

# Сравнение современных методов глубокого обучения для автоматической детекции белух в естественной среде

Булкин Антон Павлович

Московский государственный университет имени М. В. Ломоносова  
417 группа ММП ВМК  
Научный руководитель: Кравцова Ольга Анатольевна

2025

- Белуха — редкий и охраняемый вид морских млекопитающих, требующий регулярного мониторинга в естественной среде обитания.
- Традиционные методы (экспедиции, визуальный подсчёт, акустический мониторинг) трудоёмки, субъективны и плохо масштабируются.
- Методы компьютерного зрения и глубокого обучения позволяют:
  - уменьшить затраты на ручную обработку данных;
  - повысить воспроизводимость измерений;
  - масштабировать мониторинг по времени и пространству.
- Цель: автоматическая детекция и трекинг белух на аэровидео с дронов.

- Экологический контекст:
  - белухи — индикатор состояния морских экосистем;
  - необходимы объективные долгосрочные ряды наблюдений;
  - получения данных для дальнейшего исследований взаимодействия животных.
- Технические сложности:
  - малый размер животных в кадре, сильная рябь и блики на воде;
  - частичные окклузии, схожесть белух с фоновыми объектами (камни, лодки, птицы);
  - ограниченность размеченных данных и высокая стоимость аннотаций.
- В литературе используются разные датасеты и протоколы оценки, что осложняет прямое сравнение методов, а также рассматриваются старые поколения нейросетевых методов.

## Постановка задачи

- Исходный видеокорпус:  $\sim 450$  ГБ 4К-видео с бортов квадрокоптеров.
- Формируются наборы изображений с аннотацией в формате COCO:

$$\mathcal{S} = \{(x_i, Y_i)\}_{i=1}^N, \quad Y_i = \{(b_{ij}, c_{ij})\}_{j=1}^{M_i},$$

где  $b_{ij} = (x, y, w, h)$  — ограничивающая рамка,  $c_{ij}$  — класс.

- Два режима разметки:
  - **one**:  $C = \{\text{beluga}\}$  (только белухи);
  - **mlt**: многокласс  $C = \{\text{beacon, beluga, bird, person, rocks, seaweed, ship}\}$ .
- Детектор

$$f_\theta : \mathcal{X} \rightarrow 2^{\mathcal{B} \times \mathcal{C} \times [0,1]}, \quad f_\theta(x) = \{(\hat{b}_k, \hat{c}_k, \hat{p}_k)\}_{k=1}^{\hat{M}},$$

выдаёт рамки, классы и уверенности.

- Для видео дополнительно строится трекер  $h_{\theta, \varphi}$ , ассоциирующий детекции по кадрам и восстанавливающий траектории белух.

## Детекция

- Средняя точность по классам:

$$\text{mAP}@0.5 = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{AP}_c(\text{IoU} = 0.5),$$

$$\text{mAP}@[0.5 : 0.95] = \frac{1}{|\mathcal{C}| \cdot 10} \sum_{c \in \mathcal{C}} \sum_{\alpha=0.5}^{0.95} \text{AP}_c(\text{IoU} = \alpha),$$

шаг по  $\alpha$  равен 0.05.

- Временные характеристики:

- $\tau_{\text{train}}$  — среднее время одной эпохи обучения (мин);
- $\tau_{\text{inf}}$  — средняя задержка инференса на изображение (мс).

## Трекинг

- Точность многообъектного трекинга:

$$\text{MOTA}(h_{\theta,\varphi}) = 1 - \frac{\sum_{k=1}^K \sum_{t=1}^{T_k} (\text{FN}_{k,t} + \text{FP}_{k,t} + \text{IDSW}_{k,t})}{\sum_{k=1}^K \sum_{t=1}^{T_k} \text{GT}_{k,t}}.$$

- Дополнительно анализируются метрика IDF1 и число переключений IDSW.

## Трекинг

- Точность многообъектного трекинга:

$$\text{MOTA}(h_{\theta,\varphi}) = 1 - \frac{\sum_{k=1}^K \sum_{t=1}^{T_k} (\text{FN}_{k,t} + \text{FP}_{k,t} + \text{IDSW}_{k,t})}{\sum_{k=1}^K \sum_{t=1}^{T_k} \text{GT}_{k,t}}.$$

- IDF1 (F1-мера по идентичностям):

$$\text{IDF1} = \frac{2 \cdot \text{IDTP}}{2 \cdot \text{IDTP} + \text{IDFP} + \text{IDFN}},$$

где IDTP, IDFP и IDFН считаются на уровне целых траекторий.

