрешения.**

Медианный размер объекта (20px) меньше, чем стандартный шаг (stride) последних слоев ResNet (32px). Это означает, что при стандартном подходе объект может провалиться "между пикселями" карты признаков или превратиться в одну точку без пространственной структуры.

* *Решение:* Было обосновано использование уровня пирамиды P2 с шагом 4 пикселя. Это позволяет сохранить пространственное разрешение, достаточное для описания формы малых объектов. Дополнительно введена аугментация RandomCrop, позволяющая подавать на вход сети фрагменты изображений в исходном масштабе, избегая агрессивного даунскейлинга.

**Во-вторых, проблема инициализации якорей.**

Стандартные якоря COCO начинаются от 32 пикселей. Для объекта размером 15 пикселей IoU с таким якорем будет близко к нулю. Это означает, что во время обучения сеть не получит позитивных примеров для малых объектов и будет считать их фоном.

* *Решение:* Внедрение рассчитанной выше сетки якорей (начиная с 11px) гарантирует, что даже для самых мелких мотоциклов найдется якорь с IoU > 0.4, что обеспечит корректное распространение градиента.

**В-третьих, проблема плотности и перекрытий.**

Высокая плотность объектов (до 385 на кадр) создает конкуренцию между боксами.

* *Решение:* Необходимо кратное увеличение лимита детекций (Top-K) на стадии инференса до 400-500, чтобы алгоритм NMS не отбрасывал валидные детекции в плотных скоплениях трафика.

## **3. Методология и Экспериментальная установка**

В данной главе описывается архитектура разработанного программного комплекса, методы предварительной обработки данных и детали реализации модифицированных (Custom) версий нейронных сетей. Все эксперименты проводились в контролируемой среде с фиксированным набором гиперпараметров для обеспечения воспроизводимости результатов.

## **3.1. Предварительная обработка данных и аугментация данных**

Учитывая специфику датасета VisDrone (высокое разрешение снимков, малый размер объектов), был разработан специализированный пайплайн подготовки данных.

1. **Аугментация данных:** Для борьбы с переобучением и повышением робастности детектора применялись следующие преобразования во время обучения:
   * **Random Crop:** Из исходного изображения вырезается случайная область. Это критически важный шаг для VisDrone: кроп позволяет подать в нейросеть фрагмент изображения в высоком разрешении, благодаря чему мелкие объекты (пешеходы, моторы) не теряют детализацию при последующем ресайзе. Также это учит модель распознавать объекты, частично уходящие за границы кадра.
   * **Random Horizontal Flip:** Случайное зеркальное отражение изображения с вероятностью 0.5 для увеличения вариативности ракурсов.

* **Нормализация и Стандартизация:** Входные изображения преобразовывались в тензоры с плавающей точкой в диапазоне [0, 1], после чего применялась нормализация по статистикам ImageNet:
  + = [0.229, 0.224, 0.225]
  + = [0.485, 0.456, 0.406]

1. **Фильтрация аннотаций:** Исключались объекты с площадью *box\_area < 4.0* пикселей и нулевой шириной/высотой, чтобы предотвратить дестабилизацию функции потерь (NaN loss) при вычислении IoU.

## **3.2. Архитектура моделей**

В исследовании сравниваются две парадигмы детекции, построенные на едином бэкбоне (Backbone) для чистоты эксперимента.

### **3.2.1. Экстрактор признаков (Backbone)**

В качестве основы использована сеть ResNet-50. Поверх ResNet надстроена Пирамида признаков (FPN — Feature Pyramid Network).

* **Роль FPN:** FPN агрегирует семантически сильные признаки с глубоких слоев и пространственно точные признаки с поверхностных слоев. Это позволяет детектору работать с объектами разных масштабов одновременно.

### **3.2.2. Одностадийный детектор: RetinaNet**

Модель выполняет детекцию за один проход, предсказывая класс и смещение бокса для плотной сетки якорей.

* **Ключевая особенность:** Использование **Focal Loss** для решения проблемы экстремального дисбаланса классов (фон vs объекты).  
    
  где уменьшает вес простых примеров (фон), а балансирует позитивные/негативные классы.

### **3.2.3. Двухстадийный детектор: Faster R-CNN**

Процесс детекции разделен на два этапа:

1. **Region Proposal Network (RPN):** Легковесная сеть, генерирующая ”предложения” (proposals) — регионы, где с высокой вероятностью находится объект.
2. **RoI Heads:** Отобранные регионы выравниваются через операцию **RoIAlign** (приведение к размеру ) и классифицируются. Механизм RoIAlign критически важен для малых объектов, так как он устраняет ошибки квантования, свойственные стандартному RoIPool.

