DOI 10.37882/2223-2966.2024.01.29

ОБНАРУЖЕНИЕ ТРАНСПОРТНЫХ СРЕДСТВ И ПЕШЕХОДОВ ДЛЯ АВТОНОМНЫХ ТРАНСПОРТНЫХ СРЕДСТВ НА ОСНОВЕ MOBILEVIT

VEHICLE AND PEDESTRIAN DETECTION FOR AUTONOMOUS VEHICLES BASED ON MOBILEVIT

K. Parfentyev
Zhang Bohan

Summary. Since an autonomous driving information sensing and fusion system needs to identify traffic conditions and various obstacle attributes in a timely and accurate manner, it is especially important to develop an obstacle detection model that achieves both high detection speed and high accuracy. First, a target detection model combining convolutional neural network and Vision Transformer is designed to extract local and global information about vehicles in images. Secondly, an attention-grabbing mechanism module has been introduced to enhance the model's ability to focus on regions; To improve the effect of multi-scale object fusion, double triple interpolation is implemented in the upsampling module. Finally, to ensure real-time performance of the model, the lightweight MobileVit network is used as the model backbone. Experimental results show that AM-Swin Transformer, a MobileVit-based fast vehicle and pedestrian detection model in front of autonomous vehicles proposed in this paper, performs better than other models in terms of vehicle and pedestrian detection accuracy and speed.

Keywords: computer vision, artificial neural networks, object detection, self-driving cars.

Парфентьев Кирилл Викторович
Кандидат технических наук, старший преподаватель, Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)
parfentiev@bmstu.ru
Чжан Бохань
Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)
bohan-zhang@qq.com

Аннотация. Поскольку автономная система восприятия и объединения информации о вождении должна своевременно и точно идентифицировать дорожную обстановку и различные атрибуты препятствий, особенно важно разработать модель обнаружения препятствий, обеспечивающую как высокую скорость обнаружения, так и высокую точность. Во-первых, модель обнаружения цели, объединяющая сверточную нейронную сеть и Vision Transformer, предназначена для извлечения локальной и глобальной информации о транспортных средствах на изображениях. Во-вторых, введен модуль механизма привлечения внимания, повышающий способность модели фокусироваться на регионах; чтобы улучшить эффект многомасштабного объединения объектов, в модуле повышающей дискретизации реализована двойная тройная интерполяция. Наконец, для обеспечения производительности модели в режиме реального времени в качестве магистральной сети модели используется облегченная сеть MobileVit. Результаты экспериментов показывают, что AM-Swin Transformer, модель быстрого обнаружения транспортных средств и пешеходов перед автономными транспортными средствами на основе MobileVit, предложенная в этой статье, работает лучше других моделей с точки зрения точности и скорости обнаружения транспортных средств и пешеходов.

Ключевые слова: компьютерное зрение, искусственные нейронные сети, обнаружение объектов, беспилотные автомобили.

Введение

Распознавание объектов интереса — это быстрая и точная идентификация таких параметров цели, как тип и местоположение на изображении или кадре видео. Извлечение признаков методами делится на две основные категории: традиционное извлечение признаков и извлечение признаков на основе сверточных нейронных сетей. Традиционные методы обладают следующими недостатками [1]:

1. Выбор региона на основе скользящего окна является исчерпывающей стратегией, и подход с использованием скользящего окна приводит к избыточности окон. Однако метод отнимает много времени и приводит к высокой временной сложности.
2. Традиционное выделение признаков, как правило, основано на априорном опыте и не является надежным, когда целевой объект имеет множество вариаций.

Эти недостатки в значительной степени ограничивают применение функций, извлекаемых вручную; поэтому разработка традиционных моделей обнаружения целей привела к застою.

