Multimedia Tools and Applications (2022) 81:38297–38351
https://doi.org/10.1007/s11042-022-13153-y

Tools, techniques, datasets and application areas for object detection in an image: a review

Jaskirat Kaur[1] · Williamjeet Singh[2]

Received: 16 April 2021 / Revised: 24 February 2022 / Accepted: 10 April 2022 /
Published online: 23 April 2022

Abstract
Object detection is one of the most fundamental and challenging tasks to locate objects in images and videos. Over the past, it has gained much attention to do more research on computer vision tasks such as object classification, counting of objects, and object monitoring. This study provides a detailed literature review focusing on object detection and discusses the object detection techniques. A systematic review has been followed to summarize the current research work's findings and discuss seven research questions related to object detection. Our contribution to the current research work is (i) analysis of traditional, two-stage, one-stage object detection techniques, (ii) Dataset preparation and available standard dataset, (iii) Annotation tools, and (iv) performance evaluation metrics. In addition, a comparative analysis has been performed and analyzed that the proposed techniques are different in their architecture, optimization function, and training strategies. With the remarkable success of deep neural networks in object detection, the performance of the detectors has improved. Various research challenges and future directions for object detection also has been discussed in this research paper.

Keywords Computer vision · Object detection · Dataset · Deep learning

1 Introduction

There are different types of objects present in the real-world environment. Identifying those objects with the help of a machine is a complex task. Computer vision is a branch of computer science that enables the machine to see, identify, and process objects from still images and

✉ Jaskirat Kaur
jaskirat.scholar21@gmail.com

Williamjeet Singh
williamjeet@gmail.com

[1] Department of Computer Science, Punjabi University, Patiala, India

[2] Department of Computer Science and Engineering, Punjabi University, Patiala, India

videos in a visual world. A video is a sequence of continuous images (called video frames) displayed at a specific frame rate. Human beings immediately detect or recognize objects from images or video and determine their location owing to the brain's neurons interlinking. Video Object detection is an artificial intelligence task used for the same process to detect objects. Object detection is fundamental, and the longstanding computer vision problem has been a major active research area for a few decades [197]. It has been used in various computer vision applications such as face detection, face recognition, pedestrian counting, security system, vehicle detection, self-driving cars, etc. Some computer vision terms like object localization, classification, and recognition are interlinked to the object detection processing shown in Fig. 1.

Object classification defines the class of one or more objects that exist in the image and assigns the labels of the objects [56]. *Object localization* is a process that locates the position of one or more objects in the image or video with the help of a bounding box [56]. A combination of object localization and classification process is known as *Object detection* [214]. A complete *object recognition* process takes an image as input, identifies the objects, assigns labels to the object of the associated class, and gives the class probability of the recognized object [76]. Every object has unique features that help to identify the class. For example, a square with all equal sides helps detect the square-sized object. Most of the research work on object detection is divided into three categories: salient object detection, objectness detection, and category-specific object detection. The category-specific object detection's main objective is to identify the object categories from images, and it deals with interclass and intraclass variation and similarity [54]. This section discusses the basic steps of traditional object detection: Informative region, feature extraction, and classification.

Informative region Objects present in the image can vary in size and aspect ratios. The object detection model scans the whole image with different scales to detect the objects and find the recognizable pattern. The possible positions of objects can find out by this strategy, but there are some shortcomings. This strategy is expensive because several candidate windows are produced during processing that requires high computation power. Despite this, if this technique applies the fixed number of sliding windows, it may produce satisfactory regions.

Feature extraction Features are a particularly important parameter in an image that has been used for the classification and recognition of objects. In the first step, a pattern is recognized, and then this pattern is further used to extract the distinct features related to the object. Different techniques such as HOG, SIFT, and Haar are used to extract the features. However, designing a good and robust feature descriptor is challenging due to large

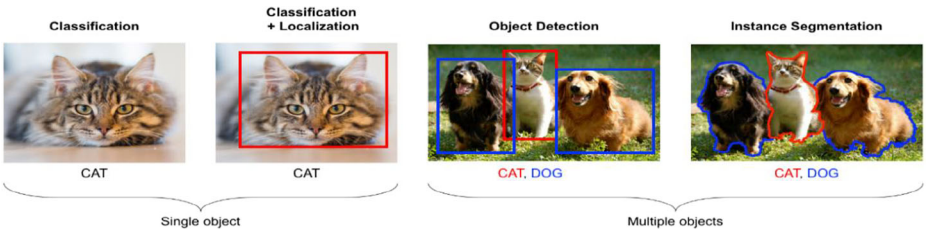

Fig. 1 Single and multiple object detection examples

variations in the images such as appearances, backgrounds, weather conditions, illumination conditions, etc.

