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

A Preprint

Булкин Антон Павлович
Московский государственный университет имени М. В. Ломоносова
Научный руководитель: Кравцова Ольга Анатольевна
bulkin261@gmail.com
<email@domain>

2025

Abstract

В работе исследуются современные архитектуры глубокого обучения для автоматического мониторинга белух — редкого и охраняемого вида, требующего внимательного и регулярного наблюдения. Цель исследования — повысить эффективность мониторинга животных, минимизируя затраты времени и ресурсов на ручную обработку данных, а также создать систему отслеживания взаимодействия животных между собой в рамках одного видеофайла. Предложен комплексный подход: классические решения CV, полуавтоматическая разметка видеоматериалов, обучение и сравнение современных детекторов и трекеров, а также анализ их точности и скорости.

Keywords детекция объектов · глубокое обучение · белуха · дроны · спутниковые снимки · трекинг · полуавтоматическая разметка

1 Введение

Белуха — редкий и охраняемый вид морских млекопитающих, для которого необходимы регулярные и объективные наблюдения в естественной среде обитания. Традиционные подходы — полевые экспедиции, акустический мониторинг, визуальный подсчёт — трудоёмки, подвержены субъективности и ограничены по пространственному и временному охвату. Автоматизация на основе методов компьютерного зрения и глубокого обучения позволяет снизить затраты, повысить воспроизводимость и масштабировать мониторинг.

Задача детекции морских млекопитающих в изображениях и видео обычно решается локализацией объектов с ограничивающими рамками и последующим трекингом. Ранние работы использовали классические техники и первые версии одноэтапных и двухэтапных детекторов, показывая базовую применимость, но чувствительность к условиям съёмки и фоновым помехам. Для видео широко применялись связки «детектор + трекер», повышающие устойчивость за счёт межкадровой ассоциации. Современные одноэтапные архитектуры семейства YOLO существенно улучшили соотношение точности и скорости и пригодны для сценариев реального времени. Трансформерные детекторы упрощают инференс, снимая необходимость в постобработке наподобие подавления немаксимумов, и повышают согласованность предсказаний на сложном фоне. Отдельное направление посвящено снижению стоимости аннотирования за счёт полуавтоматической разметки и стратегий активного обучения.

Ключевые трудности домена — малый видимый размер животных в кадре, блики и рябь воды, частичные окклюзии и близость визуально похожих фоновых объектов (лодки, буи, скалы, водоросли). Существенны доменный сдвиг между локациями и условиями съёмки и дисбаланс классов в многоклассовых сценариях. Высокая стоимость ручной аннотации приводит к ограниченным наборам и ошибкам разметки, что снижает переносимость моделей. Наконец, несогласованные протоколы оценки

в литературе — разные наборы данных, пороги пересечения, препроцессинг — затрудняют прямое сравнение результатов и выбор решений для практики.

В работе предлагается воспроизводимый бенчмарк детекции белух в реальной морской среде с унифицированным протоколом обучения и оценки. Подготовлены и сопоставлены наборы изображений разного масштаба — 400, 800, 1200 и около 8000 кадров, — где расширение корпуса достигается полуавтоматической разметкой: начальное обучение модели-помощника, автоматическая аннотация неразмеченных кадров и ручная валидация. Рассматриваются одноклассовая и многоклассовая разметки, что позволяет учитывать типичные фоновые объекты на воде.

2 Обзор литературы / Related works

Исследования, посвящённые автоматическому мониторингу белух и других морских млекопитающих, активно развиваются в последние годы и охватывают как аэровидеосъёмку, так и анализ спутниковых данных. Одной из существенных современных работ является статья Alsaidi et al. Alsaidi et al. [2024], где предложена система детекции и трекинга белух на аэровидео с использованием YOLOv7 и алгоритма DeepSORT. Модель показала высокую точность и полноту, а также устойчивость трекинга после постобработки. Аналогичные идеи развиваются в работе Harasyn et al. Harasyn et al. [2022], где YOLOv4 и DeepSORT применялись для детекции белух, каяков и лодок в видеопотоках с дронов; достигнута точность около 74% и полнота 72%. Lee et al. Lee et al. [2021] исследовали RetinaNet и Faster R-CNN на данных залива Камберленд, отметив, что использование тайлинга крупных изображений существенно повышает точность и уменьшает количество ложных срабатываний.

Особое внимание уделяется построению и масштабированию наборов данных. Araújo et al. Araújo et al. [2022] представили датасет Beluga-5k с более чем 5500 фотографиями белух и предложили полуавтоматическую разметку. Лучшая модель (YOLOv3-tiny) достигла 97,05% mAP@0.5, корректно обнаруживая белух даже в условиях сложного фона. Boulent et al. Boulent et al. [2023] предложили интерактивную схему «человек в контуре», при которой нейросеть, обученная на 100 изображениях, достигла 91% совпадения с экспертами и позволила ускорить аннотацию более чем в пять раз. Cubaynes и Fretwell Cubaynes and Fretwell [2022] опубликовали набор спутниковых изображений Whales from Space, включающий сотни размеченных примеров китов, что обеспечило возможность обучения моделей на данных высокого разрешения.