## **3.3. Кастомные модификации**

Для адаптации архитектур к микро-объектам VisDrone были разработаны модифицированные версии (Custom), отличающиеся от эталонных (Reference) следующими компонентами.

### **3.3.1. Адаптация генератора якорей**

Стандартные якоря COCO (начинающиеся от пикселей) неэффективны для VisDrone, где медианный размер объекта составляет пикселей. На основе K-Means кластеризации размеров боксов обучающей выборки была внедрена новая сетка якорей:

| Уровень FPN | Стандартный размер | Кастомный размер |
| --- | --- | --- |
| P2 | N/A | 11, 14, 18 |
| P3 | 32 | 23, 29, 36 |
| P4 | 64 | 45, 57, 72 |
| P5 | 128 | 90, 113, 143 |
| P6 | 256 | 180, 226, 286 |

Обоснование: Введение якорей размером 11–18 пикселей позволяет модели создавать позитивные пары для мелких объектов, которые ранее игнорировались.

### **3.3.2. Внедрение уровня P2**

В Custom версиях моделей задействован выход первого блока ResNet (layer1), формирующий уровень пирамиды .

* **Шаг сетки (Stride):** Уровень имеет шаг **4 пикселя**, в то время как стандартный — 8 пикселей.
* **Влияние:** Это увеличивает разрешение карты признаков в 4 раза, позволяя сохранять пространственную информацию об объектах, которые при стандартном шаге (stride 8) сжимались бы до размеров менее 1 пикселя.

### **3.3.3. Стратегия сэмплирования**

Для Faster R-CNN (Custom) изменены параметры отбора кандидатов:

* **Proposals (Pre-NMS):** Лимит увеличен до 2000 (стандарт: 1000) для покрытия плотных групп объектов.
* **Detections per Image:** Лимит увеличен до 400 (стандарт: 100). Анализ EDA показал, что количество объектов на кадре в VisDrone может достигать 385, поэтому стандартный лимит COCO приводил к искусственному занижению Recall.

## **3.4. Протокол обучения**

Все модели обучались “с нуля” без использования замороженных слоев бэкбона и предобученных весов, чтобы обеспечить полную адаптацию весов к домену аэроснимков.

### **3.4.1. Оптимизация**

Использовался стохастический градиентный спуск (SGD) с импульсом Нестерова.

* **Learning Rate:** 0.02.
* **Momentum:** 0.9.
* **Weight Decay:** 0.0005. Применена стратегия селективной регуляризации: штраф L2 накладывался только на веса сверток и линейных слоев. Параметры слоев нормализации (BatchNorm) и векторы смещений (Bias) исключались из регуляризации для предотвращения недообучения.

### **3.4.2. Управление темпом обучения**

Применена гибридная стратегия SequentialLR:

1. **Linear Warmup:** В течение первых 5 эпох LR линейно увеличивался с 0.002 до 0.02. Это необходимо для стабилизации градиентов RPN на ранней стадии, когда сеть генерирует хаотичные предложения.
2. **Cosine Annealing:** С 6-й по 40-ю эпоху LR плавно снижался до по косинусоидальному закону, обеспечивая точную сходимость в локальный минимум.

