ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2025 Управление, вычислительная техника и информатика № 72
Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 004.93`12
doi: 10.17223/19988605/72/4

Модели сверточных нейронных сетей YOLO с механизмом внимания для систем компьютерного зрения реального времени

Вадим Александрович Клековкин[1], Николай Григорьевич Марков[2], Степан Геннадьевич Небаба[3]

[1, 2, 3] Национальный исследовательский Томский политехнический университет, Томск, Россия
[1] vak37@tpu.ru
[2] markovng@tpu.ru
[3] stepanlfx@tpu.ru

Аннотация. Разрабатываются и исследуются новые модели сверточных нейронных сетей (СНС) с механизмом внимания, позволяющие решать задачи объектного детектирования малоразмерных летающих объектов (ЛО). В качестве исходных моделей выбрано две базовых модели СНС класса YOLO: YOLOv5s и YOLOv8s. На их основе создано четыре гибридных модели СНС с использованием модулей SWT и SEA, реализующих различные варианты механизма внимания. Для обучения, валидации и комплексного исследования этих моделей использовался датасет, изображения которого содержат от одного до нескольких ЛО трех классов: «Птица», «Беспилотный летательный аппарат (БПЛА) самолетного типа» и «БПЛА вертолетного типа». Исследования показали, что гибридная модель YOLOv8s + SEA является наиболее предпочтительным вариантом при создании систем компьютерного зрения реального времени для детектирования малоразмерных ЛО.

Ключевые слова: система компьютерного зрения реального времени; сверточная нейронная сеть YOLO; обнаружение и классификация летающих объектов; механизм внимания.

***Для цитирования*:** 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

Original article
doi: 10.17223/19988605/72/4

YOLO convolutional neural network models with attention mechanism for real-time computer vision systems

Vadim A. Klekovkin[1], Nikolay G. Markov[2], Stepan G. Nebaba[3]

[1, 2, 3] National Research Tomsk Polytechnic University, Tomsk, Russian Federation, vak37@tpu.ru
[1] vak37@tpu.ru
[2] markovng@tpu.ru
[3] stepanlfx@tpu.ru

Abstract. New models of convolutional neural networks (SNN) with an attention mechanism are being developed and investigated to solve the problems of object detection of small-sized flying objects (FO). Two basic CNN models of the YOLO class were selected as the initial ones for the development of new CNN models: YOLOv5s and YOLOv8s. Based on them, four hybrid CNN models were created using the SWT module and the SEA module, implementing different versions of the attention mechanism. For training, validation and research of the basic and hybrid models, a dataset with labeled images of small-sized FO of three classes was used: «Unmanned aerial vehicle (UAV) of helicopter type», «UAV of airplane type» and «Bird». Research has demonstrated that the hybrid YOLOv8s + SEA model is the most preferable option for designing real-time computer vision systems intended for the detection of small-sized FO.

Keywords: real-time computer vision system; YOLO convolutional neural network; detection and classification of flying objects; attention mechanism.

***For citation*:** Klekovkin, V.A., Markov, N.G., Nebaba, S.G. (2025) YOLO convolutional neural network models with attention mechanism for real-time computer vision systems. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 72. pp. 39–50. doi: 10.17223/19988605/72/4

Введение

В последние годы в России и за рубежом интенсивно ведутся разработки систем компьютерного зрения (СКЗ) различного назначения. Чаще всего СКЗ создаются на основе современных моделей сверточных нейронных сетей (СНС). Именно СКЗ с такими моделями СНС должны позволять решать прикладные задачи компьютерного зрения, относящиеся к четырем классам [1, 2]: семантическая сегментация (Semantic Segmentation) изображений, распознавание единичного объекта (Object Recognition) на изображении, объектное детектирование (Object Detection) и, наконец, четвертый класс задач – распознавание экземпляров объекта (Instance Segmentation) одного класса на изображении.

Отметим, что сегодня наиболее востребованными являются СКЗ, позволяющие решать задачи объектного детектирования, когда на изображениях обнаруживаются и классифицируются объекты различных классов. Именно такие СКЗ используются в системах управления автономным наземным и воздушным транспортом, в системах мониторинга воздушного пространства и мониторинга опасных технологических объектов [3, 4]. Учитывая, что во многих случаях объекты на анализируемых такими системами изображениях являются подвижными, СКЗ, чтобы обнаружить и удерживать их в поле зрения на изображении, должны функционировать в режиме реального времени. При этом масштаб реального времени в каждой конкретной задаче определяется потенциальной скоростью перемещения подвижных объектов. Важной проблемой при создании многих СКЗ реального времени для решения задач объектного детектирования подвижных объектов является необходимость детектирования объектов разных размеров (по линейным размерам, по площади). В [4, 5] показано, что детектирование объектов малого размера (часто говорят, масштаба) представляет особую сложность для современных моделей СНС. Сегодня, несмотря на общие успехи в достижении высокой точности детектирования с помощью таких моделей, все ещё сохраняется значительный разрыв между точностью детектирования объектов малого и большого масштабов. Все это указывает на актуальность создания высокоточных моделей СНС для решения задач детектирования объектов малых размеров с помощью СКЗ реального времени.

В данной статье рассматривается задача объектного детектирования на изображениях малоразмерных подвижных объектов, находящихся в воздушном пространстве и представляющих собой летающие объекты (ЛО) трех классов. Для решения такой задачи необходимо разрабатывать СКЗ реального времени, удовлетворяющие жестким требованиям по производительности и точности детектирования ЛО малых размеров. В основу этих СКЗ должны быть положены модели СНС, также удовлетворяющие таким требованиям. Для этого в работе создаются и исследуются гибридные модели СНС класса YOLO с использованием механизма внимания. По результатам их комплексных исследований выбирается наиболее эффективная с точки зрения точности и скорости объектного детектирования модель (модели) для использования в СКЗ реального времени.