В 2016 году была предложена одноступенчатая модель обнаружения цели под названием YOLOv1[2]. Хотя модель YOLOv1 обеспечивает быстрое обнаружение цели и не использует опорные рамки, точность ее обнаружения невысока. В 2016 году [3] была предложена одноступенчатая модель обнаружения цели под назва-

нием SSD (Single shot multibox detector), которая использует опорные кадры для локализации цели. Анкерные рамки (anchors) используются в моделях FasterR-CNN, YOLOv2, YOLOv2v3 и SSD, которые обеспечивают лучшее решение для целевой локализации. Однако опорные кадры также вводят больше гиперпараметров, тем самым увеличивая вычислительную мощность модели, а также проблему неравномерных положительных и отрицательных выборок.

В 2021 году был предложен swine transformer [4], который показал лучшую производительность в задачах классификации, обнаружения и сегментации, чем сверточная нейронная сеть. Хотя трансформер swine удовлетворяет требованиям к производительности в режиме реального времени, модель не может распознавать и обнаруживать небольшие цели на отдалении. Описанные выше методы с трудом обеспечивают хороший баланс между точностью обнаружения и производительностью в режиме реального времени. Поэтому в этом исследовании предлагается облегченная модель, основаная на MobileVit [5], которая обеспечивает точность обнаружения и хорошую производительность в режиме реального времени.

Механизм внимания

Обычно целью механизма внимания является эффективное изучение распределения веса различных частей карты объектов при одновременном снижении влияния фоновой информации для повышения точности распознавания и надежности модели. Например, сеть остаточного внимания [6] использует остаточный механизм для построения сети, которая вводит структуру внимания, обеспечивая при этом глубину сети. Модуль внимания сверточного блока [7] использует как канальную, так и пространственную информацию карты объектов для проектирования модуля внимания, который может фокусироваться на большем количестве полезной информации и еще больше расширять свои возможности по извлечению объектов. В работе предлагается добавить модуль внимания блока свертки в магистральную сеть, чтобы сфокусировать модель обнаружения на областях, подверженных воздействию транспортных средств или пешеходов в процессе вождения автономного транспортного средства. Канальное и пространственное внимание можно выразить в виде следующих уравнений:

$$F' = M_c(F) \otimes F, \quad (1)$$

$$F'' = M_s(F) \otimes F, \quad (2)$$

где $\otimes$ обозначает поэлементное умножение, F обозначает входную карту объектов, F'обозначает уточненную карту объектов, F'' и обозначает конечный уточненный результат.

Канальное внимание: каждое сверточное ядро можно рассматривать как детектор признаков, следовательно, каждый канал создает карту признаков, которая может представлять один объектный признак, и цель внимания канала состоит в том, чтобы сосредоточиться на наиболее значимой части всех каналов. Сначала сжимается размерность карты объектов, которая проходит через слой максимального объединения и слой среднего объединения, а затем выводятся два дескриптора объектов: слой максимального объединения, чтобы подчеркнуть важные особенности объекта, и слой среднего объединения, чтобы эффективно вычислить диапазон объекта. Затем оба дескриптора объектов перенаправлены в общую сеть, состоящую из каскадных уровней: входного уровня, выходного уровня и трех скрытых слоев. Когда два дескриптора проходят через общую сеть, используется поэлементное суммирование для объеди-