Использование спутниковых снимков для мониторинга млекопитающих подтверждается рядом работ. Green et al. Green et al. [2023] применили YOLOv5 для обнаружения серых китов на спутниковых снимках и достигли точности 80–94% при полноте 84–89%. Guirado et al. Guirado et al. [2019] предложили каскадный подход, объединяющий классификацию наличия китов и подсчёт особей, что увеличило общую точность на 36% по сравнению с одноэтапными методами. Borowicz et al. Borowicz et al. [2019] показали, что CNN, обученные на аэрофотоснимках, могут быть перенесены на спутниковые данные без значительной потери качества.

В области общих архитектур объектной детекции значительную роль сыграли двухэтапные модели (Faster R-CNN Ren et al. [2015]), обеспечившие высокую точность за счёт сети предложений регионов. Одноэтапные детекторы, начиная с YOLO Redmon et al. [2016] и YOLO9000 Redmon and Farhadi [2017], позволили объединить локализацию и классификацию в одном прогоне сети, достигнув 45–60 FPS. RetinaNet Lin et al. [2017] с функцией Focal Loss улучшил устойчивость к дисбалансу классов и повысил точность при сохранении скорости. Последующие версии YOLO (в частности, YOLOv7 Wang et al. [2022]) установили новый стандарт по соотношению точности и производительности.

Трансформерные подходы (DETR Carion et al. [2020], RT-DETR Zhao et al. [2023]) позволили отказаться от постобработки (NMS) и выполнять end-to-end детекцию. RT-DETR показал сопоставимую с YOLO точность при сохранении real-time скорости, что делает его перспективным для видеоаналитики. Модели открытого словаря (ViLD Gu et al. [2021], GLIP Li et al. [2022], YOLO-World Cheng et al. [2024]) расширили возможности детекции на неизвестные классы, включая морские объекты, что особенно важно для практических систем, где структура сцены заранее не фиксирована.

Обзор Tuia et al. Tuia et al. [2022] подчёркивает важность машинного обучения для экологического мониторинга. Авторы указывают, что интеграция глубоких моделей с полевыми данными и дроновыми системами позволяет существенно сократить стоимость и повысить масштабируемость наблюдений. Совокупно, эти исследования формируют основу для разработки универсального бенчмарка детекции белух и показывают актуальность сопоставления современных архитектур на единых данных и метриках.

3 Постановка задачи

3.1 Алгебраическая и вероятностная структура данных

Имеется исходный видеокорпус объёмом порядка 450 ГБ (4K) с бортов квадрокоптеров, из которого формируются выборки изображений размером $N \in \{400, 800, 1200, \sim 8000\}$ с аннотациями в формате COCO. Каждое изображение $x \in \mathcal{X}$ снабжено множеством объектов $Y = \{(b_j, c_j)\}_{j=1}^{M}$, где $b_j = (x, y, w, h)$ — ограничивающая рамка, $c_j \in \mathcal{C}$ — метка класса. Рассматриваются два варианта аннотаций: one (однокласс: $\mathcal{C} = \{\texttt{beluga}\}$) и mlt (многокласс: $\mathcal{C} = \{\texttt{beacon, beluga, bird, person, rocks, seaweed, ship}\}$). Разбиения $X^{\text{train}}, X^{\text{val}}, X^{\text{test}}$ фиксируются на уровне изображений ($\texttt{split} \in \{\texttt{train,val,test}\}$). Формально считаем, что $(X, Y) \sim \mathcal{D}$, где $\mathcal{D}$ — неизвестное распределение, а обучающая выборка $\mathcal{S} = \{(x_i, Y_i)\}_{i=1}^{N}$ представляет собой i.i.d. реализацию из $\mathcal{D}$; Y_i — конечное множество пар «рамка–класс» для x_i. Для расширенного набора ~ 8000 кадров аннотации получены полуавтоматически: начальная ручная разметка $\rightarrow$ предобученный детектор $\rightarrow$ авторазметка $\rightarrow$ ручная верификация.

3.2 Отображение и вычислительный конвейер

Цель — построить детектор

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

который по входному изображению выдаёт множество предсказанных боксов, классов и оценок уверенности. На практике f_θ реализуется как композиция стадий

$$f_\theta \ = \ \underbrace{\pi_{\text{post}}}_{\text{постобработка}} \ \circ \ \underbrace{d_\theta}_{\text{нейросетевой детектор}} \ \circ \ \underbrace{\phi}_{\text{препроцессинг}} \ ,$$

где ϕ приводит изображение к требуемому масштабу/формату; d_θ — параметризуемая модель; π_{post} — схема отбора и консолидации предсказаний. Для open-vocabulary варианта конвейер расширяется текстовым энкодером $t(\cdot)$, формируя совместное представление «изображение–текст» для классов. В одноклассовом режиме one пространство меток вырождается до $\{\texttt{beluga}\}$, а mlt учитывает семантически близкие фоны (лодки/буи/скалы и т. п.), снижая ложные срабатывания. Реализация конвейера, структура датасетов и сценарии one/mlt подробно заданы в курсовой работе и репозитории проекта.