1. Системы компьютерного зрения для детектирования подвижных объектов

Объектное детектирование подвижных объектов является одной из основных задач компьютерного зрения и выполняется как совокупность трех подзадач: обнаружения одного или нескольких объектов на изображении, их локализации (определение местоположения объектов на изображении) и выявления класса каждого из объектов. Далее под термином «детектирование» будем понимать процесс решения этих трех подзадач. Для решения задач детектирования в самых различных областях сегодня

на основе современных моделей СНС разрабатываются СКЗ, учитывающие специфику решения каждой задачи. Так, при создании СКЗ, осуществляющих мониторинг подвижных объектов, учитывается, что такие системы должны как обнаруживать и классифицировать объекты, так и отслеживать перемещение каждого из них в пространстве. Для этого СКЗ должны функционировать в режиме реального времени, причем масштаб реального времени в значительной степени определяется скоростью перемещения объектов в пространстве. В работах [3–5] показано, что во многих современных СКЗ, функционирующих в режиме реального времени, значительные ресурсы отводятся на вычисление модели СНС. Отсюда следует, что при создании перспективных СКЗ необходимо использовать эффективные по скорости вычисления модели СНС. Обычно для оценки минимальной достаточной скорости детектирования объектов на изображении с помощью СКЗ реального времени используется известный показатель – количество анализируемых изображений / кадров в секунду (Frames Per Second, FPS), причем в большинстве случаев детектирования подвижных объектов значение этого показателя должно быть не менее 25–30, а часто и значительно более этого порога [3, 5].

Другим основным требованием к модели СНС, планируемой для включения в состав СКЗ реального времени, является высокая точность детектирования (обнаружения и классификации) подвижных объектов на изображениях. Отметим, что эти требования к модели СНС – высокая скорость ее вычисления и высокая точность детектирования подвижных объектов на изображениях – противоречат друг другу, так как для повышения точности детектирования требуются более ресурсоемкие модели СНС, скорость вычисления которых заметно меньше, чем у более компактных моделей. Общепринятыми метриками оценки точности детектирования объектов на изображении являются AP (Average Precision, средняя точность классификации) для каждого класса объектов и mAP (mean Average Precision) – усредненное значение AP по всем классам. Чаще всего применяют метрики $AP_{0,5}$ и $mAP_{0,5}$, которые получают для порогового значения IoU (метрики, показывающей долю пересечения двух рамок на объекте), равного 0,5 [5, 6]. Точность детектирования объектов на изображениях можно считать достаточной, если значения метрики $AP_{0,5}$ для каждого класса объектов и метрики $mAP_{0,5}$ по всем классам объектов выше заданного порогового значения. Опираясь на результаты работ [2, 4–7], выявлено, что это пороговое значение обычно не менее 0,9.

На изображениях может одновременно присутствовать несколько подвижных объектов разных классов, причем они могут иметь разные размеры. Детектирование объектов, размеры которых относительно размеров анализируемого изображения малы (обычно не более 32 × 32 пикселя по занимаемой каждым из них площади, как это было предложено при формировании известного датасета MS COCO [8]), представляет особую сложность. В [6, 7] показано, что к малоразмерным объектам следует отнести, например, птиц и беспилотные летательные аппараты (БПЛА) самолетного типа, которые находятся на значительном удалении от СКЗ. Контуры таких объектов обладают высокой степенью схожести, что делает их классификацию непростой задачей. Отсюда можно сделать вывод об актуальности разработки высокоточных моделей СНС для решения задач детектирования объектов малых размеров с помощью СКЗ реального времени.

2. Постановка задачи исследований

Цель данного проекта – разработка и исследование новых моделей СНС с механизмом внимания на основе хорошо себя зарекомендовавших моделей из класса YOLO. Выбранная по результатам исследований наиболее эффективная из таких гибридных моделей СНС должна детектировать на RGB-изображениях малоразмерные подвижные объекты и удовлетворять изложенным выше требованиям по точности их детектирования и скорости вычисления модели. Это позволит рекомендовать такую модель СНС для реализации в СКЗ реального времени. В качестве подвижных объектов на изображениях будем рассматривать малоразмерные летающие объекты (ЛО) в воздушном пространстве трех классов: БПЛА вертолетного типа, БПЛА самолетного типа и класс «Птица».

Для обучения, валидации и исследования разработанных и базовых моделей СНС будем использовать датасет с малоразмерными ЛО на размеченных RGB-изображениях размером 416 × 416 пикселей

из [7]. Число объектов в этом датасете 4 540, при этом распределение по классам ЛО следующее: БПЛА вертолетного типа – 1 437 объектов, БПЛА самолетного типа – 1 663 объекта, объектов класса «Птица» – 1 440. На одном изображении может присутствовать один либо два и более ЛО. Датасет разделен на обучающую, валидационную и тестовую выборки в следующем соотношении: 70% всех изображений – обучающая выборка, 20% изображений – валидационная, 10% изображений – тестовая выборка.

Кратко остановимся на выборе исходных (далее – базовых) моделей, на основе которых будут создаваться гибридные модели СНС с механизмом внимания. Анализ моделей СНС для решения задач объектного детектирования, проведенный в работах [4–6], позволяет считать, что наиболее подходящими для детектирования ЛО на изображениях с учетом требования к высокой скорости вычисления моделей являются модели класса YOLO [9]. Модели СНС этого класса относятся к одноэтапным детекторам и поэтому являются высокопроизводительными. Кроме того, как показывает опыт их применения [5, 9], они имеют весьма высокую точность детектирования объектов на изображениях. Предлагается выбрать базовые модели СНС из этого класса, начиная с моделей с современными архитектурами из семейства YOLOv5 [10] и заканчивая хорошо себя зарекомендовавшими моделями семейства YOLOv8 [11]. Как следует из результатов исследований [11], полученных при решении одной из прикладных задач, модели семейства YOLOv5 являются более высокопроизводительными, чем модели семейства YOLOv8, однако часть моделей второго семейства показывает более высокую точность детектирования. В итоге в качестве базовых моделей из этих семейств выбраны модели YOLOv5s и YOLOv8s. Отметим, что согласно результатам исследований из [5, 7, 10, 11] эти модели являются компромиссными вариантами, поскольку каждая из них в своем семействе в значительной степени удовлетворяет взаимоисключающим требованиям по точности детектирования объектов и скорости вычисления модели. Появившиеся недавно модели YOLOv9, YOLOv10 и YOLOv11 пока недостаточно подробно исследованы, а также практически не апробированы при решении прикладных задач. По этой причине здесь они не рассматривались в качестве базовых моделей СНС.

