КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 1 С. 57–73
DOI: 10.20537/2076-7633-2023-15-1-57-73

Ки&М

МОДЕЛИ В ФИЗИКЕ И ТЕХНОЛОГИИ

УДК: 004.93'1

Двухпроходная модель Feature-Fused SSD для детекции разномасштабных изображений рабочих на строительной площадке

М. Н. Петров[1,a], С. В. Зимина[1], Д. Л. Дьяченко[2,b], А. В. Дубоделов[2,c], С. С. Симаков[1,d]

[1]Московский физико-технический институт,
Россия, 141707, г. Долгопрудный, Институтский пер., д. 9
[2]«Акселерэйшн Диджитал»,
Россия, 115114, г. Москва, Столярный пер., д. 3

E-mail: [a] mikhail.petrov@phystech.edu, [b] d.dyachenko@acceleration.ru, [c] a.dubodelov@acceleration.ru, [d] simakov.ss@phystech.edu

Получено 01.09.2022, после доработки — 31.10.2022.
Принято к публикации 13.12.2022.

При распознавании рабочих на изображениях строительной площадки, получаемых с камер наблюдения, типичной является ситуация, при которой объекты детекции имеют сильно различающийся пространственный масштаб относительно друг друга и других объектов. Повышение точности детекции мелких объектов может быть обеспечено путем использования Feature-Fused модификации детектора SSD (Single Shot Detector). Вместе с применением на инференсе нарезки изображения с перекрытием такая модель хорошо справляется с детекцией мелких объектов. Однако при практическом использовании данного подхода требуется ручная настройка параметров нарезки. При этом снижается точность детекции объектов на сценах, отличающихся от сцен, использованных при обучении, а также крупных объектов. В данной работе предложен алгоритм автоматического выбора оптимальных параметров нарезки изображения в зависимости от соотношений характерных геометрических размеров объектов на изображении. Нами разработан двухпроходной вариант детектора Feature-Fused SSD для автоматического определения параметров нарезки изображения. На первом проходе применяется усеченная версия детектора, позволяющая определять характерные размеры объектов интереса. На втором проходе осуществляется финальная детекция объектов с параметрами нарезки, выбранными после первого прохода. Был собран датасет с изображениями рабочих на строительной площадке. Датасет включает крупные, мелкие и разноплановые изображения рабочих. Для сравнения результатов детекции для однопроходного алгоритма без разбиения входного изображения, однопроходного алгоритма с равномерным разбиением и двухпроходного алгоритма с подбором оптимального разбиения рассматривались тесты по детекции отдельно крупных объектов, очень мелких объектов, с высокой плотностью объектов как на переднем, так и на заднем плане, только на заднем плане. В диапазоне рассмотренных нами случаев наш подход превосходит подходы, взятые в сравнение, позволяет хорошо бороться с проблемой двойных детекций и демонстрирует качество 0,82–0,91 по метрике mAP (mean Average Precision).

Ключевые слова: компьютерное зрение, строительная площадка, одностадийный детектор

© 2023 Михаил Николаевич Петров, Софья Васильевна Зимина, Дмитрий Львович Дьяченко, Артём Викторович Дубоделов, Сергей Сергеевич Симаков
Статья доступна по лицензии Creative Commons Attribution-NoDerivs 3.0 Unported License.
Чтобы получить текст лицензии, посетите веб-сайт http://creativecommons.org/licenses/by-nd/3.0/
или отправьте письмо в Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 1 P. 57–73
DOI: 10.20537/2076-7633-2023-15-1-57-73

Ки&М

MODELS IN PHYSICS AND TECHNOLOGY

UDC: 004.93'1

Dual-pass Feature-Fused SSD model for detecting multi-scale images of workers on the construction site

M. N. Petrov[1,a], S. V. Zimina[1], D. L. Dyachenko[2,b], A. V. Dubodelov[2,c], S. S. Simakov[1,d]

[1]Moscow Institute of Physics and Technology,
9 Institutskii Lane, Dolgoprudny, 141707, Russia
[2]Acceleration Digital,
3 Stolyarnii Lane, Moscow, 115114, Russia

E-mail: [a] mikhail.petrov@phystech.edu, [b] d.dyachenko@acceleration.ru, [c] a.dubodelov@acceleration.ru, [d] simakov.ss@phystech.edu

Received 01.09.2022, after completion — 31.10.2022.
Accepted for publication 13.12.2022.

When recognizing workers on images of a construction site obtained from surveillance cameras, a situation is typical in which the objects of detection have a very different spatial scale relative to each other and other objects. An increase in the accuracy of detection of small objects can be achieved by using the Feature-Fused modification of the SSD detector. Together with the use of overlapping image slicing on the inference, this model copes well with the detection of small objects. However, the practical use of this approach requires manual adjustment of the slicing parameters. This reduces the accuracy of object detection on scenes that differ from the scenes used in training, as well as large objects. In this paper, we propose an algorithm for automatic selection of image slicing parameters depending on the ratio of the characteristic geometric dimensions of objects in the image. We have developed a two-pass version of the Feature-Fused SSD detector for automatic determination of optimal image slicing parameters. On the first pass, a fast truncated version of the detector is used, which makes it possible to determine the characteristic sizes of objects of interest. On the second pass, the final detection of objects with slicing parameters selected after the first pass is performed. A dataset was collected with images of workers on a construction site. The dataset includes large, small and diverse images of workers. To compare the detection results for a one-pass algorithm without splitting the input image, a one-pass algorithm with uniform splitting, and a two-pass algorithm with the selection of the optimal splitting, we considered tests for the detection of separately large objects, very small objects, with a high density of objects both in the foreground and in the background, only in the background. In the range of cases we have considered, our approach is superior to the approaches taken in comparison, allows us to deal well with the problem of double detections and demonstrates a quality of 0.82–0.91 according to the mAP (mean Average Precision) metric.

Keywords: computer vision, construction site, single shot detector

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 1, pp. 57–73 (Russian).

© 2023 Mikhail N. Petrov, Sofya V. Zimina, Dmitri L. Dyachenko, Artem V. Dubodelov, Sergey S. Simakov
This work is licensed under the Creative Commons Attribution-NoDerivs 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/3.0/
or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Введение

Распознавание рабочих на изображениях строительной площадки, получаемых с камер наблюдения, является широко распространенной задачей мониторинга строительно-монтажных работ. Анализ последовательности изображений с одной камеры или наборов изображений с групп камер позволяет в автоматическом режиме отслеживать активность выполняемых работ, соответствие плановых и фактических работ и мест их проведения, соблюдение техники безопасности и др. Данная задача является особенно актуальной при мониторинге больших строительных объектов, на которых работы одновременно ведутся десятками и сотнями рабочих. Несмотря на бурное развитие в последние годы алгоритмов детекции людей, в том числе рабочих, на фото- и видео-изображениях до сих пор актуальной остается проблема обработки изображений, содержащих большое количество объектов (рабочих) различного пространственного масштаба относительно друг друга и размера самого изображения. В том числе типичными являются изображения, содержащие большое количество мелких объектов. Такие объекты часто плохо различимы даже человеческим взглядом. Автоматизация детекции объектов на изображениях с камер наблюдения значительно ускоряет процесс анализа изображения и позволяет создать прикладные программные инструменты мониторинга.