Classification Classification is used to predict the class from the given data points. In this process, relevant features are combined to represent the object, compared with the trained model. For visual recognition, classification is required to differentiate the target object from other objects, making the object's representation more semantic and informative. SVM, DPM, AdaBoost etc., are usually used for classification.

1.1 Benefits of object detection

- With increasing technical development, biometrics is important to recognize the individual identity for security. Biometrics authentication is a more reliable method to identify the individual identity. The authentication is performed based on different biological features of everyone, such as fingerprint, DNA, retina, ear, etc. [76, 205]. There are different types of object detection techniques that have been used for biometric analysis in past research work.
- Most of the living areas (metro, parks, schools, shopping centres, etc.) have been monitored by various technical video surveillance because humans cannot continuously monitor video clips. Object detection plays an important role in video surveillance to identify and track instances of a particular object in a scene at once, e.g., tracking the suspected person or vehicle from the video [124, 177].
- In recent years, autonomous robots have been proven to be one of the most interesting research areas. Object detection is the primary task that the robot performs to identify nearby objects and perform some operations such provide information, open-close the door, alarm, etc. [15, 79].
- Human detection is also challenging in computer vision because people as objects have an issue of various appearances and adopt a wide range of poses. Different object detection architecture has been proposed to identify human beings from images or videos, such as pedestrian detection [28]. With the help of object detection, crowd counting has been performed very quickly in dankly populated areas like parks, malls etc. [8].

Object detection has been used to identify an individual face, and it is the first application area in human object detection [133]. Face detection achieves high detection accuracy with minimizing computation time. Face detection has helped to permute object detection with different application areas. Many applications are currently using this idea to detect real-time smiles through cameras, facial makeup, age calculation, etc. [127]. With the advancement of technology, smart vehicle system like the driverless car has been a challenging research area [39]. These smart-systems are required to identify, locate, or track nearby objects and control the vehicle's speed. The object detection system also works on more fine-grained and region-level images like traffic lights detection and recognition [151].

1.2 Limitations of object detection

- Multi-class: Various applications require detecting more than one object class at once. Therefore, the processing speed of detection becomes an important issue as well as multiple classifications without any accuracy loss.

- Some object detection methods or trained models are limited to detecting single class objects with a specific view. They cannot detect the same object with different views, poses, or angle variations [77].
- The object in the image is presented with different aspect ratios and spatial locations. Suppose the object size is very small, like less than 5% of the image, then the system will fail to detect those smaller objects [128]. Sometimes objects are arranged very closely together like a stack of plates, so detecting those objects is very difficult [15].
- Efficiency is an important parameter measured for every object detection system. Some of the developed object detection systems are robust and fast processing, but they require high-efficiency resources (CPU and GPU). The micro-systems have very limited resources for processing; therefore, it is a challenging task [98].
- Object detection models can detect only those for which the model is trained. For example, the ball detection model detects a ball, and sometimes it detects orange as a ball because orange has the approximately same shape feature. The main reason for the detection failure is that the ball object is not included during training [166].

1.3 Motivation for work

- Object detection is an essential part of computer vision applications in the modern era. The detection of objects in various poses, sizes, viewpoints, imaging conditions, illumination etc., is an active research area.
- Several algorithms & frameworks have been developed for object detection in the past 20 years. This study explored and reviewed these object detection techniques.
- Most smart systems automatically recognize objects, such as a driverless cars. These systems can interact with the environment and provide valuable information to humans for making decisions. Google lens scans the object and provides several web links related to the detected object rather than the text input of the object
- Object detection devices such as obstacle recognition and wearable devices assist visually impaired people. These devices are also combined with natural language processing applications to alert and provide information about nearby objects in the native language of the users.

