Исследование интерпретируемости моделей детектирования пешеходов на основе векторов активации концептов (TCAV): межархитектурный эмпирический анализ

Лю Чэньфэн

***Аннотация*—В настоящей работе, направленной на решение проблемы «черного ящика» в моделях детектирования пешеходов в высокорисковых сценариях применения, таких как автономное вождение и интеллектуальные системы безопасности, внедряется метод векторов активации концептов (TCAV) для проведения эмпирического исследования кросс-уровневых представлений признаков в трех репрезентативных архитектурах: Faster R-CNN, YOLOv11 и RT-DETR. Посредством разработки метода автоматической обрезки на основе анатомических пропорций человеческого тела формируются наборы семантических концептов (голова, туловище и ноги), что позволяет систематически количественно оценивать пути эволюции семантических признаков на различных глубинах нейронных сетей. Полученные результаты показывают, что сверточная архитектура (Faster R-CNN) демонстрирует прогрессивную схему семантического моделирования — от локальных текстур к глобальной структуре, сохраняя стабильную способность к семантическому расцеплению в глубоком пространстве признаков. В детекторе реального времени YOLOv11 наблюдается явление «семантического запаздывания», при котором структурированная семантика человеческого тела формируется преимущественно в сети слияния признаков, а не в магистральной сети. Сквозная архитектура Transformer RT-DETR, несмотря на наличие высокоточной семантической разделимости, характеризуется разреженностью градиентов, вызванной насыщением решений, что ограничивает эффективность методов линейной интерпретации. Результаты исследования раскрывают границы применимости методов интерпретации в условиях различных парадигм принятия решений и служат количественной основой для построения высокопрозрачных систем детектирования пешеходов.**

***Ключевые слова*—архитектуры глубокого обучения, векторы активации концептов (TCAV), детектирование пешеходов, интерпретируемость моделей.**

I. Введение

Детектирование пешеходов, являясь ключевой задачей восприятия в системах автономного вождения и интеллектуальной безопасности, предъявляет крайне высокие требования к надежности моделей и прозрачности процесса принятия решений. Несмотря на то, что современные модели глубокого обучения демонстрируют исключительную точность детектирования в сложных условиях, их внутренние механизмы функционирования повсеместно характеризуются выраженным эффектом «черного ящика». При работе в сценариях принятия решений с высоким уровнем риска, таких как обеспечение безопасности дорожного движения, использование исключительно показателей точности не позволяет в полной мере подтвердить, опирается ли модель на подлинное понимание семантики анатомического строения человека или же принимает решения на основе локальных текстур и статистических смещений в данных. Подобная ограниченная интерпретируемость в значительной степени ограничивает дальнейшее внедрение моделей глубокого детектирования в критически важных областях[1].

Для решения проблемы непрозрачности моделей метод векторов активации концептов (Testing with Concept Activation Vectors, TCAV) предлагает эффективный подход к количественной интерпретации, позволяющий отображать внутреннее векторное пространство нейронной сети в понятные человеку высокоуровневые семантические концепты[2]. Несмотря на значительные успехи TCAV в задачах классификации изображений, применимость и интерпретационные возможности данного метода в задачах детектирования объектов, характеризующихся локализацией множественных целей и сложным фоном, требуют дополнительного глубокого подтверждения. То, каким образом в различных архитектурах детектирования формируются концептуальные представления на уровне отдельных компонентов и как трансформируется семантика в процессе извлечения признаков (feature extraction), остается важным направлением исследований в современной области интерпретируемого искусственного интеллекта[3].

В настоящей работе метод TCAV систематически внедряется в область детектирования пешеходов для проведения межуровневого эмпирического анализа трех репрезентативных архитектур: Faster R-CNN, YOLOv11 и RT-DETR. На основе сравнительного анализа функционирования различных архитектур при обработке детальной семантики (головы, туловища и ног) в рамках

Статья получена 3 марта 2026.

Лю Чэньфэн, Дальневосточный федеральный университет, Владивосток, Россия(e-mail: liuchenfeng1999@163.com).

проведенного исследования подтверждается устойчивость метода TCAV в задачах детектирования объектов. Полученные результаты показывают наличие определенных различий в подходах к семантическому моделированию между сверточными нейронными сетями и детекторами на базе архитектуры Transformer, что служит количественной экспериментальной основой для разработки заслуживающих доверия систем детектирования пешеходов.

II. Адаптация и реализация метода TCAV в задачах детектирования пешеходов

A. Выбор набора данных и построение семантических концептов

В настоящем исследовании в качестве экспериментальной базы используется набор данных CityPersons. Данный датасет, построенный на основе Cityscapes, содержит детализированную разметку, ориентированную специально на задачи детектирования пешеходов, и охватывает разнообразные городские сценарии, а также сложные случаи перекрытия объектов[4]. Наличие высококачественной разметки ограничивающих рамок (bounding boxes) послужило надежной основой для алгоритма автоматической обрезки, разработанного в рамках данной работы. Для объективной оценки восприятия моделью структуры человеческого тела в ходе эксперимента истинно положительные рамки детектирования (true positive) были разделены на четыре семантических концепта с использованием метода вертикального пропорционального сегментирования: голова (верхние 0%–25%), туловище (средние 25%–65%), ноги (нижние 65%–100%) и пешеход в полный рост. В то же время, для обучения линейных классификаторов и формирования базовых направлений был сформирован набор концептов случайного фона.

Таблица I - Сводная таблица объемов выборок наборов концептов TCAV для детектирования пешеходов

Категория концепта	**Количество образцов**	**Источник и логика построения**
Full Body	1127	Целостная обрезка истинно положительных рамок детектирования
Head	1127	Верхняя область истинно положительной рамки (от 0% до 25%)
Torso	1127	Средняя область истинно положительной рамки (от 25% до 65%)
Legs	1127	Нижняя область истинно положительной рамки (от 65% до 100%)
Random Background	826	Случайная выборка фона из CityPersons (IoU < 0,1)

Данная совокупность была получена путем случайной выборки из областей изображения, не содержащих пешеходов; при этом накладывалось строгое ограничение на пересечение над объединением (Intersection over Union, IoU) между рамкой выборки и любым известным объектом на уровне менее 0,1 для обеспечения чистоты контрольной группы[5]. В таблице I представлены объемы выборок для наборов концептов.