- Число переключений идентичностей IDSW.

## Классический baseline

- Конвейер Computer Vision:  
предобработка → генерация кандидатов → SVM-классификация → NMS.

## Современные детекторы

- **YOLOv12**: одноэтапный детектор с backbone, FPN/PAN и многомасштабными головами; оптимизирован под компромисс “качество–скорость”.
- **YOLOWorld**: open-vocabulary вариант YOLO, расширенный текстовым энкодером  $t(c)$ ; классы задаются текстовыми описаниями.

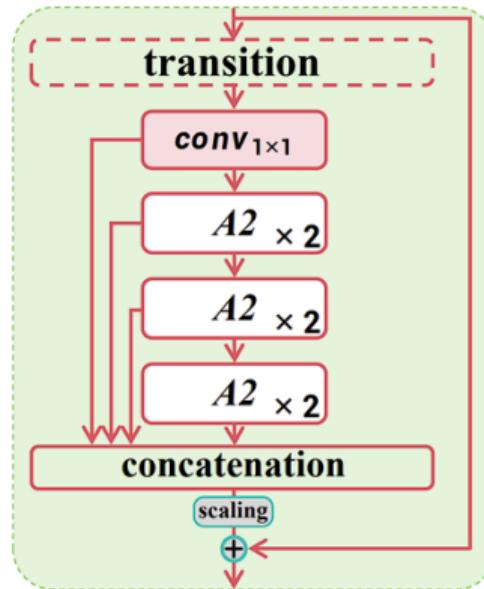


Figure 1: Residual Efficient Layer Aggregation Network as presented in paper

**Рис.:** Схема одноэтапного детектора семейства YOLO.

## Современные детекторы

- **RT-DETR**: детектор на основе трансформера (backbone + encoder-decoder), использующий объектные “queries” и глобальный контекст кадра.
- **RetinaNet**: одноэтапный детектор с FPN и Focal Loss, лучше обрабатывающий дисбаланс между фоном и объектами.

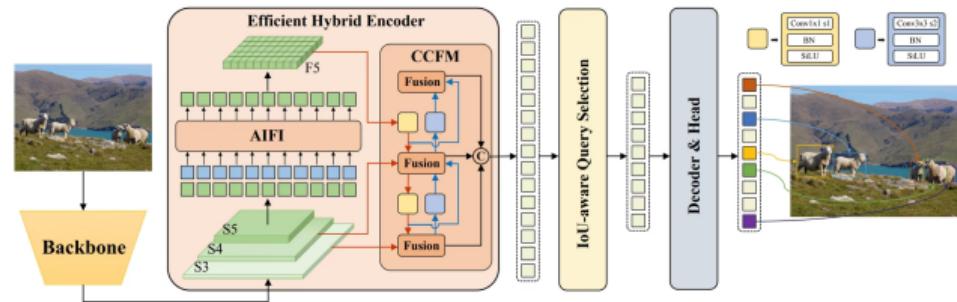


Рис.: Архитектура RT-DETR

- Вход трекера: последовательность кадров  $v^{(k)} = (x_{k,t})_{t=1}^{T_k}$  и детекции  $f_\theta(x_{k,t})$  на каждом кадре.
- Выход трекера: набор траекторий  $h_{\theta,\varphi}(v^{(k)}) = \{\tau_\ell^{(k)}\}_{\ell=1}^{L_k}$ , каждая  $\tau_\ell^{(k)}$  — отслеживаемая белуха с уникальным ID.

## Рассматриваемые трекеры

- DeepSORT
  - Калмановская фильтрация для прогноза положения объектов.
  - Венгерский алгоритм для сопоставления предсказаний и наблюдений.
  - Глубокие дескрипторы внешнего вида для устойчивости к окклюзиям.
- ByteTrack
  - Простая, но эффективная схема data association.
  - Использует детекции с высокой и средней уверенностью, уменьшая число пропусков.
- BoT-SORT
  - Развитие идей SORT/DeepSORT.
  - Более аккуратная работа с пересечениями траекторий и компенсацией движения камеры, использует дополнительные признаки скорости и формы.

# Данные

- Объёмы размеченных наборов:  
 $N \in \{400, 800, 1200, \sim 8000\}$ .
- Для  $N \in \{400, 800, 1200\}$  используются оба режима разметки (one/mlt).
- Для  $\sim 8000$  кадров аннотации получены полуавтоматически:
  - начальная ручная разметка;
  - обучение модели-помощника;
  - авторазметка неразмеченных кадров;
  - ручная валидация и исправление.
- Единый формат СОСО, единый препроцессинг и протокол обучения для всех моделей.