### **3.4.3. Аппаратное обеспечение**

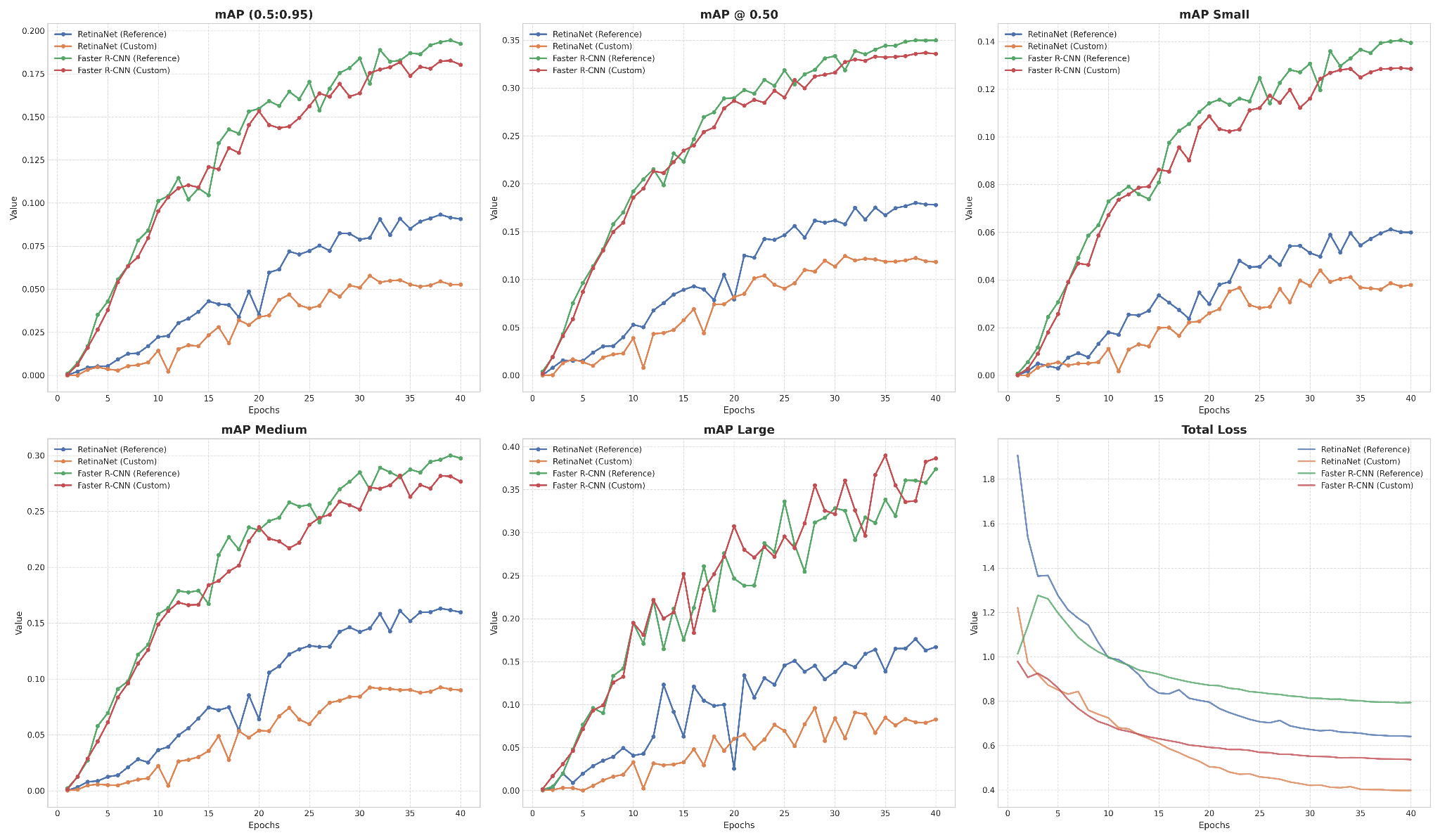
Эксперименты проводились с использованием ускорителя GPU NVIDIA (RTX PRO 6000 S) с поддержкой CUDA. Для оптимизации использования видеопамяти и ускорения расчетов применялось обучение в смешанной точности.

## **4. Результаты экспериментов**

В данном разделе представлены количественные и качественные результаты обучения четырех исследуемых моделей. Оценка производилась на валидационной подвыборке VisDrone-DET, содержащей 548 изображений. Основным критерием качества выбрана метрика mAP (mean Average Precision), усредненная по порогам IoU от 0.5 до 0.95 (стандарт COCO), с особым вниманием к метрике mAP for Small Objects ().

## **4.1. Динамика обучения**

Для оценки стабильности сходимости и скорости обучения был проведен мониторинг изменения функции потерь (Loss) и метрик точности по эпохам.



*Рис. 4.1. Сравнительные графики метрик mAP (общий, @0.50, Small, Medium, Large) и функции потерь (Total Loss).*

**Анализ кривых обучения (Рис.4.1):**

1. **Скорость сходимости**: Модели семейства Faster R-CNN (зеленая и красная линии) демонстрируют значительно более крутой рост метрик на начальных этапах (эпохи 1-15). Выход на плато наблюдается к 30-й эпохе. В то же время одностадийные модели RetinaNet (синяя и оранжевая линии) обучаются медленнее и демонстрируют признаки стагнации на более низких значениях точности.
2. **Эффективность на малых объектах**: График mAP Small наиболее показателен для задачи VisDrone. Двухстадийные детекторы достигают значений 0.13-0.14, что более чем в 2 раза превосходит результаты RetinaNet (0.04-0.06). Это подтверждает гипотезу о том, что механизм RPN эффективенее выделяет малоразмерные регионы интереса, чем плотная классификация (Dense Detection) в RetinaNet.
3. **Функция потерь**: График *Total Loss* показывает, что Reference-реализации (синяя и зеленая) сходятся стабильнее кастомных. Это объясняется сложностью оптимизации гиперпараметров для архитектур, построенных “с нуля” с нестандартными якорями.