B. Математическая логика алгоритма TCAV

Суть метода TCAV заключается в отображении понятных человеку концептов на пространство признаков промежуточных слоев нейронной сети. Данный процесс включает три основных математических этапа:

(1) Извлечение вектора активации концепта (CAV): Пусть $f_l : R^n \to R^m$ - функция отображения l -го слоя модели. Для заданного семантического концепта C сначала формируются набор образцов концепта P_C и набор образцов случайного фона P_{rand} Путем обучения линейного бинарного классификатора (например, логистической регрессии) в пространстве признаков определяется разделяющая гиперплоскость:

$$w \cdot f_l(x) + b = 0 , \quad (1)$$

При этом нормальный вектор гиперплоскости определяется как вектор активации концепта (CAV) и обозначается как $v_C^l \in R^m$. Данный вектор представляет направление наиболее быстрого роста семантического концепта C в пространстве признаков l-го слоя.

(2) Расчет чувствительности к концепту: Для оценки влияния конкретного концепта на решение модели необходимо вычислить производную по направлению целевой функции относительно активаций промежуточного слоя. Пусть непрерывная функция выхода модели для категории пешеходов обозначается как $h_k(x)$. Для единичного входного образца x его чувствительность к концепту определяется следующим образом:

$$S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_k(f_l(x) + \epsilon v_C^l) - h_k(f_l(x))}{\epsilon} = \nabla h_k(f_l(x)) \cdot v_C^l , \quad (2)$$

Данная формула количественно определяет скорость изменения результата классификации пешеходов при незначительном возмущении признаков в направлении концепта C.

(3) Статистический вывод оценки TCAV: Для всего тестового набора Q общая значимость концепта C для определения моделью категории k количественно выражается через оценку TCAV:

$$TCAV_{Q,C,k,l} = \frac{|\{x \in Q: S_{C,k,l}(x) > 0\}|}{|Q|} , \quad (3)$$

Данный показатель отражает долю образцов, для которых чувствительность к концепту является положительной. Если TCAV > 0,5, это указывает на то, что рассматриваемый концепт вносит положительный вклад в итоговое решение модели [2].

C. Анализ робастности и статистической значимости экспериментов

Ввиду того, что процесс обучения линейного классификатора подвержен влиянию случайности при выборе негативных образцов, результаты единичного обучения CAV могут носить характер статистической случайности. В настоящей работе был применен

протокол многократного случайного разбиения: путем проведения 10 независимых выборок подмножеств из набора случайного фона были обучены 10 различных векторов $v^l_{C,i}$, что позволило сформировать распределение оценок TCAV[6].

Проверка статистической значимости осуществляется с помощью двойного механизма. Биномиальный критерий используется для подтверждения того, что частота оценок TCAV значимо превышает случайный уровень (0,5); t-критерий для одной выборки применяется для оценки статистической стабильности среднего значения по результатам 10 экспериментов. Все аналитические процедуры строго соответствуют порогу значимости $p < 0,05$, что гарантирует интерпретацию выявленной семантики компонентов человеческого тела как устойчивого внутреннего основания для принятия решений моделью, а не как результат случайного шума. В рамках конкретных статистических настроек нулевая гипотеза t-критерия для одной выборки предполагает, что математическое ожидание оценки TCAV равно случайному базовому уровню (0,5); это позволяет проверить, является ли чувствительность к концепту статистически значимо и стабильно выше случайного направления. Принимая во внимание, что основное внимание в настоящей работе уделяется общему сопоставлению тенденций семантической эволюции на различных уровнях и в разных архитектурах, а не подтверждению единичной гипотезы, в данном исследовании не вводились дополнительные стратегии коррекции множественных сравнений для сценариев с множеством слоев и концептов. Соответствующие результаты тестирования используются преимущественно в качестве вспомогательного инструмента для оценки наличия устойчивого интерпретируемого сигнала в семантических концептах.

III. Условия проведения экспериментов и оценка базовых показателей

A. Экспериментальная среда и аппаратная конфигурация

Для обеспечения воспроизводимости моделей детектирования пешеходов и соответствующих экспериментов по их интерпретируемости все испытания проводились на единой вычислительной платформе. В качестве операционной системы использовалась Windows 11 (версия 23H2). Аппаратная часть основывалась на процессоре Intel Core i7-12700H 12-го поколения в сочетании с оперативной памятью DDR5 объемом 40 ГБ. Задачи графической обработки выполнялись с использованием графического процессора NVIDIA GeForce RTX 3060 Laptop GPU с 6 ГБ выделенной видеопамяти. Что касается программного стека, в качестве языка программирования использовался Python 3.11.5, версия фреймворка глубокого обучения — PyTorch 2.6.0+cu126 с библиотеками ускорения CUDA 12.6 и CuDNN 9.0.5.

B. Исследуемые модели и конфигурация их параметров

В настоящем исследовании в качестве объектов анализа были выбраны три типа архитектур, имеющих принципиальные различия в механизмах детектирования, для оценки универсальности метода TCAV[7]:

(1) Faster R-CNN: реализует классическую двухстадийную парадигму детектирования, использующую ResNet-50 в качестве магистральной сети (backbone) для извлечения признаков и интегрирующую сеть пирамид признаков (FPN) для усиления способности к многомасштабному восприятию.

(2) YOLOv11n: представляет одностадийную архитектуру детектирования в реальном времени; в данном эксперименте выбрана облегченная версия «n» с целью проверки способности модели к моделированию семантических концептов при ограниченном количестве параметров.