3. Гибридные модели СНС

Проведенный анализ исследований наиболее известных моделей нейронных сетей с механизмом внимания [12–18] позволил считать, что для разработки гибридных моделей СНС из довольно большого многообразия моделей, где реализован механизм внимания, следует выбрать модули с механизмами внимания SEAttention [16] и SWIN-Transformer [17]. Модуль SEAttention (далее – SEA) широко применяется благодаря своей простоте, вычислительной эффективности и способности повышать репрезентативную способность исходных моделей СНС. Этот модуль содержит блок сжатия и блок возбуждения (SE, Squeeze-Excitation), которые используются для сбора глобальной информации, захвата взаимосвязей по каналам и улучшения репрезентативности. Глобальная пространственная информация собирается в модуле сжатия путем глобального усреднения. Модуль возбуждения захватывает взаимосвязи по каналам и выводит вектор внимания с помощью полностью связанных слоев и нелинейных слоев (ReLU и сигмоида). Затем каждый канал входного признака масштабируется путем умножения соответствующего элемента в векторе внимания.

Модуль SWIN-Transformer – это иерархический Transformer, представление которого вычисляется с помощью смещенных окон [17]. Схема смещенных окон обеспечивает большую эффективность, ограничивая вычисления собственного внимания неперекрывающимися локальными окнами, а также допуская межоконные соединения. Модуль SWIN-Transformer (далее – SWT) извлекает визуальные признаки на основе механизма самовнимания, что позволяет захватывать глобальную и локальную контекстную информацию об объектах на изображении и в итоге улучшает извлечение признаков объектов. Подход смещенных окон обеспечивает лучшую масштабируемость и производительность этого модуля по сравнению с хорошо известными трансформерами.

Рассмотрим более детально каждую из разработанных с использованием модулей SEA и SWT гибридных моделей.

3.1. Модель YOLOv5s + SEA

Модуль SEA добавлен в архитектуру базовой модели YOLOv5 путем замены блока C3 (кросс-ступенчатый частичный блок, состоящий из трех сверточных слоев с пропусками соединений) в ее структуре Backbone. Полученная таким образом гибридная модель YOLOv5s + SEA имеет архитектуру, показанную на рис. 1. Добавленный модуль SEA выделен на этой схеме розовым цветом.

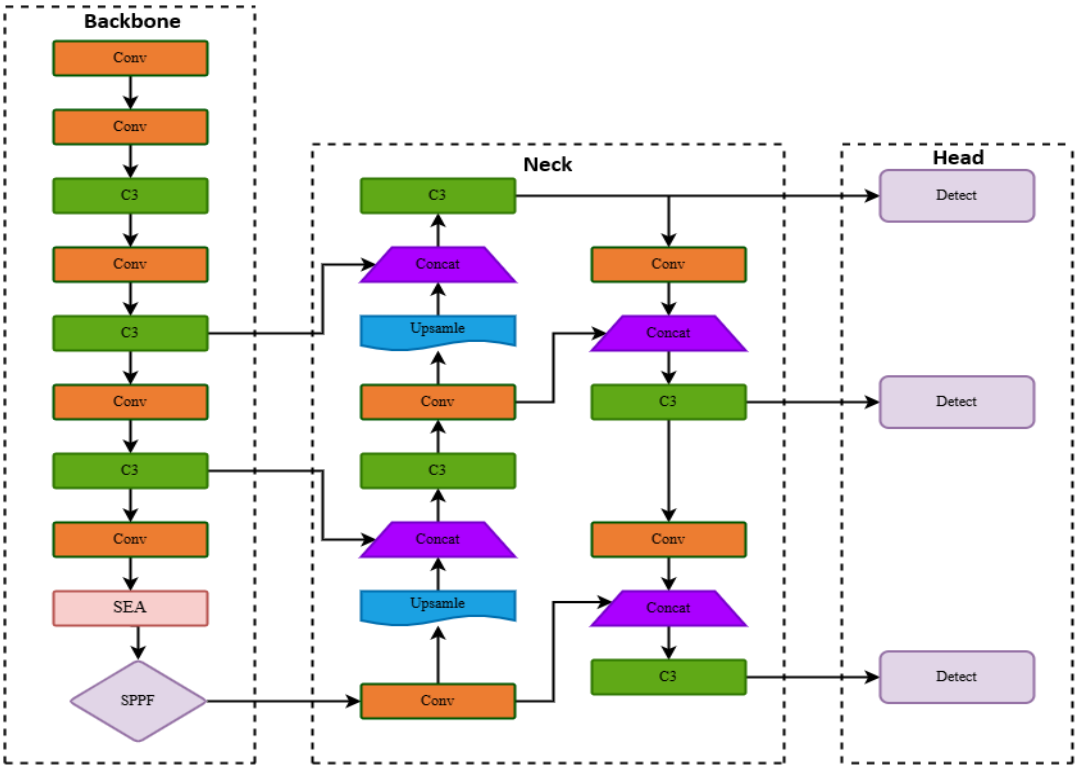

Рис. 1. Архитектура гибридной модели YOLOv5s + SEA
Fig. 1. Architecture of the hybrid model YOLOv5s + SEA

Известно, что блок C3 в классической реализации YOLOv5s обладает несколькими недостатками, ключевыми из которых являются отсутствие канального внимания (равное отношение ко всем признакам, в том числе шумовым) и слабая адаптация к мелким объектам (отсутствует механизм адаптивного усиления полезных признаков).