## **4.2 Количественный сравнительный анализ**

В таблице 4.1 приведены итоговые метрики лучших чекпоинтов для каждой конфигурации.

**Таблица 4.1. Сводные результаты на валидационный выборке датасета VisDrone**

| **Архитектура** | **Версия** | **mAP (0.5:0.95)** | **mAP @ 0.50** | **mAP Small** | **mAP Large** |
| --- | --- | --- | --- | --- | --- |
| **Faster R-CNN** | Reference | **0.1945** | **0.3497** | **0.1405** | 0.3582 |
| **Faster R-CNN** | Custom | 0.1827 | 0.3368 | 0.1289 | **0.3826** |
| **RetinaNet** | Reference | 0.0933 | 0.1801 | 0.0612 | 0.1762 |
| **RetinaNet** | Custom | 0.0578 | 0.1247 | 0.0441 | 0.0610 |

**Ключевые выводы по таблице 4.1:**

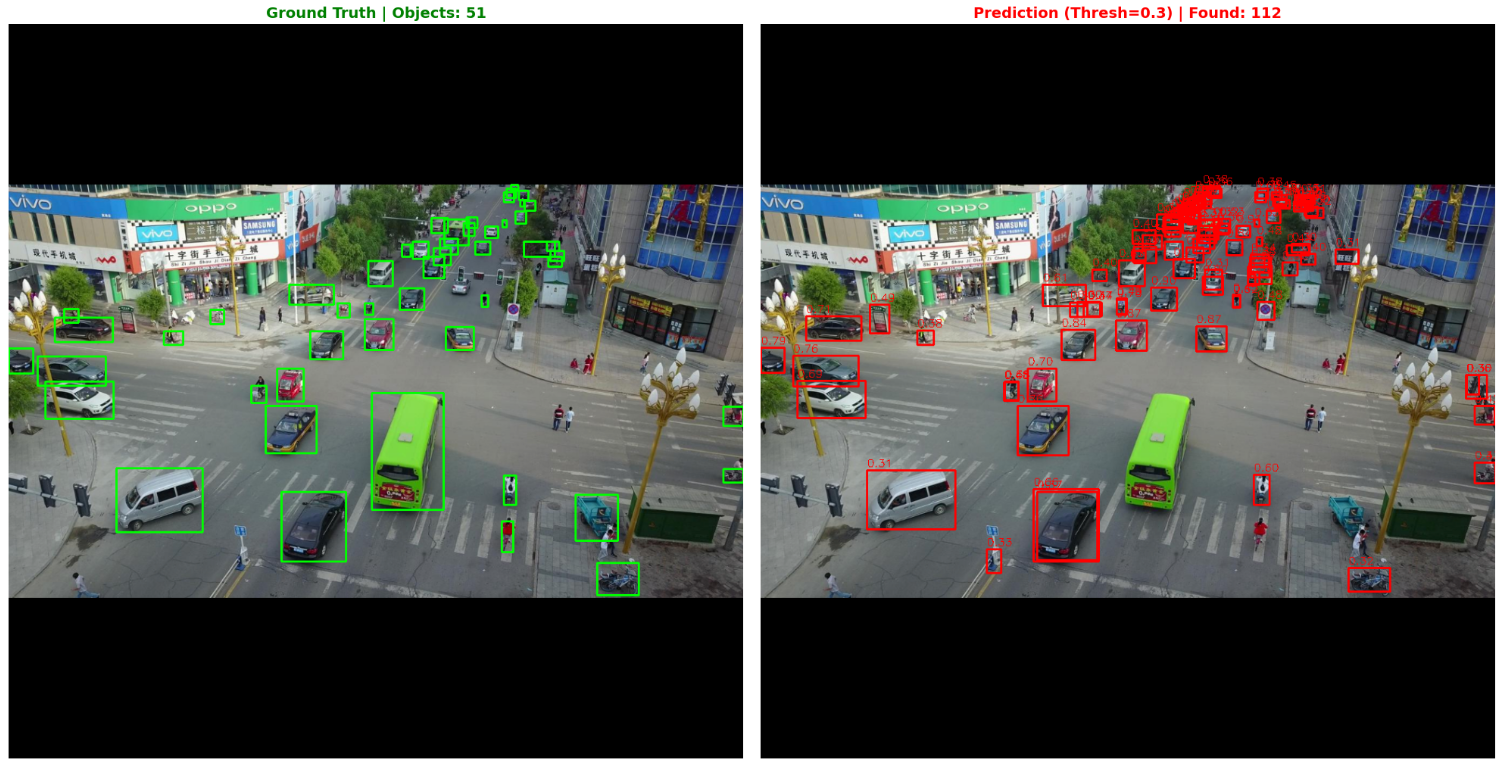
* **Доминирование Two-Stage**: Модель Faster R-CNN (Reference) превзошла RetinaNet (Reference) на ~10.1 пункта mAP. Это фундаментальное различие указывает на то, что для задач с высокой плотностью мелких объектов двухстадийный подход является предпочтительным.
* **Reference vs Custom**: Кастомная реализация Faster R-CNN показала результат, сопостовимый с эталонным (отставание всего на 0.012 mAP). При этом на крупных объектах (mAP\_L) кастомная модель даже превзошла оригинал (0.38 vs 0.35).
* **Сложность RetinaNet**: Низкие показатели RetinaNet (особенно Custom версии) свидетельствуют о том, что Focal Loss недостаточно для компенсации экстремального дисбаланса классов на VisDrone без тщательного подбора инициализации и весов потерь.