(3) RT-DETR-L: современный сквозной (end-to-end) детектор на базе архитектуры Transformer, использующий механизм глобального внимания для обработки взаимосвязей между объектами, что позволяет исследовать принципиальные различия в механизмах извлечения семантики по сравнению со сверточными нейронными сетями.

Для всех трех моделей использовались веса, предварительно обученные на наборе данных COCO. Для обеспечения целенаправленности интерпретационного анализа на этапе логического вывода параметры моделей оставались замороженными; извлекались только значения активаций промежуточных слоев для вычисления производных по направлению. Подобный кросс-архитектурный выбор охватывает основные технологические направления: от выделения областей интереса и регрессионного анализа до механизмов глобального внимания.

C. Определение оценочных показателей

Для количественной оценки базовой производительности исследуемых моделей на наборе данных CityPersons используются метрики, общепризнанные в области детектирования пешеходов[8]. Определения соответствующих показателей представлены ниже:

Полнота (Recall): измеряет долю успешно обнаруженных объектов среди всех истинных пешеходов, что отражает способность модели к выявлению максимального количества целей.

Точность (Precision): измеряет долю истинных пешеходов среди всех объектов, классифицированных моделью как пешеходы, что характеризует достоверность предсказаний.

F1-мера (F1-score): определяется как среднее гармоническое точности и полноты; используется для комплексной оценки робастности модели.

Средняя точность (Average Precision, AP): оценивает совокупную эффективность детектирования при различных порогах уверенности путем расчета площади под кривой точности-полноты (Precision-Recall curve).

Коэффициент пропусков (Miss Rate): определяется как отношение количества не выявленных истинных пешеходов к общему числу пешеходов в выборке; данный показатель является ключевым критерием безопасности в задачах детектирования пешеходов.

D. Обоснование корректировки пороговых значений и анализ базовых результатов

Перед проведением интерпретационного анализа по методу TCAV необходимо обеспечить генерацию моделью достаточного объема истинно положительных (True Positive, TP) образцов для извлечения семантических признаков. Предварительные эксперименты показали, что модель YOLOv11n при стандартном пороге достоверности 0,5 демонстрирует исключительно высокую точность (Precision), однако значение полноты (Recall) составляет лишь 0,216. Это непосредственно приводит к недостаточному количеству эффективных фрагментов изображений, пригодных для обратного распространения градиента. С целью фиксации более полных состояний семантической активации без изменения весов модели в настоящем исследовании была введена конфигурация с низким порогом (0,05). Для обеспечения семантической чистоты выборки при низком пороге уверенности (0,05) все извлеченные объекты проходили валидацию на соответствие Ground Truth (IoU≥0,5). Это исключает попадание фонового шума в набор концептов и гарантирует, что TCAV отражает внутреннюю логику модели по отношению к реальным объектам, а не случайным артефактам детектирования. Следует подчеркнуть, что корректировка порога достоверности влияет исключительно на диапазон фильтрации кандидатных результатов детектирования и направлена на расширение пула истинно положительных образцов, доступных для анализа. Процесс вычисления по методу TCAV основывается только на значениях активаций промежуточных слоев, соответствующих истинно положительным рамкам, отобранным с учетом ограничений по IoU и категориальной согласованности; ошибочно обнаруженные объекты (False Positives) не включаются в процесс обучения векторов активации концептов (CAV) или в статистический анализ чувствительности к концептам. Таким образом, снижение порога достоверности не приводит к статистическому завышению значимости семантических концептов, а служит лишь инструментом для преодоления ограничений интерпретационного анализа, обусловленных недостаточным объемом выборки при высоких пороговых значениях.

Таблица II - Сводная таблица базовых показателей эффективности детектирования исследуемых моделей

Архитектура модели	**Порог достоверности**	**Recall**	**Precision**	**F1-score**	**AP**	**Miss Rate**
Faster R-CNN	0,5	0,577	0,530	0,552	0,442	0,423
RT-DETR-L	0,5	0,437	0,690	0,535	0,359	0,563
YOLOv11n (High)	0,5	0,216	0,801	0,340	0,189	0,784
YOLOv11n (Low)	0,05	0,536	0,337	0,414	0,376	0,464

Как показано в таблице II, после корректировки значение полноты для YOLOv11n (Low) увеличилось до 0,536. Несмотря на наличие в данной конфигурации большего количества ложноположительных срабатываний с низкой степенью уверенности, это позволило существенно расширить набор истинно положительных (TP) образцов. В результате стало возможным полное извлечение семантических откликов внутренних слоев модели на компоненты тела пешехода (голова, туловище, ноги) и проведение последующего анализа статистической значимости.

IV. Результаты экспериментов по интерпретируемости Faster R-CNN с использованием метода TCAV

A. Уровни извлечения признаков и структура магистральной сети

В архитектуре Faster R-CNN в качестве магистрального экстрактора признаков (backbone) используется сеть ResNet-50, интегрированная с сетью пирамид признаков (FPN). Для фиксации характеристик эволюции семантических концептов в глубоком пространстве признаков (feature space) в настоящей работе используется оператор извлечения промежуточных слоев (IntermediateLayerGetter) для получения значений активации с различных уровней глубины магистральной сети[9]. Конкретный выбор точек зондирования и их физическая структура представлены ниже:

Слой 2: данный слой состоит из 4 последовательно соединенных бутылочных остаточных блоков (bottleneck units); в каждый модуль интегрированы замороженные слои нормализации для сохранения распределения признаков предобученных весов. Количество каналов выходных карт признаков на данном уровне составляет 512, что позволяет сохранять высокое пространственное разрешение; слой преимущественно отвечает за обработку первичных геометрических форм и текстурных признаков.

Слой 3: являясь частью магистральной сети с наибольшим количеством параметров, данный слой включает 6 бутылочных остаточных блоков, а количество выходных каналов достигает 1024. Глубокие признаки данного уровня проходят через многослойные нелинейные преобразования, что позволяет объединять базовые границы в более семантически различимые структурные единицы компонентов.