В связи с развитием методов глубокого машинного обучения получили распространение нейросетевые подходы, которые позволяют решать широкий класс задач компьютерного зрения, в частности детекции объектов [Zhao et al., 2019]. Для детекции объектов можно выделить два ключевых типа архитектур. Первый тип — это одностадийные (однопроходные, или one-stage) подходы, такие как SSD [Liu, 2016], YOLO [Redmon et al., 2016], RFBNet [Deng et al., 2019]. Второй тип — это двухстадийные (двухпроходные, или two-stage) подходы, такие как Faster-RCNN [Ren et al., 2015], Mask-RCNN [He et al., 2017], Reasoning-RCNN [Xu et al., 2019]. В двухстадийных подходах модель сначала определяет набор областей интереса, например с помощью выборочного поиска. Затем классификатор обрабатывает только кандидатов из этого набора. В одностадийных подходах этап выбора области интереса отсутствует и обнаружение объектов интереса производится непосредственно в плотной выборке возможных местоположений, которая определяется архитектурой нейронной сети. В связи с этим одностадийные подходы, как правило, требуют меньше вычислительных ресурсов и шире используются на практике. Среди таких архитектур можно выделить SSD и YOLO.

Нейросетевые подходы применяются и для решения задач компьютерного зрения на строительной площадке. В работе [Fang et al., 2018] для детекции рабочих и строительной техники на площадке используется модель IFaster R-CNN. В работе демонстрируются высокая точность представленной модели и некоторые успехи для детекции, в том числе и мелких объектов. В статье [Arabi, Haghighat, Sharma, 2020] для детекции техники предлагается использовать SSD-детектор MobileNet. Судя по примерам, представленным в работе, модель предназначена для детекции крупных объектов. Авторы работы [Kim et al., 2018] предлагают R-FCN-модель, построенную с применением transfer learning для детекции техники. В работе [Fang et al., 2018] применяется модель Faster R-CNN для детекции рабочих и строительной техники.

Базовый вариант детектора SSD сжимает входное изображение до размера 300×300 пикселей. В результате, если входное изображение имеет большее разрешение, а объекты интереса на изображении имеют малый размер, они с большой вероятностью исчезают из дальнейшего рассмотрения. Для решения проблемы детекции мелких объектов можно использовать модифицированный вариант SSD, который производит сжатие входного изображения до размеров 512×512 пикселей. В качестве следующего шага можно использовать вариант Feature Fusion SSD [Li, Zhou, 2017]. Такой подход позволяет бороться со многими проблемами базовой реализации, в частности, помимо лучшего нахождения мелких объектов, объединять разномасштабные признаки. Это достигается за счет объединения признаков с разных слоев сети с разными

масштабами и создания с помощью этого новой карты признаков. Если и этого оказывается недостаточно для обнаружения объектов, то в качестве постпроцессора может быть использована нарезка изображения с перекрытиями [Ozge Unel, Ozkalayci, Cigla, 2019]. Такой подход позволяет рассматривать одно изображение как совокупность его частей. При этом ясно, что мелкие объекты для каждой части исходного изображения будут иметь более крупный относительный размер, нежели они имели на исходном изображении. В этом случае на передний план выходит следующая проблема: параметры нарезки, подобранные для одной задачи, могу оказаться неоптимальными для другой. Также выбор параметров нарезки, оптимизированных для детекции мелких объектов, может негативно сказаться на определении более крупных объектов. В этом случае оказывается важным автоматическое определение параметров нарезки изображений. В статье [Li et al., 2020] используется сеть DMNet, которая строится на предположении, что плотность объектов на изображении зависит от интенсивности пикселей. Предполагается, что по интенсивности пикселей можно понять, есть ли в области интереса объекты или нет, что в свою очередь дает указания по статистической обрезке изображений. В статье [Wang, Yang, Zhao, 2020] для решения проблемы используется сеть CRENet, которая применяет алгоритм кластеризации для поиска областей с высокой плотностью объектов. Для автоматического определения параметров нарезки изображения в нашей работе предлагается применение двухпроходного (двухстадийного) детектора, представленного в данной работе. На первом проходе применяется усеченная версия детектора (SSD), позволяющая определять характерные размеры объектов интереса. На втором проходе осуществляется окончательная детекция объектов с параметрами нарезки, выбранными после первого прохода. Анализ результатов тестирования данного подхода показывает, что предлагаемая модель позволяет детектировать объекты (изображения рабочих) на строительной площадке с достаточно высокой точностью (0,82–0,91 по метрике mAP (mean Average Precision)).

Статья имеет следующую структуру: в разделе «Используемая архитектура» представлена основная архитектура, применяемая для детекции. В разделе «Алгоритм нарезки изображения» описан используемый алгоритм нарезки. Затем в разделе «Детекция в два прохода» изложена основная идея построения двухпроходной модели. Результаты применения предложенной модели представлены в разделе «Результаты». Заключение приводится в разделе «Заключение».

Используемая архитектура

В качестве детектора в работе используется архитектура FSSD512, основанная на архитектуре SSD (Single Shot Detector).

Базовая SSD-архитектура

В данной работе используется вариант SSD-архитектуры нейронной сети, принимающей на вход изображение размерами 512×512 (SSD512). Схематично данная SSD-архитектура представлена на рис. 1. Для извлечения признаков из входного изображения используются нейросеть VGG16 и несколько дополнительных сверточных слоев. В качестве карт признаков используются выходы некоторых последних слоев VGG16 (conv4_3, conv7) и выходы дополнительных слоев. Карты признаков для SSD512 имеют размеры (64, 32, 16, 8, 4, 2, 1). Для каждой клетки карты признаков задано несколько предопределенных ограничивающих рамок (боксов, от английского box) с центром, совпадающим с центром клетки, и разным соотношением сторон (1×1, 1×2, 2×1, 1×3, 3×1, еще один квадратный бокс с большими размерами, всего 4 или 6 боксов). Каждая карта признаков подается в соответствующий выходной слой, определяющий вероятности принадлежности классам и поправки к координатам для каждого из предопределенных боксов для этой карты признаков. Например, для выхода слоя conv4_3, имеющего разрешение 64×64