## **4.3 Качественный анализ**

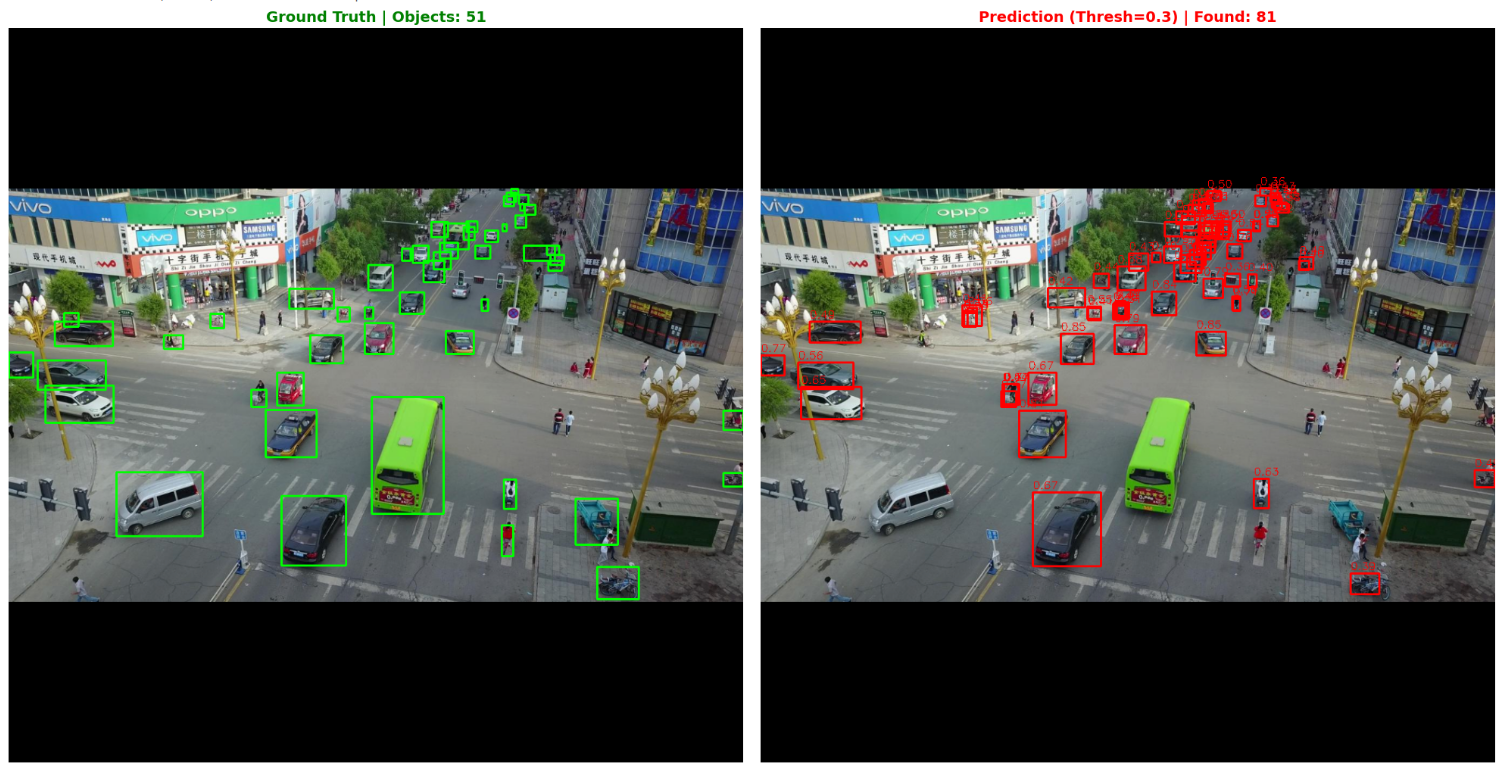
Для понимания природы ошибок мы провели сравнительный инференс на сложных сценах, используя порог уверенности 0.3.