3.3 Внешние критерии качества

Оценка проводится на тестовой части X^{test} по стандартным метрикам детекции:

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

шаг по α равен 0.05. Дополнительно фиксируются показатели производительности: средняя задержка инференса на кадр τ_{inf} (мс/кадр) и при необходимости среднее время одной эпохи обучения τ_{train} (мин/эпоха). Для сопоставимости конфигурации препроцессинга, пороги уверенности, параметры NMS (если применимо) и разметочные режимы one/mlt унифицированы по моделям и наборам N.

3.4 Оптимизационная постановка

Параметры θ оцениваются методом минимизации эмпирического риска на обучающем множестве с использованием детектор-специфичной функции потерь:

$$\hat{\theta} \ \in \ \arg\min_{\theta} \ \frac{1}{N} \sum_{i=1}^{N} \ \mathcal{L}\big(d_\theta(\phi(x_i)),\, Y_i\big) \ + \ \lambda \mathcal{R}(\theta),$$

где $\mathcal{L}$ включает классификационную компоненту и регрессию боксов, $\mathcal{R}$ — регуляризация, $\lambda \geq 0$. С учётом практических ограничений возможна многокритериальная постановка с ограничением на задержку инференса:

$$\text{minimize} \ \ -\,\text{mAP@0.5}(\theta) \quad \text{s.t.} \quad \tau_{\text{inf}}(\theta) \leq \tau^\star,$$

или эквивалентно — скаляризация качества и скорости:

$$\min_{\theta} \ \Big[-\,\text{mAP@0.5}(\theta) \ + \ \mu\, \tau_{\text{inf}}(\theta) \Big], \qquad \mu \geq 0,$$

что отражает прикладной компромисс «точность/реальное время» для мониторинга с БПЛА и береговых станций. Принятые в работе метрики и протоколы обучения/валидации/теста, а также конкретные наборы N описаны в курсовой и используются далее при сравнении моделей.

Список литературы

M. Alsaidi et al. Localization and tracking of beluga whales in aerial video using deep learning. Frontiers in Marine Science, 2024. doi:10.3389/fmars.2024.1445698.

Madison L. Harasyn, Wayne S. Chan, Emma L. Ausen, and David G. Barber. Detection and tracking of belugas, kayaks and motorized boats in drone video using deep learning. Journal of Unmanned Vehicle Systems, 2022. doi:10.1139/juvs-2021-0024.

P. Q. Lee et al. Beluga whale detection in the cumberland sound bay using convolutional neural networks. Canadian Journal of Remote Sensing, 2021. doi:10.1080/07038992.2021.1901221.

Voncarlos M. Araújo, Ankita Shukla, Clément Chion, Sébastien Gambs, and Robert Michaud. Machine-learning approach for automatic detection of wild beluga whales from hand-held camera pictures. Sensors, 22(11):4107, 2022. doi:10.3390/s22114107.

J. Boulent et al. Scaling whale monitoring using deep learning: A human-in-the-loop solution for analyzing aerial datasets. Frontiers in Marine Science, 2023. doi:10.3389/fmars.2023.1099479.

H. C. Cubaynes and P. T. Fretwell. Whales from space dataset: an annotated satellite image dataset of whales for training machine learning models. Scientific Data, 2022. doi:10.1038/s41597-022-01377-4.

K. M. Green et al. Gray whale detection in satellite imagery using deep learning. Remote Sensing in Ecology and Conservation, 2023. doi:10.1002/rse2.352.

E. Guirado et al. Whale counting in satellite and aerial images with deep learning. Scientific Reports, 2019. doi:10.1038/s41598-019-50795-9.

A. Borowicz et al. Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. PLOS One, 2019. doi:10.1371/journal.pone.0212532.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NeurIPS, 2015.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016.

Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In CVPR, 2017.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV, 2017.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

Yian Zhao et al. Detrs beat yolos on real-time object detection (rt-detr). arXiv preprint arXiv:2304.08069, 2023.

X. Gu et al. Open-vocabulary object detection via vision-and-language knowledge distillation (vild). In CVPR, 2021.

L. H. Li et al. Grounded language-image pre-training (glip). In CVPR, 2022.

T. Cheng et al. Yolo-world: Real-time open-vocabulary object detection. In CVPR, 2024.

D. Tuia et al. Perspectives in machine learning for wildlife conservation. Nature Communications, 2022. doi:10.1038/s41467-022-27980-y.