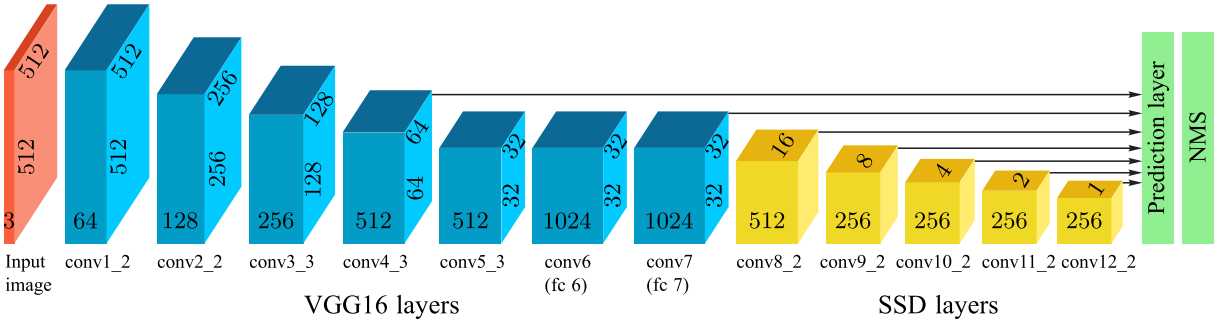


Рис. 1. Архитектура детектора изображений SSD512. Красным обозначено входное изображение, синим — базовая часть архитектуры, которая дублирует часть архитектуры модели VGG16, желтым — вспомогательная часть, зеленым — предсказательный слой и NMS

(в модели SSD512), будет предсказано 64×64×4 (разные соотношения сторон для предопределенных боксов), т. е. всего 16 384 бокса. Таким образом, модель SSD512 предсказывает координаты и классы для 24 564 боксов.

Архитектура FSSD

Для повышения качества детекции мелких объектов вместо архитектуры SSD в работе используется модель FSSD (Feature-Fused SSD), которая схематично представлена на рис. 2. В модели FSSD, в отличие от SSD, для получения предсказаний вместо выхода слоя VGG16 conv4_3 используется объединение выходов слоев conv4_3 и conv5_3. В исходной статье представлены различные способы такого объединения (feature fusion). В данной работе используется вариант, представленный на рис. 3. Для того чтобы карты признаков имели одинаковое разрешение, к выходу слоя conv5_3 применяется обратная свертка. Полученный результат конкатенируется с выходом слоя conv4_3. Так как за детекцию мелких объектов отвечают менее глубокие слои, то использование feature fusion позволяет улучшить качество их детекции.

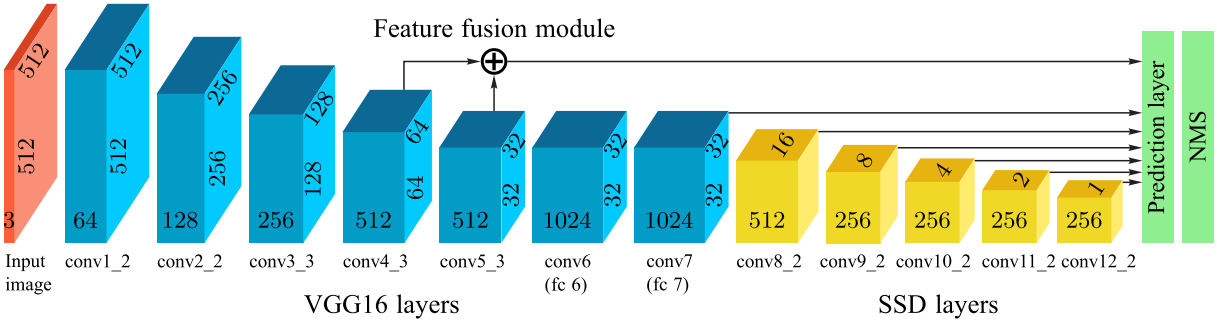


Рис. 2. Архитектура детектора изображений FSSD. Красным обозначено входное изображение, синим — базовая часть архитектуры, которая дублирует часть архитектуры модели VGG16, желтым — вспомогательная часть, зеленым — предсказательный слой и NMS. Знаком «+» отмечено следующее: выходы с каких слоев объединяются для передачи на предсказательный слой

Алгоритм NMS

Так как наша модель использует несколько карт признаков и для каждой клетки карты признаков предсказывает несколько боксов, то несколько предсказанных боксов могут соответствовать одному и тому же объекту. Для того чтобы оставить только наиболее релевантные предсказания, применяется алгоритм NMS (Non-Maximum Suppression, немаксимальное подавление), который состоит в следующем:

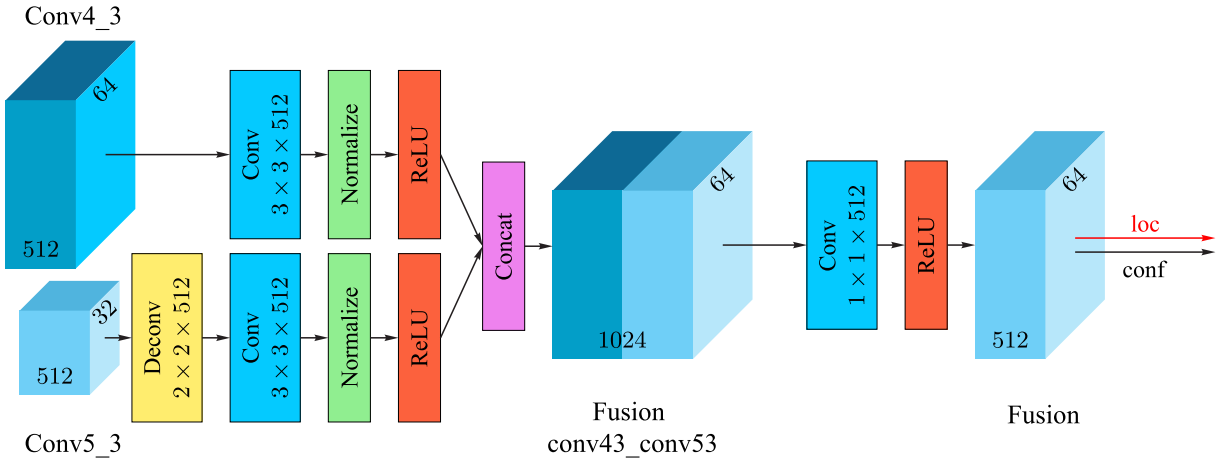
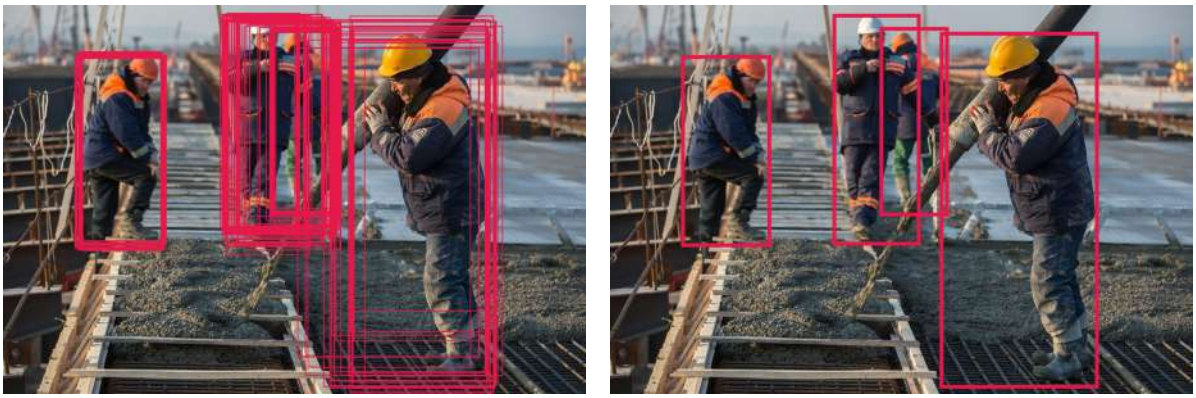