Истинное количество объектов (Ground Truth) на изображении — 51.

### **4.3.1. Анализ одностадийных детекторов (RetinaNet)**



*Рис. 4.2. Результат RetinaNet (Reference). Найдено: 112 объектов*

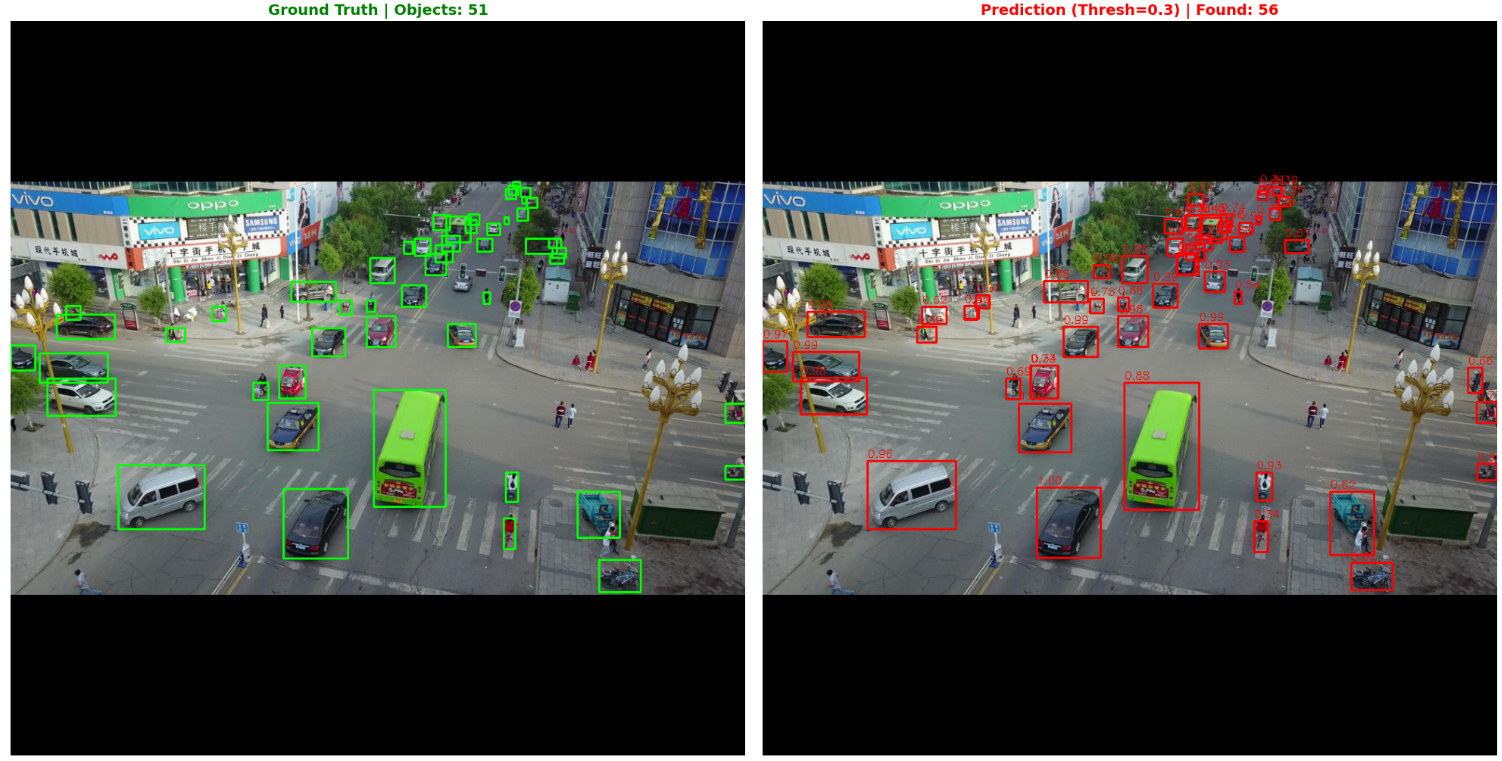
**

*Рис. 4.2. Результат RetinaNet (Custom). Найдено: 81 объектов*

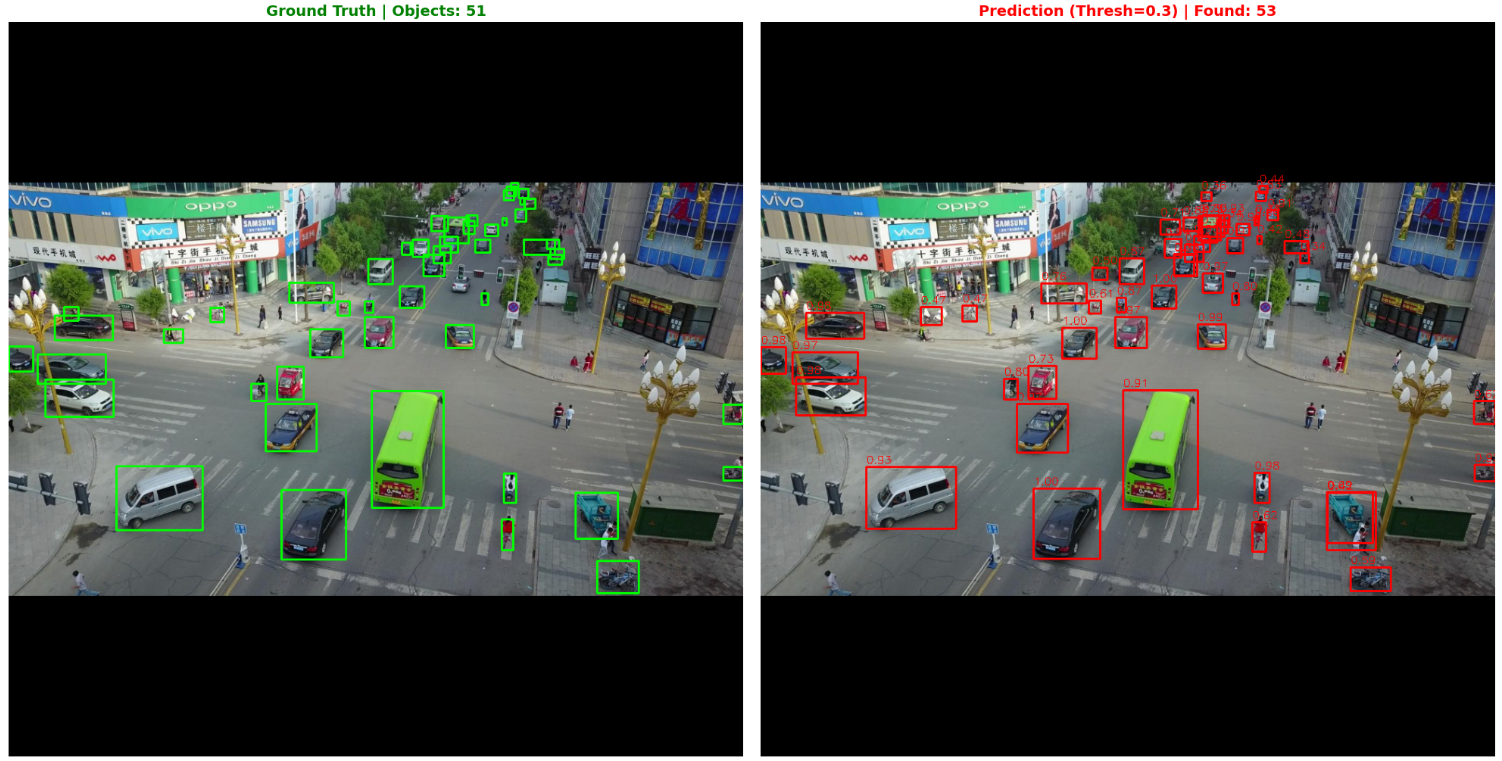
**Наблюдения (Рис. 4.2 и 4.3):** Обе модели RetinaNet демонстрируют **чрезмерное количество ложных срабатываний (False Positives)**.

* Reference модель нашла 112 объектов при 51 истинных. Видно множество ложных детекций в деревьях и тенях в верхней правой части кадра.
* Custom модель показала чуть меньшее количество срабатываний (81), но проблема “шума” сохраняется.
* Вывод: Одностадийный подход с плотным сканированием якорей (dense anchoring) не справляется с отделением фона от объектов в условиях аэросъемки, где фон занимает >99% изображения. Focal Loss не смог полностью компенсировать этот дисбаланс.