Слой 4: данный слой содержит 3 бутылочных остаточных блока, количество выходных каналов составляет 2048. Выступая в качестве конечного выхода магистральной сети, извлекаемые им признаки обладают высокой степенью абстракции; слой отвечает за синтез разрозненной информации о компонентах в глобальное представление объекта и непосредственно участвует в последующем извлечении регионов-кандидатов (region proposal) и принятии классификационных решений.

B. Анализ результатов экспериментов

В рамках архитектуры Faster R-CNN статистические результаты оценок TCAV для головы, туловища, ног и пешехода в полный рост представлены в таблице III;

значения приведены в формате Mean ± Std. Все экспериментальные данные прошли проверку с помощью биномиального критерия и t-критерия для одной выборки ($p < 0,05$), что подтверждает робастность выявленной семантики.

ТАБЛИЦА III - СВОДНАЯ ТАБЛИЦА РЕЗУЛЬТАТОВ TCAV ДЛЯ FASTER R-CNN НА РАЗЛИЧНЫХ УРОВНЯХ

Слой	Голова (Head)	Туловище (Torso)	Ноги (Legs)	В полный рост (Full Body)	Вывод о семантической значимости
Слой 2	0,851 ± 0,033	0,875 ± 0,035	0,841 ± 0,040	0,509 ± 0,083	Компоненты значимы, целое не значимо
Слой 3	0,943 ± 0,011	0,944 ± 0,008	0,920 ± 0,012	0,870 ± 0,033	Пик структурной семантики
Слой 4	0,951 ± 0,046	0,928 ± 0,078	0,968 ± 0,037	0,880 ± 0,068	Стабильная семантическая интеграция

На основе представленных количественных показателей можно сделать вывод о высокой применимости метода TCAV в двухстадийных детекторах. Линейные классификаторы, успешно обученные в рамках эксперимента, позволяют выделить независимые концепты компонентов человеческого тела из сложных сверточных признаков. Это показывает, что в условиях текущего эксперимента и с точки зрения линейного зондирования (linear probing) TCAV, на иерархических уровнях признаков Faster R-CNN сформировались семантические подпространства, характеризующиеся высокой степенью соответствия анатомической структуре человека и аппроксимативной линейной разделимостью, что обеспечивает надежную основу для анализа интерпретируемости модели.

C. Обсуждение семантической эволюции

Путем вертикального сопоставления результатов на различных глубинах зондирования можно проследить тенденции эволюции внутренней семантики пешехода в зависимости от глубины сети:

(1) Переход от локального восприятия к глобальному пониманию структуры: На раннем этапе функционирования сети (слой 2), несмотря на значительную активацию концептов локальных компонентов (таких как голова и туловище), оценка концепта «пешеход в полный рост» лишь незначительно превышает случайный базовый уровень. Данное явление указывает на то, что на этой глубине сети процесс дискриминации модели по-прежнему преимущественно опирается на отклики текстурных признаков локальных анатомических частей. С увеличением глубины сети до уровня слоя 3 оценка для пешехода в полный рост демонстрирует резкий скачок до 0,870, что позволяет предположить возможный переход модели на данном этапе от откликов на локальные признаки к более структурированным глобальным репрезентациям.

(2) Робастность семантических признаков: В глубоком пространстве признаков (слой 4) чувствительность модели к различным компонентам человеческого тела сохраняется на высоком уровне (выше 0,9), при этом показатель для концепта «ноги» достигает пикового значения 0,968. Это в определенной степени отражает высокую чувствительность Faster R-CNN к целостным характеристикам структуры человеческого тела на этапе финального определения категории пешехода, что свидетельствует о преобладании структурной информации в процессе принятия решений.

(3) Значимость интерпретационного анализа: В рамках проведенного эксперимента с количественной точки зрения продемонстрирован способ отображения абстрактных тензоров признаков промежуточных слоев в интерпретируемые показатели, имеющие анатомическую релевантность. Это не только подтверждает применимость метода TCAV к двухстадийным моделям детектирования со сложной структурой слияния признаков, но и на основе количественных экспериментальных данных раскрывает прозрачность логики принятия решений Faster R-CNN.

V. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПО ИНТЕРПРЕТИРУЕМОСТИ YOLOV11 С ИСПОЛЬЗОВАНИЕМ МЕТОДА TCAV

A. Выбор уровней зондирования и описание иерархической структуры

YOLOv11 базируется на высокоинтегрированной одностадийной архитектуре детектирования, в которой процессы извлечения признаков и семантического моделирования характеризуются нелинейным распределением в магистральной сети (Backbone) и сети слияния признаков (Neck)[10]. В настоящей работе в качестве точек зондирования были выбраны 8-й, 11-й, 13-й и 15-й слои.

Логика выбора уровней для YOLOv11 имеет существенные отличия от Faster R-CNN, рассмотренной в предыдущей главе. В то время как семантическая эволюция в Faster R-CNN происходит преимущественно в глубоких слоях магистральной сети ResNet, YOLOv11, представляя собой облегченную модель реального времени, ориентирована на сжатие признаков в рамках магистральной сети, тогда как фактическая интеграция семантики и структурное моделирование в большей степени осуществляются на этапе слияния признаков. В этой связи выбор точек зондирования охватывает не только завершение магистральной сети, но и распространяется на ключевые узлы слияния в области Neck. Физическая структура соответствующих уровней представлена ниже:

Слой 8: модуль C3k2, расположенный в завершающей части магистральной сети и состоящий из нескольких остаточных блоков (bottleneck units); отвечает за извлечение высокоабстрактных глобальных признаков.

Слой 11: выполняет операцию повышающей дискретизации (Upsample) методом ближайшего соседа, целью которой является повышение пространственного разрешения высокоуровневой семантики для последующего многомасштабного слияния.