Moreover, the following sections of this research paper provide a detailed analysis of past research work. The whole content is organized into eight sections. Section 2 represents the review method that describes the procedure followed for the planning of the study, research questions, sources of the studies, inclusion and exclusion criteria. Section 3 highlights the steps to prepare the new dataset as well as already existing standard datasets related to object detection. Furthermore, the next section discusses the traditional and state-of-the-art object detectors. Section 5 represents the various performance metrics used in past studies. Application areas of object detection are also reviewed in section 6 to provide an overview for the new researchers. Section 7 highlights the various challenges that occur during object detection processing. Section 8 provides the answers to the research questions. In the last section, summarized findings and future directions are discussed. The whole structure of this research paper is illustrated in Fig. 2.

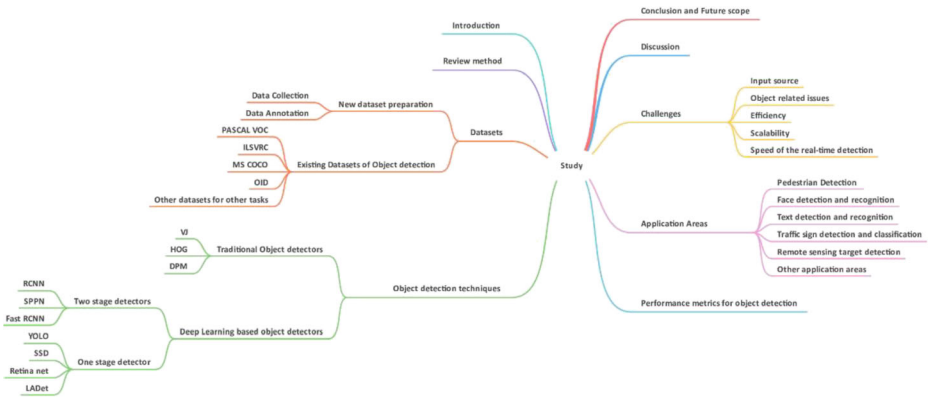

Fig. 2 Structure of our research paper

2 Review method

In this paper, the systematic review of object detection tools, techniques, datasets, and application areas is done in-depth. The steps followed to review and prepare this manuscript are review protocol, research questions, source strategies, inclusion and exclusion criteria for the selection of research studies.

Planning for review Planning is required to complete the work efficiently. Therefore, planning for the systematic review must be the first step. It includes all the further steps that divide the overall work into sub-tasks, such as the collection of content, research topics that should be covered, findings parameters that are highly required to analyze. The research articles are extracted from the various digital libraries based on object detection techniques, annotation tools, evaluation metrics, etc. Thus, some selection criteria are defined to filter millions of research articles. Hense, research questions are defined as a base for searching the papers.

Research question In this step, seven research questions directly related to object detection are defined in Table 1. The research questions are constructed using the following format: the process of object detection, challenges, data labelling, techniques, evaluation metrics, the importance of the dataset, and various application areas.

Sources Selection of the sources for collecting research papers is the most important factor in the quality outcome. So, only conference papers and journal articles are included to prepare the

Table 1 Research Questions of our study

RQ1.	What is object detection? How are objects detected from an image?
RQ2.	Which type of challenges occurred during the detection of objects?
RQ3.	Why annotation is required while training a deep learning model, and which annotations tools are widely used?
RQ4.	What are the different types of techniques mostly used for object detection?
RQ5.	Which evaluation metrics have been used to evaluate the performance of techniques?
RQ6.	How the term “dataset” is important for deep learning? What is the different dataset available for object detection?
RQ7.	What are the important application areas of Object detection?

manuscript. For quality assurance, highly recognized digital libraries "IEEEXplore Digital library", "ACM Digital library", "arXiv open access repository", "Springer", and "Elsevier" are used to collect recent papers. Some other sources are also considered to search the articles, such as reference lists of the primary studies. Search strings such as object detection, localization, object detection dataset, metrics, annotation tool are used to download the research papers. All the relevant papers are further analyzed for the systematic review paper. Different inclusion and exclusion criteria were used to select and discard the studies.