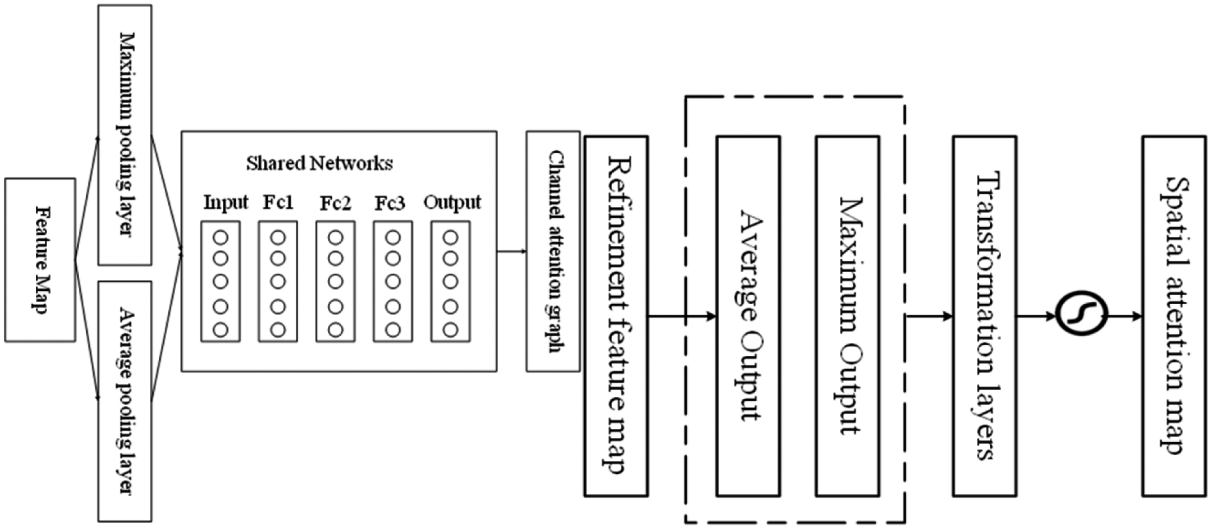

Рис. 1. Модуль внимания

нения выходных векторов признаков. Наконец, векторы признаков активируются с помощью сигмовидной функции для получения карты внимания канала.

Пространственное внимание: отношения пространственных объектов используются для создания карты пространственного внимания. Когда изображение поступает в сверточную нейронную сеть, каждый пиксель изображения участвует в вычислении. Подобно канальному вниманию, пространственное внимание фокусируется на областях изображения, которые вносят наибольший вклад в объект. Сначала карта внимания канала и уточненная карта объектов, вычисленная на основе карты объектов, пропускаются через слой максимального объединения и слой среднего объединения, соответственно, для получения двух описаний объектов. Затем оба дескриптора объектов соединяются и применяются два слоя свертки, чтобы подчеркнуть области дескрипторов. Наконец, карта внимания канала получается путем активации вектора с использованием сигмоидной функции. После использования модулей канального и пространственного внимания веса карт объектов оптимизируются, и окончательные карты объектов содержат более подробную информацию о транспортном средстве или пешеходе. Предполагается, что средним и максимальным процессами объединения являются F_{avg} и $F_{\max}$ соответственно. Att_{avg} может хорошо отфильтровывать глобальную фоновую информацию об объекте, а $Att_{\max}$ может выделять важные особенности транспортного средства или пешеходных зон. Пусть $X = [x_1, x_2, ..., x_n]$, где x_n обозначает вес n-го ядра свертки: уравнения для Att_{avg} и $Att_{\max}$ следующие:

$$Att_{avg} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_n(i,j) = F_{avg}(x_n) \quad (3)$$

$$Att_{\max} = agr\max\left(\sum_{i=1}^{H} \sum_{j=1}^{W} x_n(i,j) \right) = F_{\max}(x_n) \quad (4)$$

После совместного использования сети выходные данные канала внимание могут быть выражены следующими уравнениями:

$$output_{avg} = relu\left(FC \times Att_{avg}\right) \quad (5)$$

$$output_{\max} = relu(FC \times Att) \quad (6)$$

$$output_{channel} = \sigma\left(output_{avg} \times output_{\max}\right) \quad (7)$$

Веса, полученные путем умножения матрицы, и отфильтрованные характеристики канала равны $W = [\omega_1, \omega_2, ..., \omega_n]$, что может быть выражено в виде уравнения, показанного ниже.

$$W = \left(x_n, output_{channel}\right) = x_n \times output_{channel} \quad (8)$$

После фильтрации по каналам данные вводятся в модуль пространственного внимания. Сначала векторы объектов передаются через слой среднего объединения и слой максимального объединения. Впоследствии,