Слой 13: модуль C3k2 в сети слияния признаков, осуществляющий семантическую интеграцию карт

признаков различной глубины посредством сложных нелинейных отображений.

Слой 15: выполняет операцию объединения каналов (Concat), агрегируя признаки повышающей дискретизации и признаки магистральной сети в измерении каналов; данный уровень является ключевым информационным узлом перед этапом финального прогнозирования.

B. Анализ результатов экспериментов

В рамках архитектуры YOLOv11 статистические результаты оценок TCAV для головы, туловища, ног и пешехода в полный рост представлены в таблице IV; значения приведены в формате Mean ± Std. Все экспериментальные данные прошли проверку с помощью биномиального критерия и t-критерия для одной выборки ($p < 0,05$), что подтверждает робастность выявленной семантики.

ТАБЛИЦА IV - СВОДНАЯ ТАБЛИЦА РЕЗУЛЬТАТОВ TCAV ДЛЯ YOLOV11 НА РАЗЛИЧНЫХ УРОВНЯХ

Слой	**Голова (Head)**	**Туловище (Torso)**	**Ноги (Legs)**	**В полный рост (Full Body)**	**Характеристика этапа семантики**
Слой 8	0,569 ± 0,041	0,488 ± 0,059	0,505 ± 0,053	0,533 ± 0,027	Зарождение семантики
Слой 11	0,777 ± 0,012	0,750 ± 0,019	0,753 ± 0,023	0,734 ± 0,027	Полная активация семантики
Слой 13	0,881 ± 0,015	0,865 ± 0,018	0,883 ± 0,021	0,786 ± 0,033	Пик структурной семантики
Слой 15	0,578 ± 0,022	0,657 ± 0,031	0,596 ± 0,029	0,784 ± 0,011	Смещение к целостной дискриминации

Полученные результаты показывают, что в текущих условиях эксперимента метод TCAV сохраняет хорошую применимость в архитектурах детекторов реального времени. Даже в YOLOv11 с меньшим количеством параметров линейные классификаторы позволяют успешно обучать векторы активации концептов, обладающие физиологическим соответствием. В частности, на уровне 13-го слоя оценки значимости для всех концептов компонентов превышают 0,86, что свидетельствует о постепенном формировании в модели относительно стабильных структурированных семантических подпространств в процессе сложного слияния признаков.

C. Обсуждение характеристик семантической эволюции

Полученные данные раскрывают уникальный путь семантического моделирования YOLOv11 в условиях принятия решений при высокоскоростном детектировании:

(1) Эффект ослабления семантики на уровне магистральной сети: на 8-м слое показатели для различных компонентов сохраняются на случайном уровне около 0,5. Это свидетельствует о том, что на данном этапе магистральная сеть YOLOv11 преимущественно выполняет функцию извлечения общих признаков, а ее способность к специализированному моделированию анатомической структуры пешехода остается относительно ограниченной.

(2) Семантический всплеск на этапе слияния признаков: в интервале от 11-го до 13-го слоя оценки для всех концептов человеческого тела значительно возрастают, достигая пика на 13-м слое. Данная характеристика эволюции дополнительно подтверждает, что область Neck (сеть слияния признаков) играет ключевую роль в структурном семантическом моделировании в рамках одностадийных детекторов. Таким образом, окончательное уточненное моделирование семантики компонентов тела человека завершается моделью именно за счет взаимодействия многомасштабных признаков и нелинейного слияния.

(3) Сжатие признаков в сторону глобальной дискриминации: на 15-м слое оценка концептов локальных компонентов (например, головы) заметно снижается до 0,578, в то время как чувствительность к пешеходу в полный рост сохраняется на уровне 0,784. Это отражает тенденцию одностадийных детекторов при приближении к выходному уровню сжимать локальные детали в более глобальные дискриминационные сигналы для достижения баланса между скоростью детектирования и охватом семантической информации.

Резюмируя вышеизложенное, можно предположить, что в рамках проведенного эксперимента YOLOv11 демонстрирует тенденцию к относительному запаздыванию семантической активации. Подобная модель подтверждает, что детекторы реального времени осуществляют эффективное построение семантики человеческого тела на этапе слияния признаков.

VI. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ ПО ИНТЕРПРЕТИРУЕМОСТИ RT-DETR С ИСПОЛЬЗОВАНИЕМ МЕТОДА TCAV

A. Логика выбора уровней зондирования и описание иерархической структуры

RT-DETR, представляя собой сквозную (end-to-end) архитектуру детектирования, состоит из сверточной магистральной сети и декодера на базе механизма внимания[11]. Для исследования эволюции семантики человеческого тела в процессе межструктурной трансформации в настоящей работе были выбраны пять репрезентативных точек зондирования:

Слои 1, 3 и 5 (этап магистральной сети): данные уровни относятся к остаточным блокам (HGBlock) магистральной сети. Выбор этих уровней обусловлен необходимостью наблюдения за тем, способны ли традиционные сверточные структуры осуществлять первичное извлечение признаков компонентов пешехода перед входом в модули Transformer.

Слой 10 (слой проекции признаков): данный уровень представляет собой ключевой сверточный слой 1×1,

отвечающий за проекцию многомерных признаков магистральной сети в пространство фиксированной размерности (256 каналов) для адаптации к последующему модулю внутримасштабного взаимодействия признаков (AIFI). Выбор данного слоя направлен на верификацию состояния сохранности семантики перед ее поступлением в механизмы внимания.

Слой 28 (завершение декодера): данный слой выбран из финального этапа декодера (Decoder). Он является ключевой позицией для генерации итоговых запросов объектов (Object Queries) и формирования предиктивных значений категорий пешеходов. Выбор данного уровня продиктован необходимостью проверки валидности линейной гипотезы TCAV в условиях высоконелинейного пространства запросов.