Рис. 3. Архитектура Feature-Fusion-слоя. На рисунке показывается, как именно объединяются выходы со слоев для последующей передачи на предсказательный слой. Желтым выделена обратная свертка, необходимая для того, чтобы карты признаков имели одинаковое разрешение. Также здесь Conv — свертка, Normalize — нормализация, ReLU — функция активации, Concat — конкатенация

- предсказанные боксы сортируются по вероятности уверенности модели в предсказании объекта;
- для каждой пары боксов рассчитывается IoU (отношение площади пересечения боксов к площади их объединения);
- если два бокса имеют IoU больше порога (max_overlap), то они считаются детекцией одного и того же объекта, и объект, имеющий меньшую вероятность предсказания, подавляется.

На рис. 4 показаны предсказания детектора до и после применения алгоритма NMS.



(а) До NMS (б) После NMS

Рис. 4. Предсказания детектора до (а) и после (б) применения алгоритма NMS. Рамкой отмечен найденный объект

Алгоритм нарезки изображения

Для повышения качества детекции мелких объектов в работе используется алгоритм нарезки исходного изображения с последующим объединением результатов детекции для каждой

его части. При этом мелкие объекты для каждой части исходного изображения будут иметь более крупный относительный размер, нежели они имели на исходном изображении, что также повышает итоговое качество детекции. Алгоритм нарезки применяется как к обучающей, так и к тестовой выборке.

Обучающая выборка

Применение к обучающей выборке выглядит следующим образом. Каждое изображение разбивается на $n \times m$ равных частей с перекрытиями. При составлении аннотации для каждой части учитываются только те боксы, центры которых находятся в пределах этой части. Используются следующие параметры нарезки:

1) tiles_n, tiles_m — число разбиений по горизонтали и вертикали соответственно;

2) inter_w, inter_h — процент перекрытия между двумя соседними патчами.

Пример разбиения показан на рис. 5, на котором границы каждой части обозначены соответствующим цветом.

Рис. 5. Пример разбиения исходного изображения на шесть частей. Каждая рамка своего цвета выделяет свою часть. Видно, что части изображений перекрываются

Применение к тестовым изображениям

При применении к тестовому изображению оно аналогичным образом разбивается на $m \times n$ одинаковых частей с перекрытиями таким образом, на вход модели подается батч из $m \times n + 1$ изображений, включающий как его части, так и исходное изображение целиком. Далее к предсказанным боксам применяется алгоритм постпроцессинга из нескольких стадий. Цель постпроцессинга — объединить предсказания, полученные для разных частей исходной картинки, и выделить из предсказанных боксов наиболее релевантные (модель предсказывает фиксированное число боксов, большая часть из которых имеет низкую вероятность предсказания, также несколько боксов могут соответствовать одному и тому же объекту).

Алгоритм постпроцессинга состоит из пяти этапов:

1) пересчет координат всех боксов на каждой части изображения в абсолютные координаты боксов на основном изображении;

2) фильтрация боксов, в которых модель не уверена ($score < min\_score$);

3) использование алгоритма NMS для каждого класса внутри каждой части изображения ($overlap > max\_overlap$);

4) использование алгоритма NMS на границах частей изображения, а также между исходным изображением и его частями ($overlap > max\_overlap\_global$).

При этом имеется возможность использовать разные пороги для разных стадий NMS.

Детекция в два прохода

Алгоритм, описанный выше, имеет определенные недостатки. Так как параметры нарезки определяются заранее, он применим только к одинаковым масштабам съемки. Слишком крупная нарезка приводит к низкому качеству детекции мелких объектов, а слишком мелкая нарезка — к множественным детекциям крупных объектов (рис. 6).