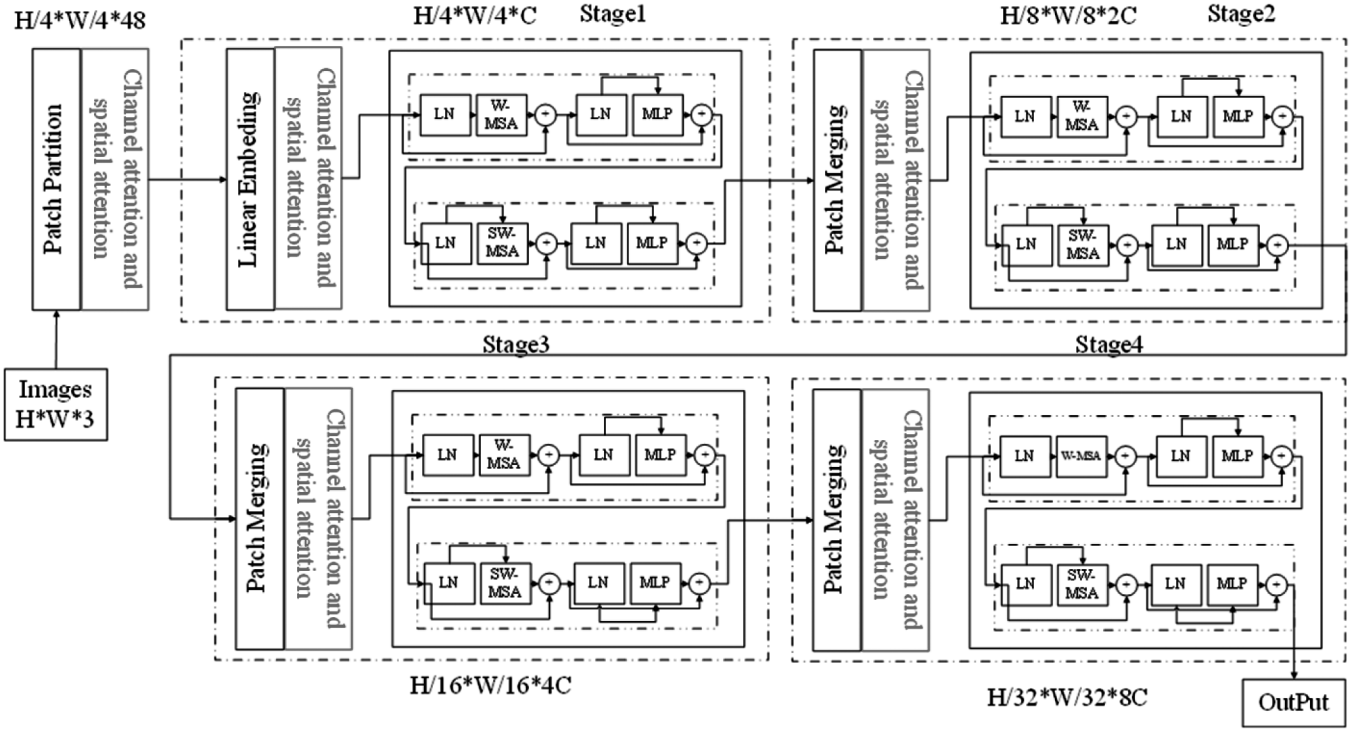

Рис. 2. Swin Transformer основанный на адаптивном механизме определения порога и внимания (A-Swing Transformer)

по размерам канала, элементы соединяются для получения $C_{conv} \in R^{1\times1\times2C}$. Чтобы получить информацию о весе объекта, требуется операция свертки, позволяющая $F_{5\times5}$ обозначать операцию свертки с входным каналом, равным 2, выходным каналом, равным 1, и размером ядра 5×5. Окончательные отфильтрованные веса могут быть выражены следующим образом

$$output_{cb\&sp} = F_{5\times5}\left(C_{conv}\right) \times W \quad (9)$$

Результатом работы всего модуля внимания является $output_{cb\&sp} + X$, который пересчитывает пропорции различных частей исходного входного вектора. Благодаря такой структуре модель может выборочно улучшать объекты, содержащие транспортные средства или пешеходные зоны, и подавлять несущественные или слабые объекты. Модель алгоритма показана на следующем рисунке.