B. Анализ результатов экспериментов и верификация применимости

В рамках проведенного эксперимента в архитектуре RT-DETR оценки TCAV для различных компонентов человеческого тела, а также результаты проверки статистической значимости (значения *p*), представлены в таблице V. Все экспериментальные данные получены на основе 10 серий рандомизированных контрольных экспериментов; численные показатели приведены в формате Mean ± Std.

Таблица V - Сводная таблица результатов TCAV для RT-DETR на различных уровнях

Уровень	Голова (Head)	Туловище (Torso)	Ноги (Legs)	В полный рост (Full Body)	Вывод о значимости
Слой 1	0,546 ± 0,023 (p=0,001)	0,434 ± 0,020 (p=1,000)	0,436 ± 0,024 (p=1,000)	0,427 ± 0,018 (p=1,000)	Значимость только для головы
Слой 3	0,544 ± 0,031 (p=0,011)	0,383 ± 0,016 (p=1,000)	0,412 ± 0,027 (p=1,000)	0,382 ± 0,014 (p=1,000)	Сохранение зависимости от головы
Слой 5	0,460 ± 0,012 (p=1,000)	0,414 ± 0,012 (p=1,000)	0,421 ± 0,010 (p=1,000)	0,412 ± 0,011 (p=1,000)	Исчезновение значимости
Слой 10	0,469 ± 0,013 (p=1,000)	0,459 ± 0,014 (p=1,000)	0,463 ± 0,011 (p=1,000)	0,446 ± 0,014 (p=1,000)	Семантическое размытие
Слой 28	0,000 ± 0,000 (p=1,000)	0,000 ± 0,000 (p=1,000)	0,000 ± 0,000 (p=1,000)	0,000 ± 0,000 (p=1,000)	Семантическое «непопадание» в декодере

Полученные результаты показывают, что метод TCAV сталкивается с существенными проблемами применимости в архитектуре RT-DETR. В отличие от сверточных нейронных сетей, где семантические оценки демонстрируют непрерывный рост, в RT-DETR слабая семантическая корреляция наблюдается только на ранних уровнях, при этом значимость быстро утрачивается с увеличением глубины.

Для исключения ошибок программной реализации и исследования потенциальных физических причин эффекта затухания активаций (TCAV = 0) на выходе декодера (28-й слой), в настоящей работе дополнительно проведены численные диагностические эксперименты.

Таблица VI - Сводная таблица диагностики разделения семантики и градиентной чувствительности на ключевых уровнях RT-DETR

Уровень зондирования	Семантический концепт	Точность обучения CAV	Оценка TCAV	Кол-во эффективных образцов градиента	Диагностический вывод
Слой 3	Голова (Head)	95,21%	0,4600	50 / 50	Стабильный градиентный поток, линейная гипотеза применима
	Туловище (Torso)	92,84%	0,4400	50 / 50	Первичное разделение семантики, слабая корреляция
Слой 28	Голова (Head)	99,33%	0,0000	0 / 50	Высокая степень разделения семантики, насыщение градиента
	Туловище (Torso)	98,05%	0,0000	0 / 50	Высокая определенность решения, отказ линейного зондирования

Посредством сопоставления различий между сверточным магистральной сетью (3-й слой) и декодером Transformer (28-й слой) в контексте линейной разделимости в пространстве признаков (Feature Space), измеряемой точностью классификации CAV, и градиентной чувствительности на выходе (измеряемой оценкой TCAV), ставится цель определить границы применимости методов интерпретации в различных парадигмах принятия решений. Результаты

диагностических экспериментов представлены в таблице VI.

C. Обсуждение характеристик семантической эволюции

Полученные данные раскрывают принципиальные различия архитектуры сквозного Transformer в моделировании пешеходов:

(1) Механизм приоритета головы («Head-first»): на 1-м и 3-м уровнях концепт головы является единственной семантической единицей, прошедшей статистическую проверку ($p < 0,05$). Это свидетельствует о том, что на ранних стадиях извлечения признаков RT-DETR обладает более высокой чувствительностью к признакам контура головы, которые могут использоваться в качестве важного ориентира для локализации пешеходов; при этом для морфологически сложных туловища и ног в модели не сформировалось устойчивого представления в линейном подпространстве.

(2) Явление семантического разбавления (Semantic Dilution): начиная с 5-го уровня, показатели TCAV для всех компонентов опускаются ниже случайного базового уровня 0,5. Это показывает, что в признаках средних и глубоких слоев RT-DETR высокоабстрактные представления, генерируемые механизмом глобального внимания, в определенной степени ослабляют разделимость семантики на уровне компонентов, в результате чего линейный классификатор не может выделить из них конкретные анатомические концепты.

(3) Эволюция от локального градиентного управления к глобальному дискретному принятию решений: сопоставление экспериментальных данных таблиц 5 и 6 показывает, что в глубоком пространстве признаков (Feature Space) RT-DETR по-прежнему наблюдаются линейно разделимые представления, связанные с семантикой человеческого тела. Точность CAV на 28-м уровне (достигающая 99,33%) значительно превосходит показатели ранних сверточных слоев, что доказывает реализацию внутри модели высокоточного линейного разделения семантики. Явление приближения показателей TCAV на 28-м уровне к нулю может быть связано с насыщением выходных данных при принятии решений или ослаблением градиентного отклика. В сверточных слоях модель следует логике гладкого линейного отображения, что позволяет фиксировать стабильный градиентный поток; в то же время в конце декодера определение категории пешехода на основе запросов объектов (Object Queries) достигает состояния насыщения с высокой степенью определенности, что приводит к стремлению производной по направлению к нулю.

Резюмируя вышеизложенное, RT-DETR демонстрирует паттерн «ранней чувствительности и последующего распада» семантики человеческого тела. Данный феномен указывает на то, что применимость линейного метода TCAV в глубоком пространстве принятия решений сквозных моделей Transformer может иметь определенные ограничения.