Inclusion and exclusion criteria A step-by-step filtration process is followed to discard irrelevant studies. Initially, collected research papers are discarded based on the keywords, abstract, and title that are not relevant to our review. After that, irrelevant research papers are discarded based on the introduction and conclusion parts of the studies. Further, all these studies are analyzed in detail in the context of the review paper. The inclusion criteria are defined as follows:

- Studies deal with the concepts related to identifying the objects from the image, video, or real world.
- Studies that are relevant to object detection.
- The irrelevant studies according to the research questions.
- Studies that are published in journals or conferences.

3 Datasets

A dataset is a collection of relevant data (i.e., text, image, video) that is prepared in a specific format according to the algorithm's requirements. Artificially trained machines require a mathematically and logically prepared model. These models are developed with the help of the training process. Technique and dataset both are the major components of the training process. The technique defines the learning steps followed by the machine, and the dataset help to learn from the input samples. The performance of the recognition model highly depends upon the dataset quality and volume. These datasets also used to evaluate and calculate accuracy of the proposed techniques [47, 95, 140, 141, 186, 190]. So, this section discusses the steps to create a new dataset and available existing datasets.

3.1 New dataset preparation

Various researchers have developed and distributed several models over the years. Sometimes models are not suitable for performing the specific task due to their trained features. Therefore, training a new model is not an easy task. It requires a sufficient amount of dataset. If the required dataset is unavailable, the dataset must be created according to the specific problem or area of interest. The process of creating a dataset is divided into two steps, as described below.

3.1.1 Data collection

Computer vision algorithms work on data, and the result of these algorithms is purely based on the quality of datasets. Initially, object categories need to be considered for data (images, videos, etc.) collection. Resources such as online websites and manual image capturing using

the camera can be used to collect data. Most of the dataset has contained the images downloaded from the "Flickr photo-sharing website" [35, 155]. Simple query strings return a limited number of images because of the limited use of specific words in that search engine (shows only a few hundreds of images to thousand). Therefore, a large collection of images can be retrieved by expanding the query set or translating the queries into other languages such as Italian, Dutch, Spanish, etc. [26]. While developing the dataset few factors such as lighting, orientation, pose, environmental condition (sunny, rainy, day or night) are considered carefully for inclusion and exclusion of samples [38]. Data cleaning step is required after the collection of a sufficient amount of data. The data quality affects the model's accuracy and performance. Perfect data collection is a challenging task for large models. Therefore, collecting data from different sensors or cameras is not directly used for training. Some pre-processing tasks are required to clean or improve the data's quality.

3.1.2 Data annotation

Images stored in the dataset with additional formal representation helps to maintain and scale the database very easily. The formal representation is metadata that comprises various information such as date, location, physical properties, symbolic description, etc. This information is presented in a free or constrained format [173]. In the object detection and recognition research area, a large collection of data is required with the ground truth labels, and further data is used to perform learning and evaluation. Labelling an object in the image provides information such as shape, location, identity, and other information like pose. Building a large dataset is costly as well as lengthy due to annotations of many objects in images. Therefore existed dataset contains a limited class of data [145]. In the annotation process, the bounding box is drawn around the identified objects and stores that object's properties [110]. The quality of annotated data is key to the success of any computer vision project. For example, if a person shows a pencil to a child and says it is a pen, then after some time, the child sees a pencil, the child will classify it as a pen. In the same way, the machine learns by example (input samples). Therefore, the result of the object detection model depends on the annotated labels that are fed during the training phase. If the erroneous data is fed into the model during training time, then the result of the trained model is also wrong.