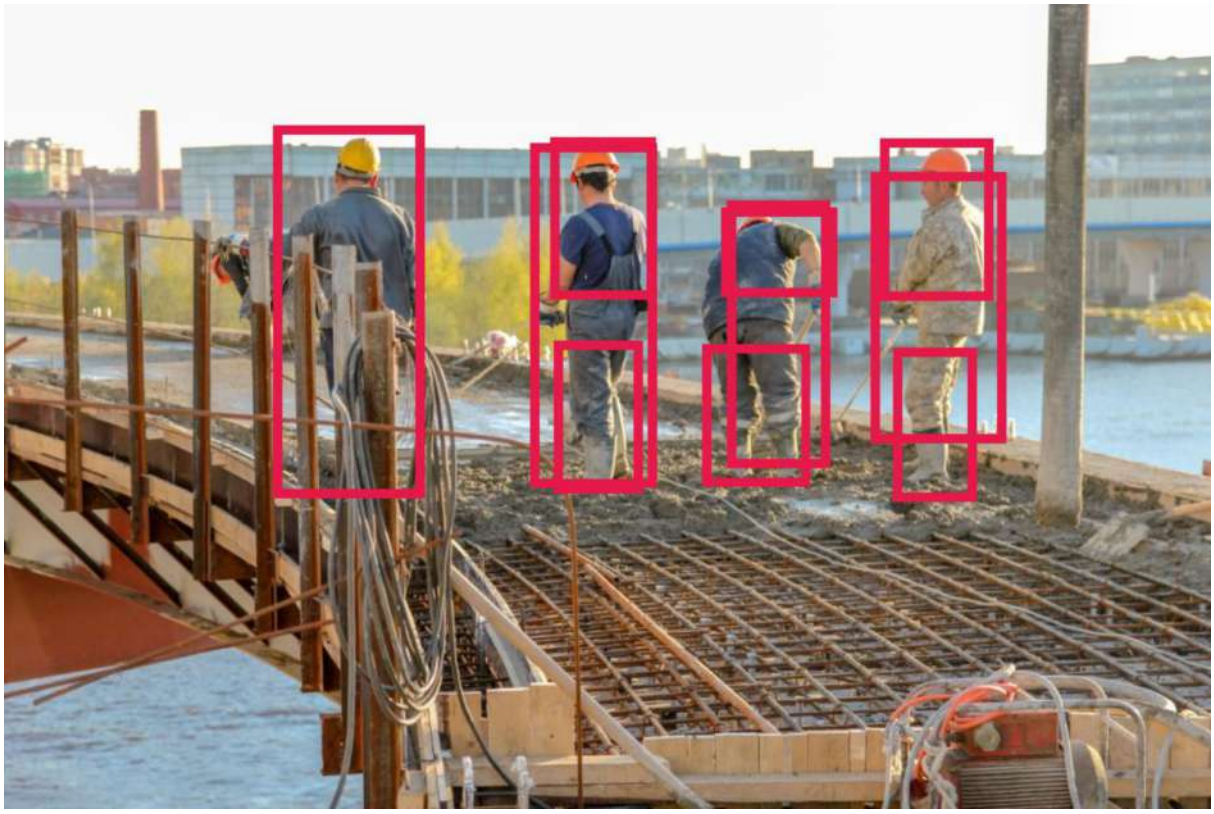


Рис. 6. Пример множественных детекций при слишком мелком разбиении. На рисунке видно, что трое из четырех рабочих отмечены сразу тремя рамками вместо одной

Для избежания множественных детекций предлагается применять двухпроходный алгоритм. На первом проходе входное изображение режется на фиксированное количество частей по умолчанию и подается в детектор. Среди полученных боксов отбираются те, в которых модель наиболее уверена ($score > min\_score\_prev$). Постобработка (NMS) на первом проходе не применяется. На основе полученных боксов рассчитывается средний размер бокса ($mean\_size$) по каждой из осей x, y, который используется для вычисления оптимального размера патча: $patch\_size_{x,y} = mean\_size_{x,y} \cdot rec\_coeff$. Уточненное число разбиений рассчитывается

как $\frac{image\_size_{x,y}}{patch\_size_{x,y}}$ (здесь $image\_size_{x,y}$ — это ширина и высота входного изображения соответственно, а деление подразумевается целочисленным). В рассмотренной задаче детекции рабочих на строительной площадке использовались значения гиперпараметра *rec_coeff* в диапазоне 6–8. На втором проходе изображение разбивается согласно подобранным на первом проходе параметрам и дальше применяется стандартный алгоритм детекции (аналогично детекции в один проход).

Результаты

Датасет

Для обучения и тестирования модели использовался датасет рабочих на строительной площадке, собранный вручную. Датасет состоит из 1882 фотографий рабочих (1750 — обучающая выборка, 132 — тестовая). Объекты имеют различный масштаб относительно изображения (в обучающей выборке ≈ 600 фотографий с крупными и средними объектами, ≈ 750 — с мелкими и ≈ 400 — с разноплановыми; в тестовой выборке ≈ 72 изображения с крупными и средними объектами, ≈ 30 — с мелкими и ≈ 30 — с разноплановыми). Примеры изображений из датасета представлены на рис. 7.

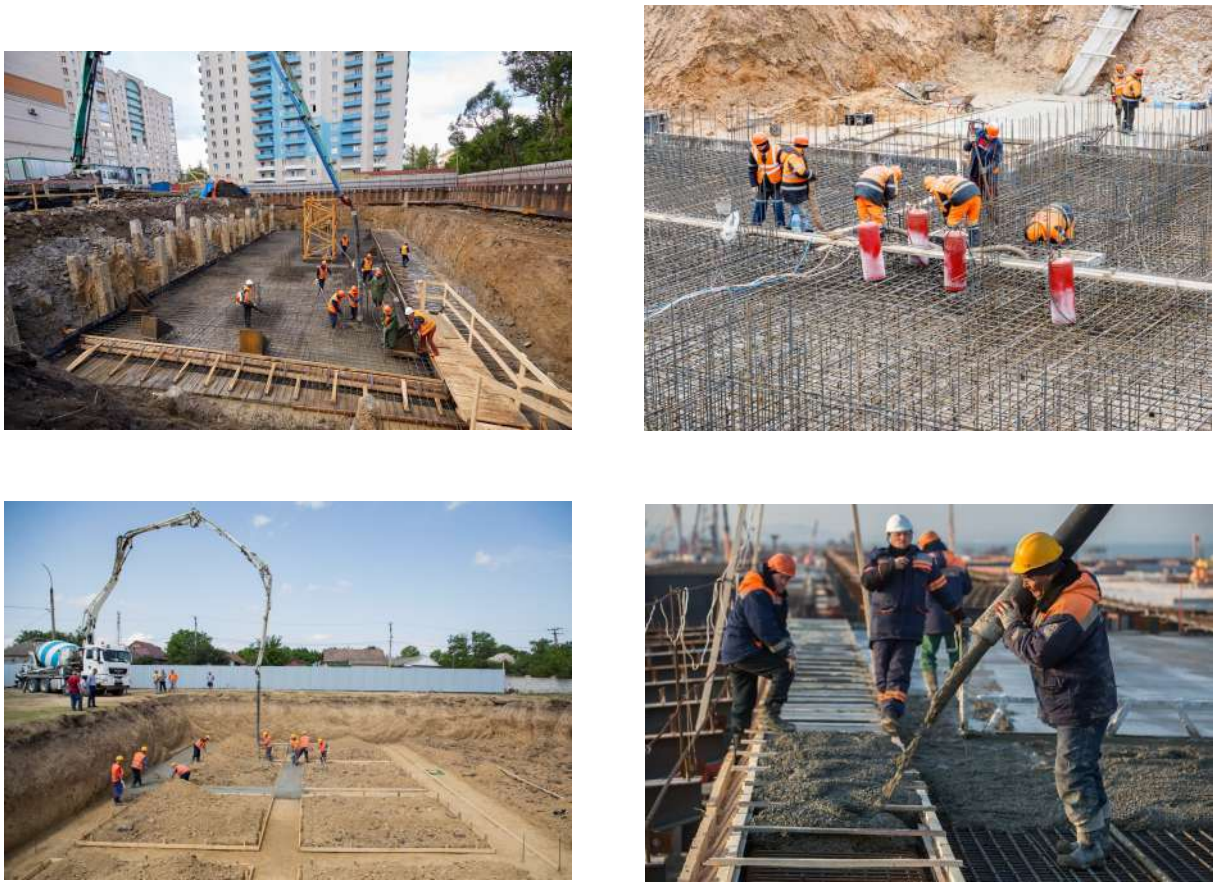


Рис. 7. Примеры изображений из датасета. Представлены варианты как с крупными, средними, мелкими объектами, так и с разноплановыми