Модуль SEA выполняется в три этапа: этап Squeeze – глобальное усреднение по пространственным измерениям, этап Excitation – поиск зависимостей между каналами путем обучения весов двуслойного перцептрона, наконец, Scale – масштабирование полученного вектора внимания на пространство признаков. Это позволяет усилить информативные каналы и подавить шумовые. При этом количество вычислений растет незначительно, т.е. на скорости вычислений замена блока сказывается минимально.

Использование гибридной модели для решения задачи объектного детектирования малоразмерных объектов позволит улучшить фокусировку на целевой информации о таких объектах, а также подавить нерелевантную информацию о признаках ЛО на изображениях, что должно в итоге привести к повышению точности детектирования малоразмерных объектов.

3.2. Модель YOLOv5s + SWT

Ключевыми особенностями модуля SWT являются:

– иерархическая структура – работа с изображениями как с последовательностью патчей, которые постепенно объединяются в более крупные блоки (аналогично блокам модели CHC), сохраняя свойство мультимасштабности;

– оконное внимание (Window-based Self-Attention) – разбиение изображения на неперекрывающиеся локальные окна и вычисление параметра внимания только внутри окна, что снижает вычислительную сложность модуля;

– смещенные окна (Shifted Windows) – на каждом следующем слое окна сдвигаются на половину их размера, что позволяет учитывать связи между разными областями;

– линейная сложность – благодаря локальным окнам и иерархии SWIN-Transformer масштабируется лучше классических модулей внимания.

Указанные особенности позволяют предположить, как и в случае модуля SEA, что гибридная модель за счет модуля SWT должна быть более эффективной при детектировании малоразмерных ЛО, чем использование классических моделей СНС. Точность детектирования объектов может повышаться также за счет лучшего захвата глобальных зависимостей на изображении.

Вместе с тем гибридная модель, использующая модуль SWT, требует больше памяти, чем базовая модель СНС, и вычислительно она более сложна.

Поскольку в модуле SWT схема смещенных окон обеспечивает большую эффективность выделения признаков объектов на изображении, ограничивая вычисления собственного внимания неперекрывающимися локальными окнами, а также допуская межоконные соединения, добавим модуль SWT в архитектуру базовой модели YOLOv5s путем замены блока C3 в структуре Backbone. Архитектура полученной таким образом гибридной модели, получившей название YOLOv5s + SWT, приведена на рис. 2.

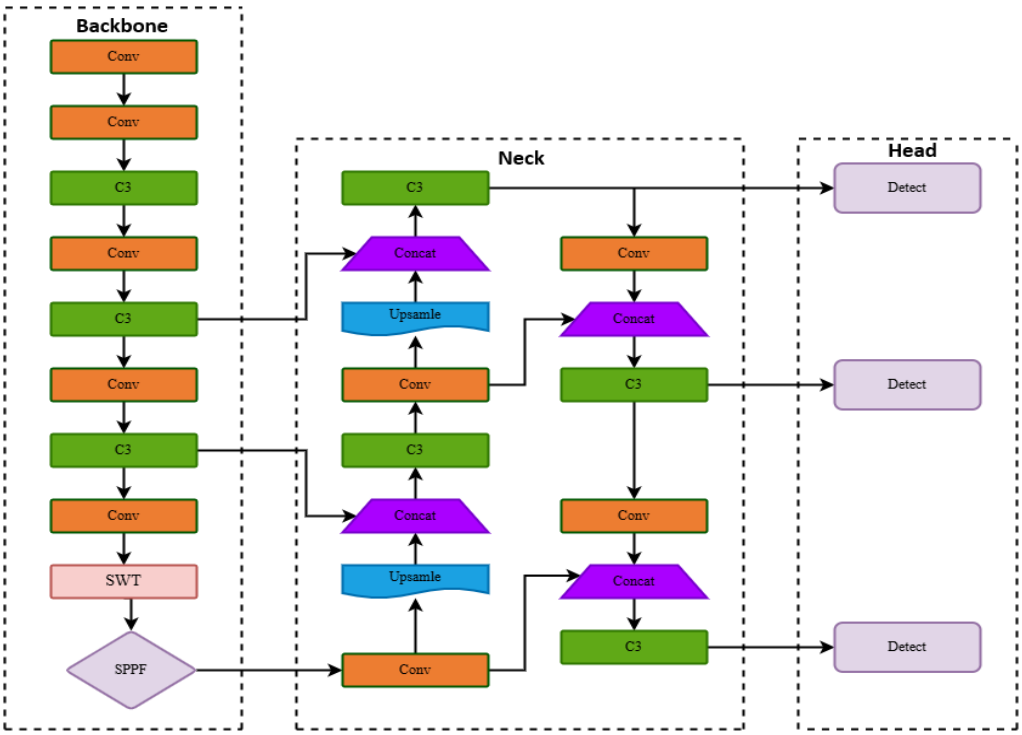

Рис. 2. Архитектура гибридной модели YOLOv5s + SWT
Fig. 2. Architecture of the hybrid model YOLOv5s + SWT

Добавленный модуль SWT выделен на этой схеме розовым цветом. Реализованный в модели механизм самовнимания, который позволяет захватывать глобальную и локальную контекстную информацию об объектах, ведет к улучшению извлечения признаков. Это должно позволить лучше сохранить контекстную информацию об ЛО и в итоге повысить точность детектирования объектов.

3.3. Модель YOLOv8s + SEA

При построении гибридной модели на основе базовой модели YOLOv8s возможны два основных варианта использования модуля SEA – замена блоков C2F (аналогично тому, как в модели YOLOv5s

заменяется блок C3), либо включение модуля SEA как дополнительного блока без изменения существующей последовательности блоков архитектуры базовой модели. Последний вариант имеет ряд преимуществ, в частности подразумевает минимальное вмешательство в оригинальную архитектуру базовой модели, обратную совместимость с этой архитектурой и улучшение работы с мультимасштабными объектами на изображении.