Bounding Box, semantic segmentation, polygon segmentation, line, splines, etc. annotations are used to annotate the images. Bounding Box is the most commonly used annotation to define the boundary or location of the objects in an image. Most of the objects are not rectangular, so a polygon shape is used to locate that type of object. Semantic segmentation is a pixel-wise annotation means in this every pixel is classified in the image. Line and Spline: annotate the object using lines and spline, for example, in lane detection. Therefore, many annotation tools are used to annotate the object in images according to the object properties. For example, LabelImg is used to annotate indoor objects [1]. Different annotation formats (such as JSON, XML, plain text) are used by existing techniques, such as COCO, PASCAL VOC, and YOLO. Some authors have used the annotation pipeline in which annotators, inspectors, examiners perform all annotation processes [155]. Different annotation tools have been developed in recent years, which are free or require a premium licence. Some of the tools are briefly described in the next section:

- **MakeSense** AI is an open-source and free annotation tool that has been used under the GPV3 Licence. Therefore, there is no need for the installation of this software. In addition,

this tool support multiple label types such as rects, lines, points, and polygons and many outputs file formats such as YOLO, XML, VGG, JSON, and CSV [23].

- **LabelImg** is an image annotation tool that generates the output in XML files. It is free and open-source software used to label the image graphically. PASCAL VOC, YOLO, supports exported file formats [175].
- **VOTT** (Visual Object Tagging Tool) has been used to label images as well as video frames. Microsoft developed this annotation tool, and it is freely available. This tool is written in TypeScript programming, and it can maintain data from Cloud and local storage [182].
- **VIA** (VGG Image Annotator) is offline software used to perform manual annotation on images, audio, and video. This software is very lightweight and does not require additional software libraries or installation. Web programming languages named CSS, JavaScript, and HTML are used to develop this tool [32].
- **KAT** (K-Space Annotation Tool) was introduced during the K-Space project, supporting structural and descriptive annotations [24]. This tool provides flexible and low-level semantic annotation using the COMM, i.e., Core Ontology of Multimedia. It also provides an API and other services used with the help of plugins [147].
- **PhotoStuff** is proposed by the Mindswap group, which was initially used to annotate images. The descriptive and structural annotations are primarily discussed in the PhotoStuff [24]. It provides the ability to browse, search, and manage image annotations from the semantic web portal [53].
- **AktiveMedia** is proposed during the projects such as AKT and X-Media. Labelling an image or text can refer to a region or image-level [24]. It is an ontology-based annotation, and the main objective is to automate the annotation process by minimizing user efforts. The entire system has been worked in the background by interacting through web services and context-specific knowledge queries sent to the central annotation store.
- **Caliph and Emir** developed Java-based applications that have been used for image annotations and retrieval. MPEG-7 XML files are generated, searched, and retrieved with the help of the user interface in this application. Caliph (Common and lightweight photo annotation) tool has been used to perform manual annotation of images. It is also used to extract low-level features and existing metadata automatically. The Emir (experimental meta-data-based image retrieval) tool performs the linear search to retrieve the MPEG-7 document collection [105].
- **LabelMe** is a database and digital image annotation tool that helped in many research areas of computer vision. It is a free, open annotation tool that helps to create a large image database. LabelMe is developed at MIT CSAIL (Computer Science and Artificial Intelligence Laboratory. The annotation files are exported in XML file formats [145].
- **LabelBox** is a versatile labelling tool that annotates digital images using the line, polygon, rectangle, etc. The LabelBox annotation tool is paid and provides a platform for data management, labelling, and data science [83].

There are some other tools that are used to annotate the objects in an image, such as RectLabel (jamtsho), maskEditor, GTTool, ByLabel (qin2018), SWAD, M Ontomat-Annotizer, PerLA, Scalable etc. The annotation tool is selected according to the annotated file format required for training, easiness, and licence.

3.2 Existing datasets

Over a decade, numerous datasets and benchmarks have been released, like PASCAL VOC challenges, ILSVRC challenges, etc. The most commonly used datasets are discussed in this section.