Метрика

Для получения количественной оценки качества предсказаний и сравнения результатов, полученных разными методами, в работе использовались две метрики: *Accuracy* (точность) и *Mean Average Precision* (mAP) (усредненная точность). Для расчета этих метрик вначале определялось, является ли каждый предсказанный моделью бокс истинно положительной или ложноположительной детекцией. Совпадение предсказанного и реального бокса определялось по величине *IoU* (Intersection Over Union) — отношение площади пересечения боксов к площади их объединения:

$$IoU = \frac{S_{intersection}}{S_{union}}.$$

Истинно положительная детекция (true positive, TP) означает совпадение предсказанного бокса с реальным боксом объекта (IoU между предсказанным и реальным боксом выше порога), ложноположительная детекция (false positive, FP) означает отсутствие совпадений предсказанного бокса со всеми реальными боксами объектов данного класса (IoU ниже порога), ложноотрицательная (false negative, FN) — отсутствие для реального бокса объекта совпадающего с ним предсказанного бокса (примеры этих случаев представлены на рис. 8).

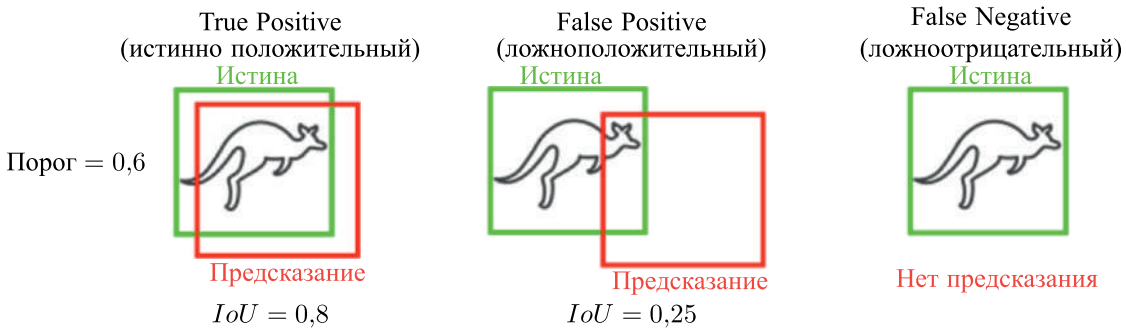


Рис. 8. Примеры истинно положительного (true positive), ложноположительного (false positive) и ложноотрицательного (false negative) результатов для задачи детекции объектов. Зеленым представлена истинная рамка найденного объекта, красным — предсказанная. Название результата детекции зависит от того, как эти рамки пересекаются

Метрика *Accuracy* считается как отношение числа истинно положительных детекций к сумме чисел истинно положительных, ложноположительных и ложноотрицательных детекций:

$$Accuracy = \frac{TP}{TP + FP + FN}.$$

Метрика *mAP* (Mean Average Precision) рассчитывалась способом, аналогичным расчету метрики в PascalVOC2007 [Everingham et al., 2010]. Вначале для каждого предсказанного бокса выяснялось, является ли он истинной (true positive) или ложной детекцией (false positive). Затем боксы сортировались по значению уверенности в предсказании (некий аналог вероятности) и считалась кумулятивная сумма true positive и false positive:

$$TP_i^{cumulative} = \sum_{j=0}^{i} TP_j,$$

$$FP_i^{cumulative} = \sum_{j=0}^{i} FP_j.$$

Затем на основе полученных кумулятивных значений TP и FP вычислялись precision (точность) и recall (полнота). Precision рассчитывается как отношение числа истинно положительных детекций к сумме чисел истинно положительных и ложноположительных детекций (или, что то же самое, отношение числа истинно положительных детекций к числу предсказаний модели). Recall — как отношение числа истинно положительных детекций к сумме чисел истинно положительных и ложноотрицательных детекций (или, что то же самое, отношение числа истинно положительных детекций к реальному числу объектов):

$$precision = \frac{TP}{TP + FP} = \frac{TP}{n_{predictions}},$$
$$recall = \frac{TP}{TP + FN} = \frac{TP}{n_{objects}}.$$

Average Precision (AP, средняя точность) вычислялось следующим образом: для каждого из порогов recall в диапазоне [0; 1] с шагом 0,1 бралось максимальное значение precision для всех recall выше порога и затем вычислялось среднее значение для всех порогов:

$$AP = \frac{1}{11} \sum_{recall \in \{0,\, 0{,}1,\, \ldots,\, 1\}} \max_{\widetilde{r} \geqslant recall} (precision(\widetilde{r})).$$

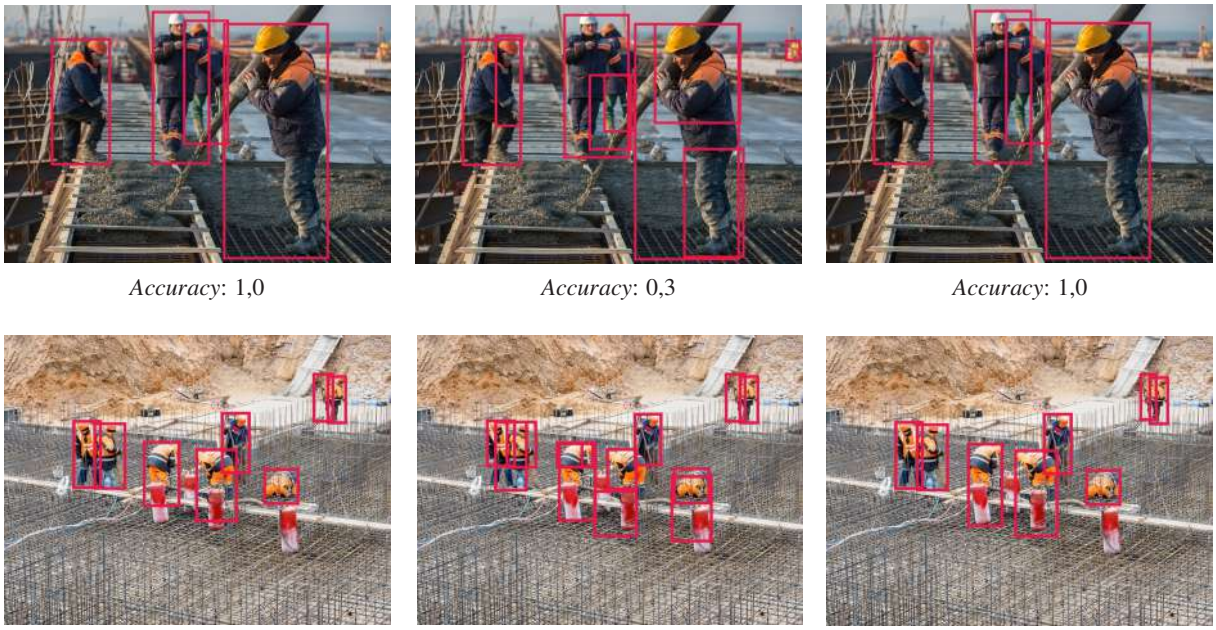
Обе использованные метрики рассчитывались при значении порога IoU, равном 0,5.