В случае модели YOLOv8s было решено выбрать вариант добавления модуля SEA в структуру Neck ее архитектуры. Благодаря включению модуля SEA в структуру Neck механизм внимания помогает модели определить, какие признаки наиболее важны, прежде чем передавать их на уровни обнаружения объекта. Архитектура гибридной модели YOLOv8s + SEA приведена на рис. 3. Добавленный модуль SEA выделен выделен на схеме розовым цветом.

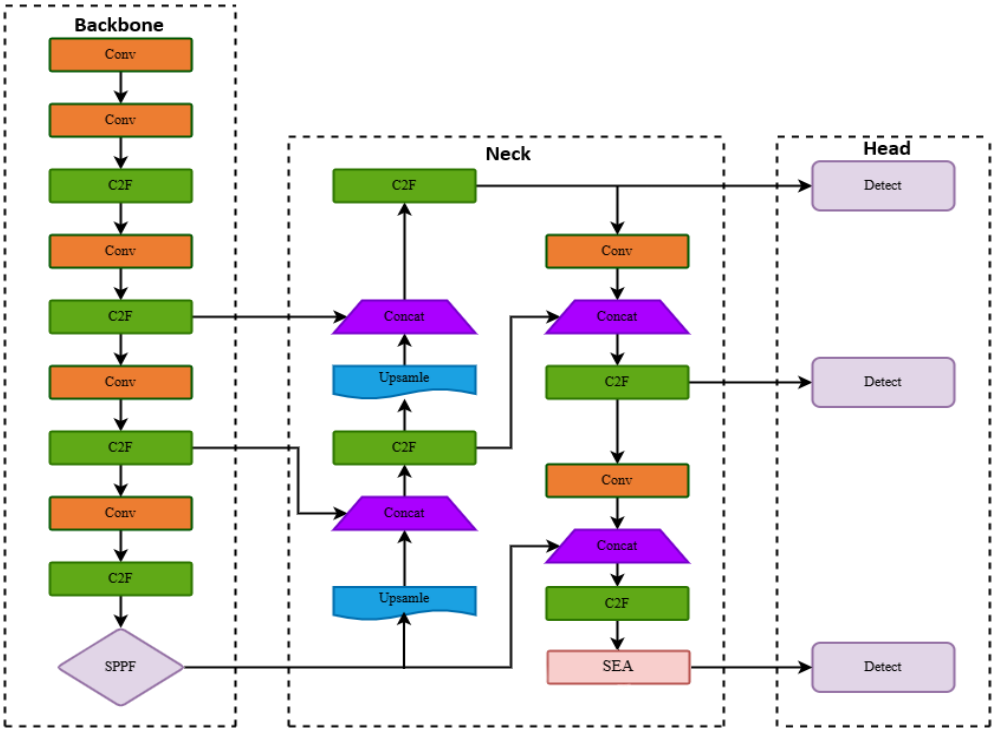

Рис. 3. Архитектура гибридной модели YOLOv8s + SEA
Fig. 3. Architecture of the hybrid model YOLOv8s + SEA

Рассмотрим особенности этой гибридной модели:

– SEA-модуль располагается в месте слияния признаков, что минимизирует влияние на вычислительную сложность гибридной модели;

– количество SEA-модулей можно относительно легко регулировать;

– сохраняется оригинальная архитектура базовой модели YOLOv8s.

3.4. Модель YOLOv8s + SWT

Блок C2F в архитектуре модели YOLOv8s представляет собой усовершенствованный вариант блока C3 в архитектуре модели YOLOv5s с дополнительными skip-соединениями (более быстрая реализация узкого места CSP с двумя свертками). Основной его функцией является агрегация мультимасштабных признаков на разных уровнях модели. Этот блок, как и блок C3 в архитектуре модели YOLOv5s, обладает рядом ограничений, среди которых локальная природа сверточных операций, ограниченное рецептивное поле и фиксированные шаблоны извлечения признаков объектов.

При этом модуль SWT позволяет реализовать глобальное взаимодействие признаков через механизм внимания, поддерживает адаптивную оценку значимости регионов и иерархическое представление признаков объектов.

Учитывая изложенное, было решено добавить модуль SWT в архитектуру модели YOLOv8s путем замены четвертого блока C2F в структуре Backbone. Архитектура гибридной модели YOLOv8s + SWT приведена на рис. 4. Добавленный модуль SWT на рис. 4 выделен розовым цветом.

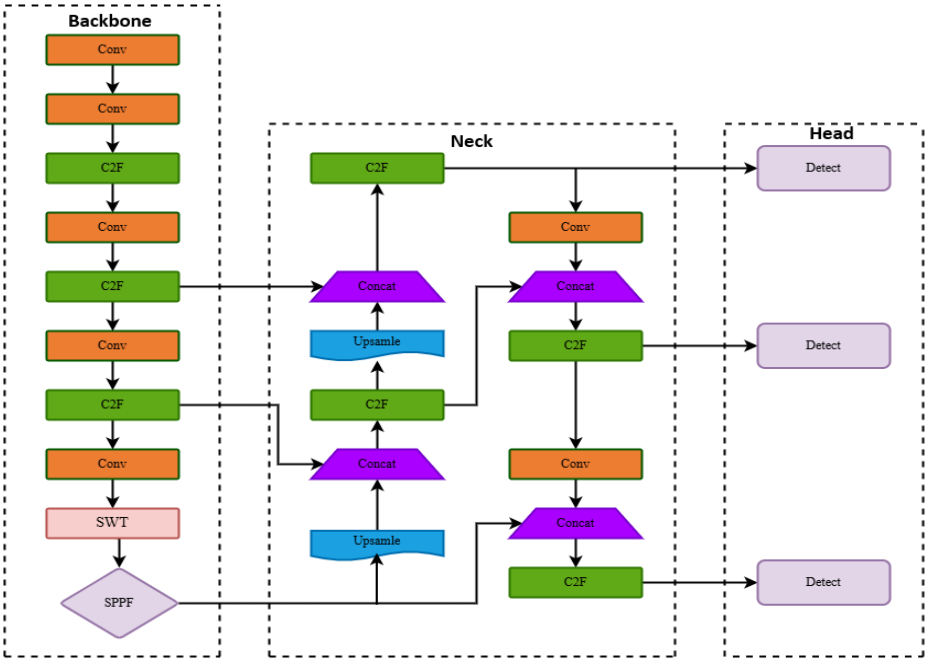

Рис. 4. Архитектура гибридной модели YOLOv8s + SWT
Fig. 4. Architecture of the hybrid model YOLOv8s + SWT