3.2.1 Pascal VOC

PASCAL VOC has been a popular dataset in computer vision technology that helped develop and evaluate new algorithms. Those algorithms perform variety of tasks like object detection, image classification, action detection and segmentation [34–36]. PASCAL VOC challenges were held between the years 2005 to 2012, and different versions were developed each year, but two of the highly used versions are VOC07 and VOC12. All the images were gathered from the existing data source or Flickr. VOC 2005 detected only four classes, i.e., motorbikes, bicycles, cars, and people. But the PASCAL VOC 12 dataset consists of 11,530 training images and 27,450 annotated objects to detect 20 object classes from images. The 20 objects classes (Person: person; Indoor: chair, dining table, sofa, potted plant, tv/monitor, bottle; Vehicle: bus, bicycle, car, boat, motorbike, aeroplane, train; Animal: cat, dog, cow, bird, sheep, horse) have been selected and annotated for creating the dataset which is common in human life.

3.2.2 ImageNet large scale visual recognition challenge (ILSVRC)

ILSVRC is a computer vision challenge, and it was held every year between 2010 and 2017 [144]. ILSVRC included publicly available datasets, challenges, and related workshops. It collected images from ImageNet (used the WorldNet) Hierarchy for detection challenges [26]. In 2010, it started only on the image classification challenge of 1000 categories. With the change in time and technology, challenges have also grown. In 2017, the dataset was developed to detect 30 fully labelled categories differentiated based on movement type, level of video cluttered, average no. of object instances, and several other factors. The image dataset contains more than 15 million images of high resolution [144].

3.2.3 Microsoft common objects in context (MS COCO)

MS-COCO is one of the challenging datasets for object detection, image captioning, and segmentation [178]. The competition of dataset MS-COCO has been held every year since 2015. In contrast to the dataset ImageNet, it contains a smaller number of objects but has more object instances. In MS-COCO, the annotation file is stored in a JSON file. Despite bounding box annotations in the COCO dataset, the instance segmentation of objects has been used for precise localization. MS-COCO dataset has very small and dense objects compared to the above-discussed dataset. These feature of MS-COCO helps to close the real-world objects.

3.2.4 Open image dataset (OID)

The Open Images dataset V4 contains 9.2 million images with unified annotations for three tasks as visual relation detection, object detection and image classification [82]. All the images have been downloaded from Flicker without the use of prefined class names. This dataset

contains 600 object classes with 15.4 M bounding boxes, 57 classes with 375 visual relationship annotations and 30.1 M image-level annotations involving 19.8 k concepts. Dataset provides accuracy and consistency because all the objects were labelled by professional annotators. All the images of this dataset are of high quality and contain several objects.

3.2.5 Other datasets

In the above sections, all the existing datasets are directly related to object detection. But some other datasets are interlinked with the object detection. Table 2 shows some of the other popular datasets used in the past research work like face detection, text detection, pedestrian detection, traffic sign, light detection, and remote sensing target detection.

4 Techniques for object detection

Object detection is one of the powerful areas of computer vision, and its main aim is to locate the object and classify them in the given image. The object detection framework is divided into two types of categories. Figure 3 shows the categorization of various object detection techniques into two groups: (i) traditional detectors and (ii) deep learning-based detectors. Deep learning detectors are further divided into sub-parts such as "Two-stage detector" and "One stage detector".

4.1 Traditional detectors

Manually crafted features were used in traditional object detection techniques. Most of the researchers only used the complex features vectors because advanced image representation methods were not developed completely, and the computational resources were limited at that time. Traditional detectors identify the object using approaches like VJ (Viola-Jones) detector, HOG, SIFT, etc. Traditional object detection methods were successful because features descriptors were carefully designed to achieve the regions of interest in the image. The remarkable result was attained on the Pascal VOC dataset using feature representation and classifier [192].

4.1.1 Viola-Jones detectors (VJ).

The first traditional object detection technique was VJ Detector to detect the human face 21 years ago. This technique has been performed face detection without constraints [181]. VJ

Table 2 Standard datasets and the referenced studies

Sr. No.	Datasets	Count	Citations
1.	Pedestrian detection	10	[12, 13, 21, 28, 29, 45, 126, 131, 185, 211]
2.	Face detection	8	[77, 78, 87, 113, 122, 201, 202, 206]
3.	Text detection	19	[14, 49, 65, 70, 74, 103, 104, 119, 121, 143, 149, 150, 154, 157, 162, 178, 183, 203, 207]
4.	Traffic light and sign detection	9	[6, 25, 30, 33, 38, 62, 120, 172, 220]
5.	Remote sensing target detection	11	[20, 60, 84, 89, 94, 96, 138, 169, 196, 219, 221]