Улучшения модели на снове MobileviT

Чтобы обеспечить точность при одновременном повышении производительности в режиме реального времени, предлагается модель облегченной сети MobileVit. Сверточная нейронная сеть эффективна для извлечения информации о локальных объектах, а преобразователь зрения, основанный на механизме самообучения, эффективен для извлечения информации о глобальных объектах. Выборка объектов на десятом уровне сети MobileVit была уменьшена в 32 раза. Полученная структура приведена в таблице 1.

Таблица 1.

Оптимизированная структура MobileVit

Размерность входа	Слой	Количество выходов
2562x3	conv2d	16
1282x16	MV2	32
1282x32	MV2	64
642x64	MV2	64
642x64	MV2	64
642x64	MV2	96
322x96	MVIT	96
322x96	MV2	128
162x128	MVIT	128
162x128	MV2	160

Следовательно, сеть MobileVit отбрасывается после десятого уровня, а оставшаяся сеть используется в качестве магистральной сети извлечения признаков MobileVit-10 модели. Предложенная структура модели Attention-Mobilevit-Swin Transformer (AM— Swin Transformer) представлена на рис. 3.

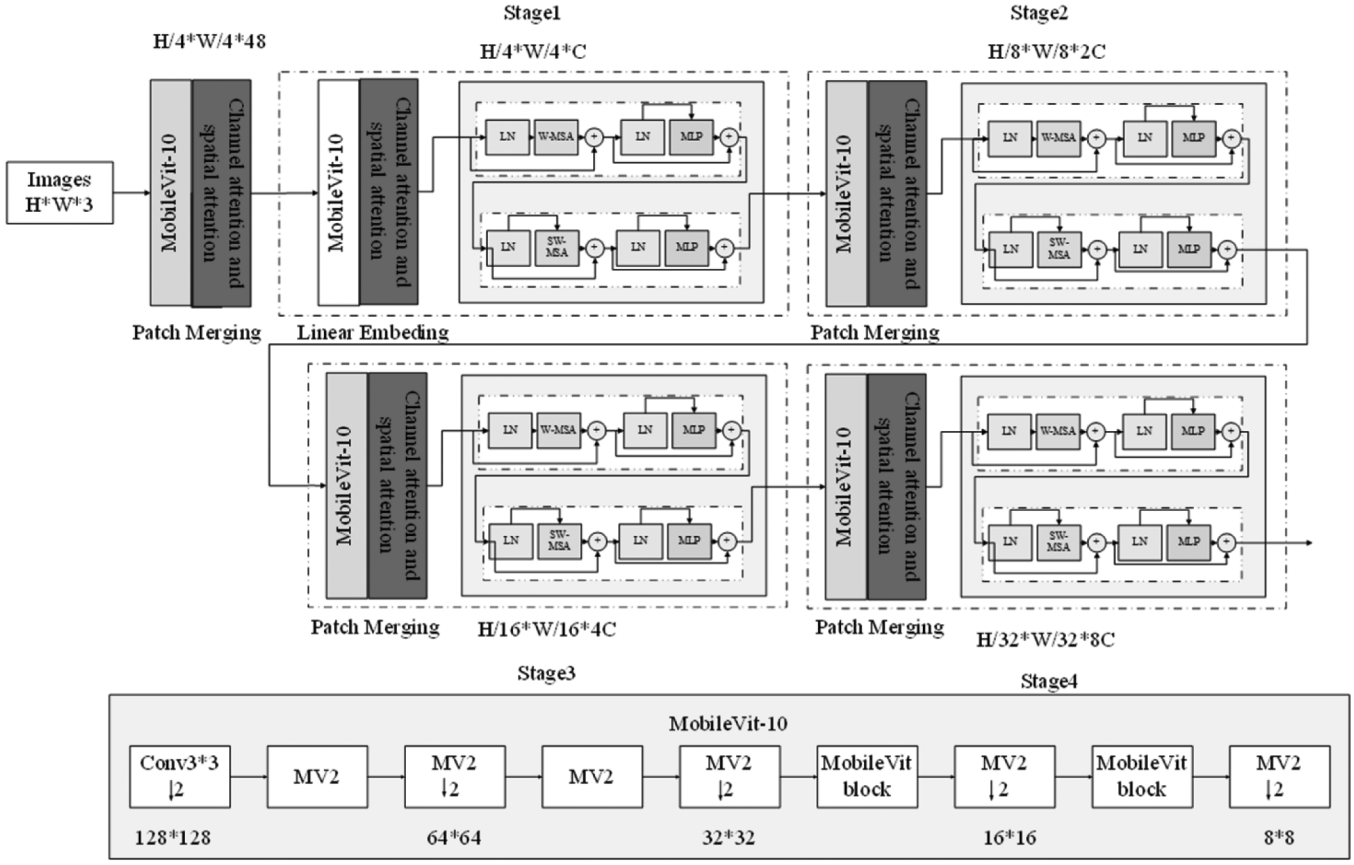

Рис. 3. Структура AM-Swin Transformer

Постановка эксперимента

Набор данных для обучения и тестирования получен с автомобиля Ola Black Cat с различными автомобильными датчиками и бортовыми вычислительными устройствами в качестве экспериментальной платформы для автономного вождения. Автомобиль оснащен 16-линейным лидаром, радаром миллиметрового диапазона, высокоточной автомобильной комбинированной навигационной системой, системой автомобильного зрения CalmCar и платой разработки Nvidia JetsonTX2. Шина CAN используется для связи между датчиками, двигателями и устройствами управления. На рис.4 приведены примеры набора данных:

В эксперименте используется 7481 изображений обучающей выборки из набора данных KITTI. После того, как процесс обучения завершился, модель была оценена с использованием 7518 изображений тестовой выборки. Согласно результатам тестирования на наборе данных KITTI, модель, разработанная в этом исследовании, и другие модели были получены для сравнения. В качестве оценки использовались следующие метрики: precision, recall, mAP и f1 score Результаты приведены в таблице 2.

Таблица 2.

Результаты тестирования набора данных

Методы	mAP	f1 score	recall	precision
Unet	0.83	0.37	0.68	0.82
YOLOv5	0.91	0.40	0.74	0.89
Faster R-CNN	0.78	0.33	0.61	0.73
AM-Swin Transformer	0.96	0.42	0.78	0.93

Рис. 4. Набор данных для обучения и тестирования

Тестирование демонстрирует, что алгоритм AM-Swin Transformer улучшает вероятность обнаружения препятствий в автономных транспортных средствах на 0,13, 0,05 и 0,18 по сравнению с Unet, YOLOv5 и более быстрым R-CNN соответственно. Алгоритм AM-Skin Transformer улучшает F1 обнаружения препятствий в автономном транспортном средстве на 0,05, 0,02 и 0,09 по сравнению с Unet, YOLOv5 и более быстрым R-CNN соответственно. Алгоритм преобразования AM-Skin повышает точность обнаружения препятствий на 0,11, 0,04 и 0,2 по сравнению с Unet, YOLOv5 и более быстрым R-CNN соответственно. Улучшена общая производительность обнаружения дорожных препятствий, а увеличение частоты отклика указывает на то, что алгоритм преобразования AM-Swin имеет больше преимуществ для обнаружения небольших целей.

Кривые изменения функции потерь во время обучения и валидации модели AM-Swin представлены на рис. 5.