4. Результаты исследования моделей СНС и их обсуждение

После обучения и валидации разработанных и базовых моделей СНС на обучающей и валидационной выборках из датасета [7] были проведены исследования этих моделей с использованием его тестовой выборки. Исследования проводились по точности детектирования ЛО и по скорости вычисления моделей. В табл. 1 приведены результаты по точности детектирования по метрике $AP_{0,5}$ объектов каждого класса и по метрике $mAP_{0,5}$ объектов всех классов для обученных базовой модели YOLOv5s и гибридных моделей на ее основе.

Таблица 1

Результаты исследования базовой модели YOLOv5s и гибридных моделей на ее основе по точности детектирования ЛО

Класс	$AP_{0,5}$, $mAP_{0,5}$		
	YOLOv5s	YOLOv5s + SWT	YOLOv5s + SEA
БПЛА самолетного типа	0,935	0,947	0,941
Птица	0,930	0,907	0,920
БПЛА вертолетного типа	0,961	0,967	0,956
Все классы	0,942	0,940	0,939

Анализ результатов, представленных в табл. 1, позволяет считать, что все модели демонстрируют довольно высокую точность детектирования ЛО по метрикам $AP_{0,5}$ и $mAP_{0,5}$, превышающую их пороговое значение 0,9.

Гибридная модель YOLOv5s + SWT показывает лучшие результаты для объектов классов «БПЛА самолетного типа» и «БПЛА вертолетного типа» по сравнению с результатами для объектов этих классов, полученных с помощью гибридной модели YOLOv5s + SEA и базовой модели YOLOv5s. Это указывает на эффективность использования SWIN-Трансформера для обнаружения и классификации малоразмерных объектов данных классов. Однако для объектов класса «Птица» обе гибридные модели

дают точность детектирования ниже, чем у базовой модели YOLOv5s. Более того, результаты по точности детектирования объектов по метрике $mAP_{0,5}$, получаемые с помощью этих моделей, несколько ниже, чем у базовой модели YOLOv5s.

В табл. 2 приведены результаты по точности детектирования по метрике $AP_{0,5}$ ЛО каждого класса и по метрике $mAP_{0,5}$ объектов всех классов для обученных базовой модели YOLOv8s и гибридных моделей на ее основе. Анализ этих результатов позволяет сделать следующие выводы.

Таблица 2

Результаты исследования базовой модели YOLOv8s и гибридных моделей на ее основе по точности детектирования ЛО

Класс	$AP_{0,5}$, $mAP_{0,5}$		
	YOLOv8s	YOLOv8s + SWT	YOLOv8s + SEA
БПЛА самолетного типа	0,937	0,925	0,942
Птица	0,935	0,925	0,933
БПЛА вертолетного типа	0,966	0,956	0,971
Все классы	0,946	0,935	0,948

Базовая модель и гибридные модели демонстрируют высокую точность детектирования объектов по метрикам $AP_{0,5}$ и $mAP_{0,5}$ в диапазоне значений от 0,925 до 0,971 (значительное превышение порогового значения 0,9 этих метрик). Отметим, что все результаты в табл. 2 лучше, чем соответствующие результаты для модели YOLOv5s и гибридных моделей на ее основе, приведенные в табл. 1.

Гибридная модель YOLOv8s + SEA демонстрирует лучшие результаты по точности детектирования объектов по метрикам $AP_{0,5}$ и $mAP_{0,5}$ по сравнению с моделями YOLOv8s + SWT и YOLOv8s, а для объектов класса «БПЛА вертолетного типа» – наилучший результат (значение метрики $AP_{0,5} = 0,971$). Однако эта модель незначительно уступает базовой модели YOLOv8s по точности детектирования в случае объектов класса «Птица». Результаты для всех классов объектов в случае гибридной модели YOLOv8s + SWT хуже, чем у базовой модели YOLOv8s. Это указывает на то, что гипотеза о повышении точности детектирования малоразмерных объектов путем замены четвертого блока C2F в структуре Backbone модели YOLOv8s на модуль SWT не подтвердилась.

Из анализа результатов исследования, приведенных в табл. 1 и 2, следует вывод о том, что гибридная модель YOLOv8s + SEA является наиболее предпочтительным вариантом при создании СКЗ реального времени с повышенным требованием к точности детектирования малоразмерных ЛО.

В табл. 3 приведены результаты исследований на тестовой выборке обученных базовых моделей YOLOv5s и YOLOv8s и гибридных моделей на их основе в виде усредненной скорости вычисления этих моделей. Результаты показаны в виде значений метрик времени вычисления Inference и NMS (в миллисекундах) и FPS для RGB – изображения размером 416 × 416 пикселей.