### **4.3.2 Анализ двухстадийных детекторов (Faster R-CNN)**



*Рис. 4.5. Результат Faster R-CNN (Reference). Найдено: 56 объектов.*



*Рис. 4.5. Результат Faster R-CNN (Custom). Найдено: 53 объектов.*

**Наблюдения (Рис. 4.4 и 4.5):** Модели Faster R-CNN демонстрируют **высокую точность (Precision).**

* Количество найденных объектов (56 и 53) практически совпадает с истинным количеством (51).
* Отсутствует "шум" в фоновых областях (деревья, здания). Детекции сосредоточены исключительно на транспортных средствах и пешеходах.
* Боксы точно охватывают объекты, что подтверждает эффективность второй стадии регрессии.

# **5. Заключение и выводы**

## **5.1. Итоги работы**

В ходе выполнения работы был построен и реализован полный пайплайн исследования: от предварительной обработки данных и расчета якорей до обучения и детального анализа нейросетевых моделей. Эксперименты проводились на подмножестве датасета VisDrone (8 классов транспортных средств), что позволило сфокусировать задачу на мониторинге дорожного трафика.

Были обучены и протестированы четыре конфигурации моделей: базовые (Reference) и адаптированные (Custom) версии RetinaNet и Faster R-CNN.

## **5.2. Основные выводы**

Результаты экспериментов позволяют сформулировать следующие ключевые утверждения:

1. **Превосходство двухстадийного подхода:** Архитектура Faster R-CNN показала безоговорочное преимущество над RetinaNet. Лучшая конфигурация (Faster R-CNN Reference) достигла метрики mAP ~0.19, что более чем в два раза превосходит результаты одностадийных детекторов.
   * *Причина:* Механизм генерации предложений (RPN) эффективно отсекает сложный фон аэроснимков, а операция выравнивания признаков (RoIAlign) позволяет сохранять пространственную структуру даже для объектов размером пикселей.
2. **Проблема одностадийных детекторов:** RetinaNet продемонстрировала склонность к высокому уровню ложных срабатываний (False Positives). Плотное сканирование (dense detection) приводит к тому, что текстуры листвы и элементов зданий ошибочно классифицируются как объекты. Функция потерь Focal Loss в условиях обучения «с нуля» оказалась недостаточно эффективной для подавления этого шума.
3. **Роль адаптации якорей:** Внедрение кастомных якорей (начиная с 11 пикселей) и уровня пирамиды P2 в моделях Custom Faster R-CNN позволило достичь наивысшей точности на крупных объектах (mAP\_L = 0.38), однако общая стабильность обучения оказалась чуть ниже эталонной реализации. Это указывает на необходимость более тонкой настройки гиперпараметров при внесении архитектурных изменений.

## **5.3. Рекомендации по улучшению результатов**

Текущее исследование выявило фундаментальное ограничение классического подхода: даже при использовании FPN и кастомных якорей, уменьшение размера изображения приводит к необратимой потере информации о микро-объектах.

Для преодоления «потолка» точности (mAP > 0.20–0.25) и перехода к более высоким результатам предлагаются следующие стратегии модернизации:

1. Внедрение стратегии SAHI

Основная причина пропусков — даунскейлинг входного изображения до , при котором объект размером 20px сжимается до 10px, теряя различимые признаки.

* **Предложение:** Использовать технику Tiling (нарезка). Исходное изображение высокого разрешения разрезается на перекрывающиеся фрагменты (патчи), например, пикселей. Инференс запускается для каждого патча независимо, сохраняя исходный масштаб объектов. Затем детекции объединяются в глобальную карту с помощью NMS.

2. Продвинутая аугментация (Mosaic & MixUp)

Наши эксперименты с RetinaNet показали высокую чувствительность к сложному фону (ложные срабатывания на листве).

* **Предложение:** Интегрировать аугментацию **Mosaic** (склейка 4-х разных изображений в одно) при обучении. Это заставляет модель учиться распознавать объекты в неожиданном контексте и на стыках кадров, а также автоматически решает проблему обучения на малых объектах (так как мы видим 4 сцены в одном батче). Данная техника является стандартом для современных детекторов (семейство YOLO).

3. Смена архитектуры бэкбона (Backbone Transition)

Использованная сеть ResNet-50 является сверточной архитектурой, которая имеет ограниченное поле восприятия (receptive field).

* **Предложение:** Перейти на иерархические трансформеры, например, Swin Transformer. Механизм внимания позволяет модели лучше учитывать глобальный контекст сцены, что критически важно для отличения плотных групп людей от текстуры деревьев или брусчатки.

4. Test Time Augmentation

Для повышения метрик без переобучения модели.

* **Предложение:** Во время инференса подавать в сеть не только оригинальное изображение, но и его зеркальные отражения и версии с разным масштабом (Multi-scale testing). Усреднение предсказаний от разных вариаций одного снимка позволяет отфильтровать случайные выбросы и повысить уверенность (confidence) на сложных объектах.