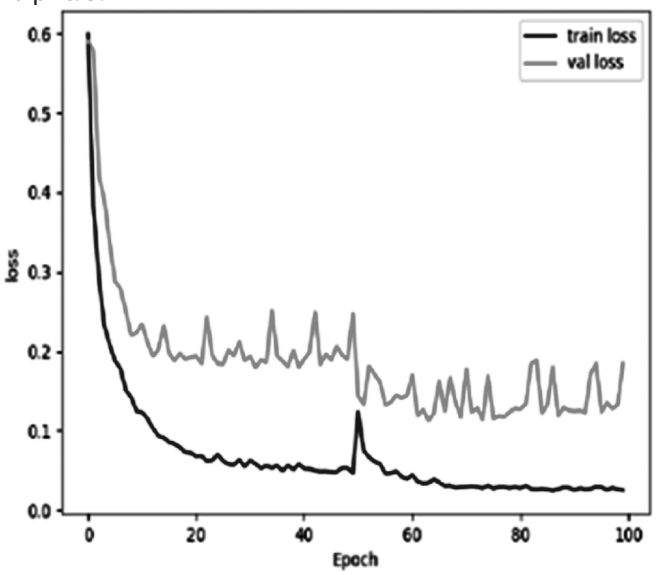

Рис. 5. Обучение нейросетевой модели

Некоторые примеры результатов алгоритма AM-Swin Transformer для обнаружения препятствий показаны на рисунке 6.

Для дальнейшего анализа производительности модели обнаружения препятствий самоуправляемым транспортным средством, построенной с использованием преобразователя AM-Swin, скорость вывода модели была сравнена со скоростью основных алгоритмов, и результаты приведены в таблице 3.

Согласно сравнению, ясно, что модель алгоритма преобразования AM-Swin намного быстрее других моделей с точки зрения скорости вывода. Поскольку модель U-Net проще, скорость ее вывода выше. Однако модель трансформера AM-Swin достигла аналогичной скорости вывода при гораздо более высокой точности.

Таблица 3.

Сравнение скорости работы нейросетевых моделей

Метод	Скорость обнаружения (fps)
Unet	23.2
YOLOv5	20.5
Faster R-CNN	22.6
AM-Swin Transformer	24.1

Заключение

В работе предложена нейросетевая модель на основе swin-transformer для обнаружения дорожных транспортных средств и пешеходов во время вождения автономных транспортных средств. Объясняется принцип работы механизма внимания и метод его моделирования, и предлагается модуль внимания, который может внедрять механизм внимания в канальную и пространственную области сверточной нейронной сети. Вместо традиционного сверточного модуля был внедрен MobileviT, чтобы повысить эффективность работы модели и удовлетворить требования к обнаружению целей в режиме реального времени.

Производительность и точность алгоритма быстрого обнаружения транспортных средств в режиме реального времени были доказаны экспериментально на данных, полученных с реального автономного транспортного средства.

Направление дальнейших исследований связано с улучшением структуры сети, дальнейшем повышении точности обнаружения за счет обеспечения производительности в режиме реального времени, снижении частоты ошибок обнаружения и повышении безопасности самоуправляемых транспортных средств.

Рис. 6. Результат работы AM-Swin

ЛИТЕРАТУРА

1. Rosenberg C., Hebert M., Schneiderman H. Semi-Supervised Self-Training of Object Detection Models //Applications of Computer Vision and the IEEE Workshop on Motion and Video Computing, IEEE Workshop on. — IEEE Computer Society, 2005. — Т. 1. — С. 29–36.
2. Redmon J. et al. You only look once: Unified, real-time object detection //Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 779–788.
3. Liu W. et al. Ssd: Single shot multibox detector //Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. — Springer International Publishing, 2016. — С. 21–37.
4. Liu Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows //Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — С. 10012–10022.
5. Mehta S., Rastegari M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv 2021 //arXiv preprint arXiv:2110.02178.
6. Woo S. et al. Cbam: Convolutional block attention module //Proceedings of the European conference on computer vision (ECCV). — 2018. — С. 3–19.
7. Deng J. et al. Imagenet: A large-scale hierarchical image database //2009 IEEE conference on computer vision and pattern recognition. — Ieee, 2009. — С. 248–255.