Таблица 3

Результаты исследования базовых моделей YOLOv5s и YOLOv8s и гибридных моделей по усредненной скорости вычисления

Модель СНС	Inference, мс	NMS, мс	FPS
YOLOv5s	3,2	2,4	179
YOLOv5s + SWT	3,0	2,4	185
YOLOv5s + SEA	3,0	2,4	185
YOLOv8s	4,6	1,8	156
YOLOv8s + SWT	4,9	1,8	149
YOLOv8s + SEA	4,5	2,3	147

На основе представленных в табл. 3 результатов можно сделать следующие выводы. У модели YOLOv8s время вычисления одного изображения Inference значительно выше (4,6 мс) по сравнению со значением этого показателя у модели YOLOv5s (3,2 мс), что дает более низкое значение FPS, равное 156, по сравнению со значением 179 у модели YOLOv5s. То есть модель YOLOv5s демонстрирует лучшую производительность, чем модель YOLOv8s.

Гибридные модели YOLOv5s + SWT и YOLOv5s + SEA показывают максимальное значение FPS, равное 185, что делает данные модели наиболее эффективными среди всех исследуемых здесь моделей.

Гибридная модель YOLOv8s + SWT имеет время Inference 4,9 мс и значение FPS, равное 149, а модель YOLOv8s + SEA показывает соответственно 4,5 мс и FPS, равное 147. Это указывает на снижение усредненной скорости их вычислений по сравнению с базовой моделью YOLOv8s.

Модель YOLOv5s и гибридные модели на ее основе демонстрируют лучшие значения усредненной скорости вычислений по сравнению с моделью YOLOv8s и гибридными моделями на ее основе. Это позволяет считать их более предпочтительными для решения задач, требующих высокой производительности СКЗ реального времени.

Заключение

Анализ ряда исследований по точности объектного детектирования ЛО на изображениях показал, что существует актуальная проблема повышения точности детектирования таких объектов малых размеров. Для ее решения предлагается разрабатывать и исследовать новые модели СНС с механизмом внимания. В качестве исходных для разработки таких новых моделей СНС выбраны две базовые модели класса YOLO: YOLOv5s и YOLOv8s. На их основе создано четыре гибридных модели СНС с использованием модуля SWT и модуля SEA, реализующих два варианта механизма внимания.

По результатам обучения, валидации и исследования базовых и гибридных моделей на датасете с размеченными изображениями малоразмерных ЛО трех классов выявлено, что обе базовые модели и гибридные модели на их основе по точности детектирования ЛО по метрикам $AP_{0,5}$ и $mAP_{0,5}$ превышают весьма высокий заданный порог 0,9 и могут использоваться в качестве основы в СКЗ реального времени для детектирования малоразмерных объектов. Однако гибридная модель YOLOv8s + SEA является наиболее предпочтительным вариантом при создании СКЗ реального времени с повышенным требованием к точности детектирования малоразмерных ЛО.

Результаты исследования моделей по скорости вычислений показали, что все они позволяют превысить пороговое значение метрики FPS, равное 25, и поэтому могут использоваться в составе СКЗ реального времени. Модель YOLOv5s и гибридные модели на ее основе по скорости вычислений эффективнее, чем модель YOLOv8s и гибридные модели на ее основе. Это позволяет считать их более предпочтительными для решения задач, требующих высокой производительности от СКЗ реального времени. Однако выбор из них конкретной модели зависит от выдвигаемых требований к точности детектирования малоразмерных ЛО каждого класса и от масштаба реального времени.

Список источников

1. Tan M., Pang R., Le Q.V. EfficientDet: Scalable and Efficient Object Detection // CVPR. 2020. Art. 09070. URL: https://arxiv.org/abs/1911.09070 (accessed: 10.04.2025).
2. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. М. : ДМК-Пресс, 2018. 652 с.
3. Zoev I.V., Markov N.G., Ryzhova S.E. Intelligent computer vision system for unmanned aerial vehicles for monitoring technological objects of oil and gas industry // Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering. 2019. V. 330 (11). P. 34–49. doi: 10.18799/24131830/2019/11/2346
4. Alzubaidi L., Zhang J., Humaidi A.J., Al-Dujaili A., Duan Y., Al-Shamma O., Santamaría J., Fadhel M.A., Al-Amidie M., Farhan L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions // Journal of Big Data. 2021. V. 8 (53). P. 1–77. doi: 10.1186/s40537-021-00444-8
5. Небаба С.Г., Марков Н.Г. Сверточные нейронные сети семейства YOLO для мобильных систем компьютерного зрения // Компьютерные исследования и моделирование. 2024. Т. 16, № 3. С. 615–631. doi: 10.20537/2076-7633-2024-16-3-615-631
6. Wu S., Lu X., Guo C., Guo H. Accurate UAV Small Object Detection Based on HRFPN and EfficentVMamba // Sensors. 2024. V. 24 (5). Art. 4966. doi: 10.3390/s24154966
7. Клековкин В.А., Марков Н.Г., Небаба С.Г. Обнаружение и классификация малоразмерных летающих объектов на изображениях с использованием сверточных нейронных сетей семейства YOLOv5 // Доклады ТУСУР. 2024. Т. 27, № 4. С. 103–110. doi: 10.21293/1818-0442-2024-27-4-103-110
8. Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Zitnick C.L. Microsoft COCO: Common objects in context // Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014 : proc. Springer International Publishing, 2014. Pt. V 13. P. 740–755. doi: 10.48550/arXiv.1405.0312