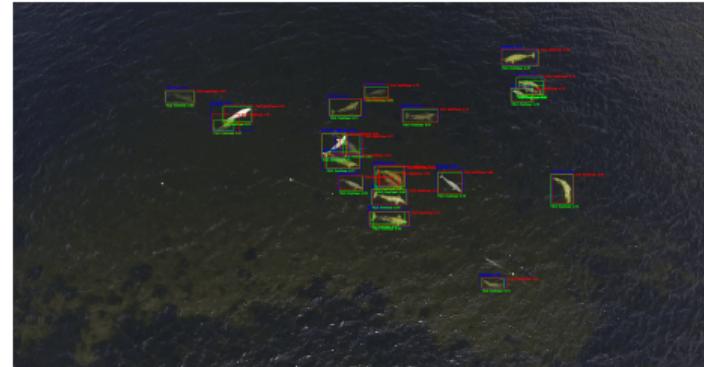


Рис.: Пример работы детекторов на реальном кадре.

# Анализ результатов: детекция

## Детекторы ( $N \approx 8000$ , лучший режим разметки)

Модель	mAP@0.5	mAP@[0.5:0.95]	$\tau_{\text{train}}$ (мин)	$\tau_{\text{inf}}$ (мс)
YOLOv12	0.989	0.652	13.850	17.600
YOLOWorld	0.951	0.626	16.350	10.560
RT-DETR	0.894	0.508	37.400	58.600
RetinaNet	0.869	0.473	0.830	4.200

- Увеличение объёма данных приводит к росту mAP для всех моделей.
- YOLOv12 даёт наивысший mAP@0.5 при умеренной задержке инференса.
- RT-DETR выигрывает по согласованности на сложном фоне, но заметно медленнее.
- RetinaNet — лёгкая и быстрая альтернатива при жёстких ограничениях по задержке.

# Анализ результатов: трекинг

## Трекеры (YOLOv12, режим mlt)

Трекер	MOTA	IDF1	IDSW	$\tau_{\text{det+trk}}$ (мс/кадр)
BoT-SORT	0.742	0.711	41	32.5
ByteTrack	0.789	0.770	35	30.8
DeepSORT	0.826	0.812	29	33.1

- DeepSORT показывает наибольшие MOTA и IDF1 при умеренном числе IDSW.
- BoT-SORT немного уступает по MOTA, но лучше обрабатывает пересечения траекторий.

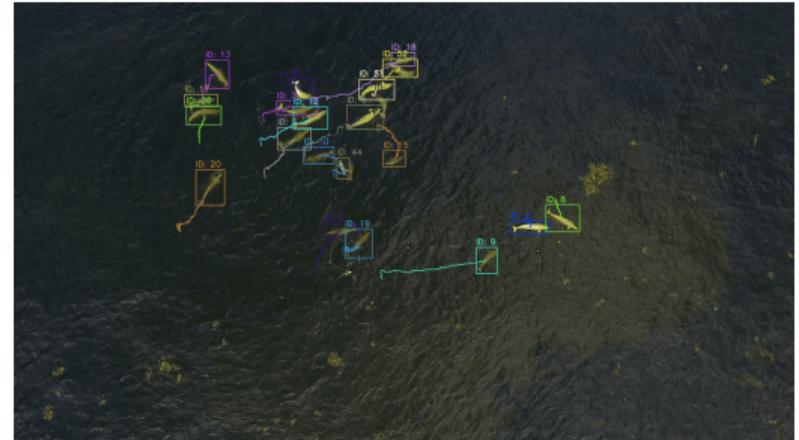


Рис.: Пример траекторий белух

# Заключение

- Сформирован воспроизводимый бенчмарк детекции и трекинга белух:
  - несколько масштабов выборки (от 400 до  $\sim 8000$  кадров);
  - два режима аннотаций: одноклассовый (one) и многоклассовой (mlt);
  - единый протокол обучения и оценки (mAP, MOTA, временные характеристики).
- Показано, что современные детекторы семейства YOLO значительно превосходят классический baseline и трансформерные/двуэтапные альтернативы на рассматриваемых данных.
- Интеграция детектора YOLOv12 с трекерами DeepSORT, ByteTrack и BoT-SORT позволяет получать устойчивые траектории белух.

**Спасибо за внимание!**