Сравнение методов

Результаты детекции в два прохода (с определением оптимальных параметров разбиения) сравнивались с детекцией в один проход (фиксированные параметры разбиения) и детекцией без разбиений. В качестве детектора во всех трех случаях использовалась архитектура FSSD512. Ниже представлены результаты для каждого из методов при разном размере объектов относительно изображения (крупные, мелкие и разноплановые). Также для каждого результата детекции была рассчитана метрика *Accuracy* и указана на представленных рисунках. На рис. 9 представлены результаты детекции относительно крупных объектов на изображении для модели FSSD без нарезки, с нарезкой и одним проходом сети и с оптимальной нарезкой двухпроходной модели. Из рисунка видно, что базовая модель, как и двухпроходная модель, одинаково хорошо справляется с задачей. С другой стороны, использование нарезки для детекции крупных объектов снижает качество предсказания. Это происходит из-за того, что при выборе неоптимальных параметров нарезки часто случаются ситуации двойной детекции объекта. Поскольку часто априори мелкость объектов на строительной площадке неизвестна, нарезка все же необходима, что будет видно из следующих примеров. Но фиксированное ручное определение параметров нарезки может приводить к неудовлетворительному результату. Пример оправдывает использования двухпроходной модели для подбора оптимальных параметров нарезки при ее использовании.

На рис. 10 представлены примеры использования модели FSSD без нарезки, с нарезкой и одним проходом сети и с оптимальной нарезкой двухпроходной модели для детекции очень мелких объектов. На рисунке в первом ряду представлено основное изображение и для каждого из изображений по вертикали представлены некоторые наиболее показательные его части. Этот пример демонстрирует преимущество двухпроходной модели для детекции мелких объектов. Рис. 11 показывает, что двухпроходная модель не ухудшает предсказание, даже если начальная нарезка задана оптимально.

Следующим примером в этой серии служат рис. 12, 13. Этот пример характеризуется высокой плотностью объектов как на переднем, так и на заднем плане. Как можно видеть, снова предсказания объектов заднего плана лучше в случае использования двухпроходной модели. Как



(а) Без разбиения (б) Один проход, 4×3 (в) Два прохода, 4×3

Рис. 9. Результаты детекции для (а) однопроходного алгоритма без разбиения входного изображения, (б) однопроходного алгоритма с равномерным разбиением 4×3 и (в) двухпроходного алгоритма с подбором оптимального разбиения (с разбиением 4×3 для первого прохода). В качестве детектора использовалась модель $FSSD512$, изображение взято из тестовой части датасета

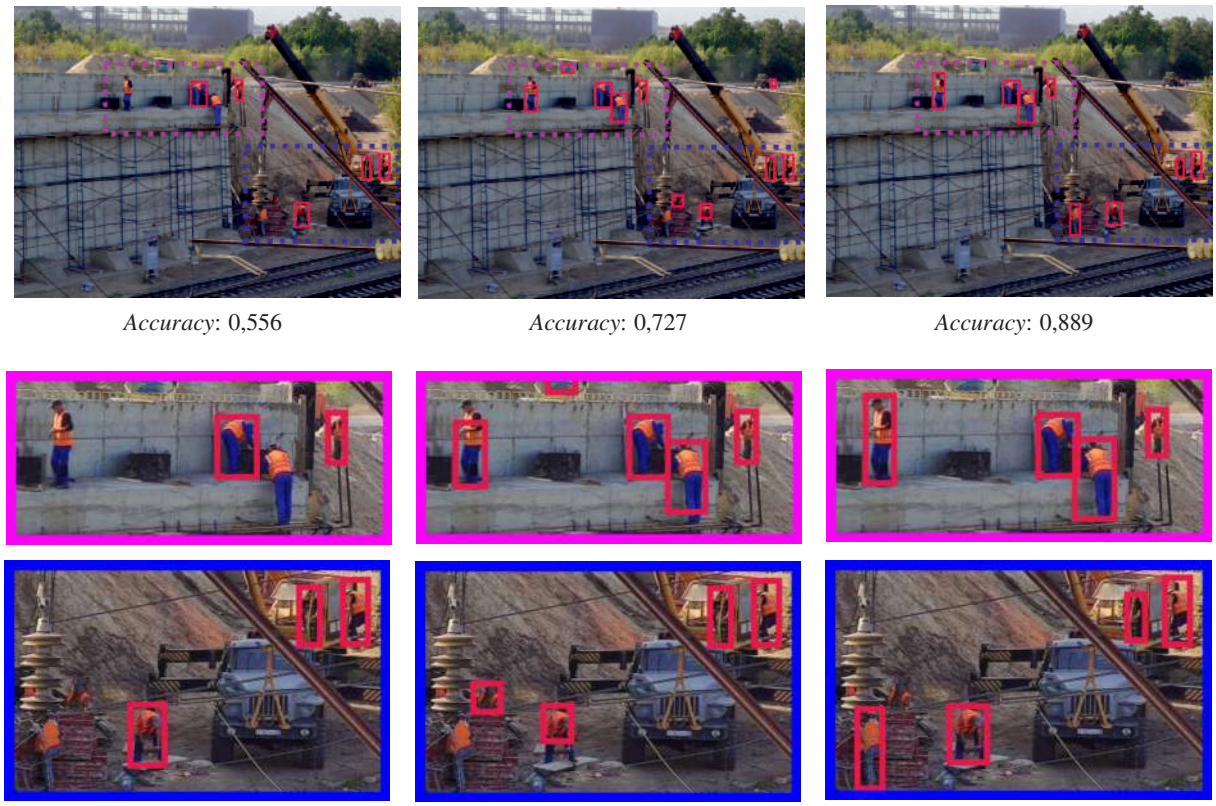
можно видеть из примера, неоптимальное задание параметров нарезки для однопроходной модели приводит к ухудшению предсказания на переднем плане.

Завершающий пример (рис. 14) в этой серии показывает, что двухпроходная модель позволяет избавить от проблемы двойной детекции объектов уже на заднем плане.