9. Bochkovskiy A., Wang C.Y., Liao H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection // arXiv. 2020. Art. 10934v1. P. 1–17. doi: 10.48550/arXiv.2004.10934
10. Olorunshola O.E., Irhebhude M.E., Evwiekpaefe A.E. A Comparative Study of YOLOv5 and YOLOv7 Object Detection Algorithms // Journal of Computing and Social Informatics. 2023. V. 2. P. 1–12. doi: 10.33736/jcsi.5070.2023
11. Филичкин С.А., Вологдин С.В. Сравнение эффективности алгоритмов YOLOv5 и YOLOv8 для обнаружения средств индивидуальной защиты человека // Интеллектуальные системы в производстве. 2023. Т. 21, № 3. С. 124–131.
12. Vaswani A. et al. Attention is all you need // 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. V. 30. URL: https://arxiv.org/abs/1706.03762 (accessed: 10.04.2025).
13. Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale // arXiv. 2020. Art. 11929. doi: arXiv:2010.11929. 2020
14. Han K. et al. A survey on visual transformer // // arXiv. 2012. Art. 12556. doi: 10.48550/arXiv.2012.12556
15. Li J., Zhang J., Shao Y., Liu F. SRE-YOLOv8: An Improved UAV Object Detection Model Utilizing Swin Transformer and RE-FPN // Sensors. 2024. V. 24 (12). Art. 3918. doi: 10.3390/s24123918
16. Hu J., Shen L., Sun G. Squeeze-and-excitation networks // IEEE / CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 7132–7141. doi: 10.1109/CVPR.2018.00745
17. Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. Swin transformer: Hierarchical vision transformer using shifted windows // IEEE / CVF International Conference on Computer Vision (ICCV). 2021. P. 9992–10002. doi: 10.48550/arXiv.2103.14030
18. Yang J. et al. Focal modulation networks // Advances in Neural Information Processing Systems 35 (NeurIPS 2022). 2022. V. 35. P. 4203–4217.

References

1. Tan, M., Pang, R. & Le, Q.V. (2020) EfficientDet: Scalable and Efficient Object Detection. *CVPR 2020.* Art. 09070. [Online] Available from: https://arxiv.org/abs/1911.09070 (Accessed: 10th April 2025).
2. Goodfellow, I., Bengio, Y. & Courville, A. (2018) *Glubokoe obuchenie* [Deep Learning]. Translated from English. Moscow: DMK-Press.
3. Zoev, I.V., Markov, N.G. & Ryzhova, S.E. (2019) Intelligent computer vision system for unmanned aerial vehicles for monitoring technological objects of oil and gas industry. *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering.* 330(11). pp. 34–49. DOI: 10.18799/24131830/2019/11/2346
4. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. & Farhan, L. (2021) Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*. 8(53). pp. 1–77. DOI: 10.1186/s40537-021-00444-8
5. Nebaba, S.G. & Markov, N.G. (2024) Convolutional neural networks of the YOLO family for mobile computer vision systems. *Komp'yuternye issledovaniya i modelirovanie*. 16(3). pp. 615–631. DOI: 10.20537/2076-7633-2024-16-3-615-631
6. Wu, S., Lu, X., Guo, C. & Guo, H. (2024) Accurate UAV Small Object Detection Based on HRFPN and EfficentVMamba. *Sensors.* 24(5). Art. 4966. DOI: 10.3390/s24154966
7. Klekovkin, V.A., Markov, N.G. & Nebaba, S.G. (2024) Detection and classification of small flying objects in images using convolutional neural networks of the YOLOv5 family. *Doklady TUSUR*. 27(4). pp. 103–110. DOI: 10.21293/1818-0442-2024-27-4-103-110
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D. & Zitnick, C.L. (2014) Microsoft COCO: Common objects in context. *Computer Vision–ECCV 2014.* 13th European Conference, Zurich, Switzerland, September 6-12. Proceedings, Part V 13. Springer International Publishing. pp. 740–755. DOI: 10.48550/arXiv.1405.0312
9. Bochkovskiy, A., Wang, C.Y. & Liao, H.Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. *Journal arXiv, preprint arXiv.* 10934v1. pp. 1–17. DOI: 10.48550/arXiv.2004.10934
10. Olorunshola, O.E., Irhebhude, M.E. & Evwiekpaefe, A.E. (2023) A Comparative Study of YOLOv5 and YOLOv7 Object Detection Algorithms. *Journal of Computing and Social Informatics*. 2. pp. 1–12. DOI: 10.33736/jcsi.5070.2023
11. Filichkin, C.A. & Vologdin, S.V. (2023) Comparison of the effectiveness of YOLOv5 and YOLOv8 algorithms for detecting personal protective equipment. *Intellektual'nye sistemy v proizvodstve*. 21(3). pp. 124–131.
12. Vaswani, A. et al. (2017) Attention is all you need. *Advances in Neural Information Processing Systems*. 30. [Online] Available from: https://arxiv.org/abs/1706.03762
13. Dosovitskiy, A. et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*
14. Han, K. et al. (2020) A survey on visual transformer. *arXiv preprint arXiv:2012.12556*
15. Li, J., Zhang, J., Shao, Y. & Liu, F. (2024) SRE-YOLOv8: An Improved UAV Object Detection Model Utilizing Swin Transformer and RE-FPN. *Sensors*. 24(12). Art. 3918. DOI: 10.3390/s24123918
16. Hu, J., Shen, L. & Sun, G. (2018) Squeeze-and-excitation networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9992–10002. DOI: 10.48550/arXiv.2103.14030
18. Yang J. et al. (2022) Focal modulation networks. *Advances in Neural Information Processing Systems*. 35. pp. 4203–4217.

***Информация об авторах*:**
Клековкин Вадим Александрович – аспирант отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета (Томск, Россия). E-mail: vak37@tpu.ru
Марков Николай Григорьевич – доктор технических наук, профессор отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета (Томск, Россия). E-mail: markovng@tpu.ru
Небаба Степан Геннадьевич – кандидат технических наук, доцент отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета (Томск, Россия). E-mail: stepanlfx@tpu.ru

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

***Information about the authors*:**
Klekovkin Vadim A. (Post-Graduate Student, National Research Tomsk Polytechnic University, Tomsk, Russian Federation). E-mail: vak37@tpu.ru
Markov Nikolay G. (Doctor of Technical Sciences, Professor, National Research Tomsk Polytechnic University, Tomsk, Russian Federation). E-mail: markovng@tpu.ru
Nebaba Stepan G. (Candidate of Technical Sciences, Associate Professor, National Research Tomsk Polytechnic University, Tomsk, Russian Federation). E-mail: stepanlfx@tpu.ru

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Поступила в редакцию 05.05.2025; принята к публикации 02.09.2025

Received 05.05.2025; accepted for publication 02.09.2025