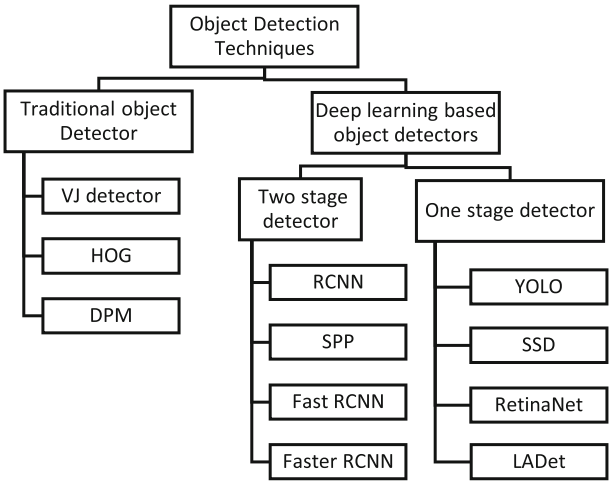

Fig. 3 Classification of various object detection techniques

uses a straightforward approach for detection, such as a sliding window. In this technique, all the pixel locations from the image are checked and scaled, so detecting the human face in the window becomes easy. At that time, this algorithm has detected the faces without any kind of constraints [133, 180, 181]. The name of this technique is given to the author, i.e., "Paul Viola and Micheal Jones" because authors of these techniques significantly contribute to the area of face detection. It has a straightforward process, but the calculation behind this process was too away from the computer power of that time. As a result, the detection speed is fast, but the training for this technique is slow. This entire process has divided into three steps:

Integral image This is a computational method, but it has been used to speed up the performance of the box filtering. For calculating the rectangular features, the integral representation of the input image acts as an intermediate form for further steps. VJ detector has used Haar wavelet for feature representation of an image.

Feature selection Integral representation has been used to select the small and key features. Adaboost procedure has been performed the searching to select the small number of good features.

Detection cascade This is the last step of the process that uses cascade classifiers to discard the background region of an object from an image. More computations are worked upon an object rather than on the whole image. Therefore, it reduces the computational overhead on the background and focuses only on the object (faces).

4.1.2 Histograms of Oriented Gradients (HOG)

In 2005, N. Dalal and B. Triggs [22] had proposed the HOG (Histograms of Oriented Gradients) descriptors with improvements in the SIFT (scale-invariant feature transform) and shape context. It creates the histogram of edges using patches, where a patch contains anything such as it may be a person, any meaningful background, an object, etc. In this technique, pre-

processing is performed on selected patches such as translating to the fixed aspect ratio, i.e. 1:2, and if the patch size is larger than 64 × 128, then it resizes to it in that size. Gamma correction is used in pre-processing. This method calculates the histogram of gradients, i.e., vertical and horizontal gradients are calculated. It can be done with the use of the Sobal operator, and after that magnitude and direction of gradients can find out by using the following formula as $g = \sqrt{g_x^2 + g_y^2}$ and $\theta = arctan\frac{g_y}{g_x}$ non-essential information is removed with the help of the gradient's image and highlights the outlines. If the input image is coloured, then three channels of colours are evaluated. Therefore, gradients are calculated for three colour channels on each image pixel, and the maximum gradient from the three-channel and corresponding angle is selected for further steps. A histogram of gradients vector is created and normalized to the vectors to take the final result. After that, information is forwarded to machine learning algorithms to train the classifier, such as SVM used to train the classifier. Non-Maximum Suppression (NMS) is applied to remove the object's maximum redundant bounding box. HOG is used for the detection of different class objects. It detects all distinct sizes of different objects by using multiple times of re-scaling of the image as well as keeping the window size unchanged. HOG has been highly used for the detection of pedestrians.