VII. Общее обсуждение

A. Анализ эффективности семантического расцепления индуктивных смещений и семантики структуры человеческого тела

Полученные результаты показывают, что сверточные архитектуры, представленные Faster R-CNN, демонстрируют высокую робастность при моделировании семантики компонентов человеческого тела. Суть данного явления заключается в том, что сверточные нейронные сети обладают мощными индуктивными смещениями, такими как локальность (Locality) и инвариантность к сдвигу, что в высокой степени согласуется с когнитивной логикой «уровня компонентов» в анатомии человека. Исходя из тенденции изменения оценок TCAV, можно наблюдать, что модели данного типа постепенно проявляют характеристики семантической декупляции на этапах формирования признаков в средних и глубоких слоях магистральной сети, преобразуя сложные тензоры изображений в линейно разделимое пространство анатомических концептов.

В отличие от этого, явление «семантического разбавления» (Semantic Dilution), наблюдаемое в RT-DETR, указывает на то, что механизм глобального внимания в текущих экспериментальных условиях обладает относительно слабой способностью к линейному извлечению семантики локальных компонентов. Получая более широкое поле зрения (receptive field), модель жертвует способностью к линейному извлечению семантики локальных деталей, что приводит к невозможности формирования стабильных концептуальных представлений на уровне компонентов на средних и поздних этапах функционирования магистральной сети.

B. Анализ механизма семантической компенсации в условиях ограничений реального времени

Эмпирический анализ одностадийных архитектур детектирования в реальном времени выявил специфический феномен «семантического запаздывания». В рамках проведенного эксперимента с YOLOv11 было обнаружено, что магистральная сеть демонстрирует низкую степень активации анатомических концептов, в то время как точки семантического всплеска преимущественно сосредоточены в сети слияния признаков (Neck). Данное явление отражает стратегию компромисса между вычислительной эффективностью и полнотой семантического представления в детекторах реального времени.

В облегченных моделях, таких как YOLOv11, магистральная сеть реализует быстрое снижение размерности признаков посредством глубокого сжатия для удовлетворения требований реального времени, однако это также приводит к потере локальных семантических признаков. Сеть слияния признаков через многомасштабное взаимодействие не только обеспечивает выравнивание пространственной информации, но и выполняет роль семантической «компенсации» и «реконструкции». Можно

предположить, что уровень слияния признаков в моделях детектирования реального времени является не только преобразователем масштаба, но и ключевым механизмом генерации структурированной семантики человеческого тела.

C. Вызовы градиентной разреженности для линейной интерпретируемости

В настоящем исследовании на основе диагностических данных таблицы 6 дополнительно раскрываются ограничения линейной гипотезы TCAV в условиях высоконасыщенного пространства принятия решений. В сверточных архитектурах, таких как Faster R-CNN и YOLOv11, благодаря непрерывности процесса принятия решений метод TCAV позволяет эффективно количественно оценивать вклад отдельных компонентов. Однако на уровне декодера RT-DETR, несмотря на то что точность CAV подтверждает четкое присутствие семантики, явление затухания градиента создает иллюзию «семантического промаха». Это указывает на то, что основания для принятия решений сквозными детекторами могли эволюционировать от локальных активаций на уровне пикселей к дискретным вероятностям, выведенным на основе глобальных связей. Феномен сосуществования «высокой точности CAV и нулевого показателя TCAV» показывает, что в сложных архитектурах между семантической разделимостью и градиентной чувствительностью может существовать определенная степень разобщенности.

Для исключения возможности возникновения артефактов метода были проведены контрольные эксперименты (табл. 7).

Таблица VII - Результаты контрольных экспериментов (Sanity Checks) для векторов TCAV (Mean ± Std)

Модель	**Режим (Mode)**	**Голова (Head)**	**Туловище (Torso)**	**Ноги (Legs)**
YOLOv11	Shuffle	0.517 ± 0.071	0.491 ± 0.060	0.508 ± 0.068
	Random Weights	0.520 ± 0.000	0.547 ± 0.042	0.520 ± 0.029
RT-DETR	Shuffle	0.420 ± 0.224	0.382 ± 0.224	0.380 ± 0.207
	Random Weights	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Faster R-CNN	Shuffle	0.499 ± 0.199	0.471 ± 0.285	0.461 ± 0.206
	Random Weights	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

При рандомизации меток концептов (Shuffle) и весов моделей (Random Weights) оценки TCAV для всех архитектур снизились до случайного уровня (~0.5) или исчезли (0). Это подтверждает, что выявленные семантические зависимости являются результатом обучения на реальных данных, а не следствием случайных шумов.

D. Рекомендации по проектированию робастных и интерпретируемых детекторов пешеходов

Эмпирическое сопоставление различных архитектур дает следующие рекомендации для оптимизации высокопроизводительных и интерпретируемых детекторов пешеходов в будущем:

(1) Усиление признаков под управлением семантики: при проектировании моделей реального времени, таких как YOLOv11, можно рассмотреть возможность введения явных сигналов контроля на уровне компонентов в областях слияния признаков для углубления восприятия структуры человеческого тела и повышения робастности дискриминации в сложных фоновых условиях.

(2) Решение проблемы семантического размытия в Transformer: учитывая выявленную в RT-DETR зависимость от признаков головы на ранних этапах, на стадии декодера можно внедрить механизмы ограничения пространственной локальности для предотвращения избыточной потери ключевой анатомической информации, тем самым улучшая показатели сквозных моделей в аспекте интерпретируемости.

(3) Адаптивная эволюция инструментов интерпретации: принимая во внимание ограничения методов линейной интерпретации в глубоких слоях Transformer, в будущих исследованиях следует изучить нелинейные фреймворки интерпретации, сочетающие карты внимания (Attention Map) с производными по направлению TCAV.