В таблице 1 приводится общая статистика по метрике mAP для каждого метода на всем датасете и для каждой его части отдельно, характеризующейся размером объектов (крупные объекты, мелкие объекты и разноплановые объекты). Из таблицы видно, что двухпроходной детектор дает лучшие предсказание как в общем, так и для каждой подгруппы в отдельности.

Таблица 1. Статистика качества предсказания по метрике mAP на всем датасете и отдельных его частях. Каждая часть характеризуется размером объектов, представленных в нем (крупные, мелкие и разноплановые)

Метод	Размер объекта			
	крупные	мелкие	разноплановые	все
1 проход, без разбиения	0,848	0,631	0,72	0,725
1 проход, 4×3	0,818	0,711	0,786	0,777
2 прохода, начальное разбиение 4×3	0,91	0,82	0,82	0,906



(а) Без разбиения (б) Один проход, 4×3 (в) Два прохода, 4×3

Рис. 10. Результаты детекции для (а) однопроходного алгоритма без разбиения входного изображения, (б) однопроходного алгоритма с равномерным разбиением 4×3 и (в) двухпроходного алгоритма с подбором оптимального разбиения (с разбиением 4×3 для первого прохода). В качестве детектора использовалась модель $FSSD512$, изображение взято из тестовой части датасета. По вертикали — основное изображение и некоторые увеличенные его части (выделены пунктиром на основном изображении)

(а) Без разбиения (б) Один проход, 4×3 (в) Два прохода, 4×3

Рис. 11. Результаты детекции для (а) однопроходного алгоритма без разбиения входного изображения, (б) однопроходного алгоритма с равномерным разбиением 4×3 и (в) двухпроходного алгоритма с подбором оптимального разбиения (с разбиением 4×3 для первого прохода). В качестве детектора использовалась модель $FSSD512$, изображение взято из тестовой части датасета. По вертикали — основное изображение и некоторые увеличенные его части (выделены пунктиром на основном изображении)

Accuracy: 0,7 *Accuracy*: 0,692 *Accuracy*: 1,0

(а) Без разбиения (б) Один проход, 4×3 (в) Два прохода, 4×3

Рис. 12. Результаты детекции для (а) однопроходного алгоритма без разбиения входного изображения, (б) однопроходного алгоритма с равномерным разбиением 4×3 и (в) двухпроходного алгоритма с подбором оптимального разбиения (с разбиением 4×3 для первого прохода). В качестве детектора использовалась модель $FSSD512$, изображение взято из тестовой части датасета. По вертикали — основное изображение и некоторые увеличенные его части (выделены пунктиром на основном изображении)

Accuracy: 0,667 *Accuracy*: 0,857 *Accuracy*: 1,0

(а) Без разбиения (б) Один проход, 4×3 (в) Два прохода, 4×3

Рис. 13. Результаты детекции для (а) однопроходного алгоритма без разбиения входного изображения, (б) однопроходного алгоритма с равномерным разбиением 4×3 и (в) двухпроходного алгоритма с подбором оптимального разбиения (с разбиением 4×3 для первого прохода). В качестве детектора использовалась модель $FSSD512$, изображение взято из тестовой части датасета. По вертикали — основное изображение и некоторые увеличенные его части (выделены пунктиром на основном изображении)

Accuracy: 0,75 *Accuracy*: 0,778 *Accuracy*: 0,875

(а) Без разбиения (б) Один проход, 4×3 (в) Два прохода, 4×3

Рис. 14. Результаты детекции для (а) однопроходного алгоритма без разбиения входного изображения, (б) однопроходного алгоритма с равномерным разбиением 4×3 и (в) двухпроходного алгоритма с подбором оптимального разбиения (с разбиением 4×3 для первого прохода). В качестве детектора использовалась модель $FSSD512$, изображение взято из тестовой части датасета. По вертикали — основное изображение и некоторые увеличенные его части (выделены пунктиром на основном изображении)

Заключение

В работе представлена идея построения модели для улучшения качества детекции разномасштабных рабочих на строительной площадке. В качестве основы архитектуры модели использовалась архитектура FSSD512. На первом проходе представленная модель применяла усеченную версию SSD-детектора для определения характерных размеров объектов и задания оптимальных параметров нарезки изображения для лучшей детекции на втором проходе. В качестве примеров применения модели были рассмотрены случаи детекции крупных объектов, мелких объектов и объектов, представленных на разных планах. Во всех представленных примерах модель позволяла добиться лучшей точности по сравнению с базовой моделью и моделью, использующей нарезку на инференсе.

Список литературы (References)

Arabi S., Haghighat A., Sharma A. A deep-learning-based computer vision solution for construction vehicle detection // Computer-Aided Civil and Infrastructure Engineering. — 2020. — Vol. 35, No. 7. — P. 753–767.

Deng L., Yang M., Li T., He Y., Wang C. RFBNet: deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation // arXiv preprint arXiv:1907.00135. — 2019.

Everingham M., Van Gool L., Williams C. K., Winn J., Zisserman A. The pascal visual object classes (voc) challenge // International journal of computer vision. — 2010. — Vol. 88, No. 2. — P. 303–338.

Fang W., Ding L., Zhong B., Love P. E., Luo H. Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach // Advanced Engineering Informatics. — 2018. — Vol. 37. — P. 139–149.

He K., Gkioxari G., Dollár P., Girshick R. Mask r-cnn // In Proceedings of the IEEE international conference on computer vision. — 2017. — P. 2961–2969.

Kim H., Kim H., Hong Y. W., Byun H. Detecting construction equipment using a region-based fully convolutional network and transfer learning // Journal of computing in Civil Engineering. — 2018. — Vol. 32, No. 2. — 04017082.

Li C., Yang T., Zhu S., Chen C., Guan S. Density map guided object detection in aerial images // In proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. — 2020. — P. 190–191.

Li Z., Zhou F. FSSD: feature fusion single shot multibox detector // arXiv preprint arXiv:1712.00960. — 2017.

Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C. Y., Berg A. C. Ssd: Single shot multibox detector // In European conference on computer vision. — Springer, Cham., 2016. — P. 21–37.

Ozge Unel F., Ozkalayci B. O., Cigla C. The power of tiling for small object detection // In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. — 2019.

Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection // In Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — P. 779–788.

Ren S., He K., Girshick R., Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks // Advances in neural information processing systems. — 2015. — Vol. 28.

Wang Y., Yang Y., Zhao X. Object detection using clustering algorithm adaptive searching regions in aerial images // In European Conference on Computer Vision. — Springer, Cham, 2020. — P. 651–664.

Xu H., Jiang C., Liang X., Lin L., Li Z. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection // In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2019. — P. 6419–6428.

Zhao Z. Q., Zheng P., Xu S. T., Wu X. Object detection with deep learning: A review // IEEE transactions on neural networks and learning systems. — 2019. — Vol. 30, No. 11. — P. 3212–3232.