VIII. Заключение

В настоящей работе, направленной на преодоление характера «черного ящика» в моделях детектирования пешеходов, метод векторов активации концептов (TCAV) был внедрен в данную область прикладных исследований. В рамках проведенного эксперимента было реализовано межуровневое эмпирическое исследование трех репрезентативных архитектур: Faster R-CNN, YOLOv11 и RT-DETR. Путем разработки метода автоматической обрезки на основе анатомических пропорций пешехода была проведена количественная оценка различий в механизмах обработки семантики человеческого тела различными глубокими архитектурами. Основные выводы исследования могут быть сформулированы следующим образом:

Во-первых, полученные результаты подтверждают высокую применимость метода TCAV в рамках архитектур сверточных нейронных сетей. Модель Faster R-CNN демонстрирует послойную прогрессивную схему моделирования — от отклика на локальные текстуры к пониманию глобальной структуры; при этом её глубокие признаки обладают высокой робастностью по отношению к анатомическим концептам человека. В то же время YOLOv11 проявляет уникальную характеристику «семантического запаздывания», что доказывает определяющую роль сети слияния признаков (Neck), а не магистральной сети (Backbone), в генерации структурированной семантики в условиях ограничений реального времени.

Во-вторых, исследование раскрывает феномен диссоциации между «семантическим расцеплением» и «выражением решения» в архитектуре сквозного Transformer. В ходе эксперимента было установлено, что, несмотря на достижение в RT-DETR исключительно высокой точности расцепления семантики компонентов на выходе декодера (точность CAV > 97%), традиционное линейное градиентное детектирование демонстрирует состояние разреженного обнуления вследствие насыщения решения, обусловленного механизмом глобального внимания. Данный результат определяет границы применимости инструментов линейной интерпретации в архитектурах Transformer и служит эмпирическим обоснованием для разработки в будущем фреймворков нелинейной интерпретации сквозных моделей.

Резюмируя вышеизложенное, результаты настоящего исследования показывают, что метод TCAV обладает достаточной применимостью в определенных архитектурах детектирования и позволяет с количественной точки зрения описать путь эволюции внутренних семантических признаков в различных моделях. Это создает экспериментальную основу для дальнейших исследований в области проектирования более прозрачных систем детектирования пешеходов.

Библиография

[1] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature machine intelligence, 2019, 1 (5): 206-215.

[2] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]//International conference on machine learning. PMLR, 2018: 2668-2677.

[3] Seyedmomeni F S, Keyvanrad M A. Explaining What Machines See: XAI Strategies in Deep Object Detection Models[J]. arXiv preprint arXiv:2509.01991, 2025.

[4] Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3213-3221.

[5] Chen V, Yang M, Cui W, et al. Best practices for interpretable machine learning in computational biology[J]. Biorxiv, 2022: 2022.10. 28.513978.

[6] Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps[J]. Advances in neural information processing systems, 2018, 31.

[7] Naufaldihanif R, Kurniawan D, Tania K D. Performance Analysis of YOLO, Faster R-CNN, and DETR for Automated Personal Protective Equipment Detection[J]. Journal of Applied Informatics and Computing, 2025, 9 (6): 3810-3820.

[8] Padilla R, Netto S L, Da Silva E A B. A survey on performance metrics for object-detection algorithms[C]//2020 international conference on systems, signals and image processing (IWSSIP). IEEE, 2020: 237-242.

[9] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[10] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv preprint arXiv:2410.17725, 2024.

[11] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.

Investigating the Interpretability of Pedestrian Detection Models via Testing with Concept Activation Vectors (TCAV): A Cross-Architectural Empirical Analysis

Liu Chenfeng

***Annotation*—This study addresses the "black-box" issue in pedestrian detection models deployed in high-risk scenarios, such as autonomous driving and intelligent security systems, by employing the Testing with Concept Activation Vectors (TCAV) framework. We conduct a cross-layer empirical analysis of feature representations across three representative architectures: Faster R-CNN, YOLOv11, and RT-DETR. An automatic cropping strategy based on human anatomical proportions is introduced to construct semantic concept sets (head, torso, and legs), enabling quantitative evaluation of semantic evolution at different network depths. Results show that Faster R-CNN follows a progressive semantic modeling pattern from local textures to global structures while maintaining stable semantic disentanglement in deeper layers; YOLOv11 exhibits semantic latency, with structured human-body representations emerging mainly within the feature fusion module rather than the backbone network; and RT-DETR, despite strong semantic separability, demonstrates gradient sparsity induced by output saturation, limiting the effectiveness of linear interpretability methods. These findings delineate the applicability boundaries of TCAV-based interpretation across heterogeneous architectural paradigms and provide a quantitative basis for designing highly transparent pedestrian detection systems.**

***Keywords*—deep learning architectures, testing with Concept Activation Vectors (TCAV), pedestrian detection, model interpretability.**

REFERENCES

[1] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature machine intelligence, 2019, 1 (5): 206-215.

[2] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]//International conference on machine learning. PMLR, 2018: 2668-2677.

[3] Seyedmomeni F S, Keyvanrad M A. Explaining What Machines See: XAI Strategies in Deep Object Detection Models[J]. arXiv preprint arXiv:2509.01991, 2025.

[4] Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3213-3221.

[5] Chen V, Yang M, Cui W, et al. Best practices for interpretable machine learning in computational biology[J]. Biorxiv, 2022: 2022.10. 28.513978.

[6] Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps[J]. Advances in neural information processing systems, 2018, 31.

[7] Naufaldihanif R, Kurniawan D, Tania K D. Performance Analysis of YOLO, Faster R-CNN, and DETR for Automated Personal Protective Equipment Detection[J]. Journal of Applied Informatics and Computing, 2025, 9 (6): 3810-3820.

[8] Padilla R, Netto S L, Da Silva E A B. A survey on performance metrics for object-detection algorithms[C]//2020 international conference on systems, signals and image processing (IWSSIP). IEEE, 2020: 237-242.

[9] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[10] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv preprint arXiv:2410.17725, 2024.

[11] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16965-16974.