4.1.3 Deformable Part-based Model

In 2008, DPM (Deformable Part-based Model) was proposed by P. Felzenszwalb et al., which was the apex of traditional object detection methods [37]. DPM is the extended version of the HOG Detectors. DPM follows the divide and conquer strategy for the detection process. Here divide is just like training, and conquer is inference. In training, learning has been performed so that it helps to decompose the object into parts. After detection of the parts of the object, results are combined to compute the overall inference. This model consists of three components, i.e., root filter, part filter, and spatial model. A root filter is a detection window that approximately covers the entire object. Therefore, a filter with some specific weights is used for the "region feature vector". Part filter: Multiple-part filters are used to cover the small parts of the objects in an image. A spatial model is used to score the part filters locations relative to the root.

Further, this work is extended as the 'Star Model' done by P. Felzenszwalb, D. McAllester, and D. Ramanan [37]. After some time "Star model" is converted into a mixture model by the author R. Girshick to detect real-world objects under different but significant variations. In the mixture model, weakly supervised learning has been performed to configure the part filters. In other words, part filters have been automatically learned by using weakly supervised learning. PASCAL VOC gave a "lifetime achievement" award to P. Felzenszwalb and R. Girshick in 2010.

4.2 Deep Learning based object detectors

With the advancement of technology, object detection has become one of the main research areas of computer vision. Before 2010, Object detection has done through the classical algorithm. Traditional detectors had some limitations, such as a huge amount of proposals generated, from which many proposals were redundant. These redundant proposals resulted in many false positives [192]. Therefore, when the CNN architecture was proposed, it gave a big

change in the area of object detection and the deep neural network. The deep neural network has been used for feature representation of the image on a deep level so that the detection of an object is done effectively with less error rate. Deep neural architecture takes lots of time because it has a huge amount of data for training. Different types of the backbone are applied to detect objects, such as AlexNet, VGG Net, GoogleNet etc. [209]. Object detection techniques based on deep learning are divided into *two-stage and one-stage detectors*. The *two-stage detectors* used two stages to detect the objects from the image, and these detectors often provide state-of-the-art results or high accuracy on available datasets. But these detectors have a lower inference speed as compared to one-stage detectors. The one-stage detector is mostly used in real-time object detection and provides the desired result much faster than two-stage detectors.

4.2.1 Two-stage Detectors

In this, two stages have been used to detect the objects, such as proposal generation and, from these generated proposals, make predictions about the objects. In the first stage, regions of all the objects are identified using detectors. The image detector generates the regions which have a high recall rate; further, objects belong to at least one of these generated regions. In the second stage of these detectors, classification is performed by using the deep learning models. The generated regions either contain the objects of predefined labels or a background. Despite this, localizations may also be refined by the model proposed in the first stage. Further, this section describes the prominent two-stage detectors techniques.

Region-based Convolutional Neural Network (RCNN). R. Girshick, J. Donahue, T. Darrell, et al. have used CNN to make a new object detection technique, RCNN, a two-stage technique [47]. *RCNN* algorithm performs three steps for object detection: extract region, compute CNN features and classify the region. In the Region Extraction step, 2000 cropped and wrapped regions were generated using the selective search approach. Objects are present on any scale, so the selective search approach finds all objects on all scales with fast computation speed. Similar regions are grouped based on size, texture, colour, and shape. Compute CNN: After selecting regions, each region is resized to the fixed size and sent to CNN for feature extraction. Classification of the region: SVM is the score used for each class and applies NMS (Non-Maximum Suppression) on each class for rejecting those regions whose IOU (Intersection Over Union) has larger than the learned threshold.

Therefore, the traditional detectors used the hand-crafted features descriptor to detect the object from the image. In the comparison of this, the deep neural network produced the hierarchical features from the image, and different layers of different scale information is captured for object detection. As discussed in the previous paragraph, Fig. 4 displays the steps that are performed in the R-CNN architecture. This method is used to classify the object as well as show the bounding box over the detected object. In addition to this, RCNN faces some limitations, such as (i) using a deep convolution network, features are extracted from each proposal separately, making many duplicated computations. Therefore, training and testing of RCNN was a very time-consuming process (ii) all the three steps of this technique were independent which has run independently makes it difficult for getting the globally optimal solution (iii) in complex image backgrounds, selective search does not perform well for generating the proposals because selective search uses only low-level signs.

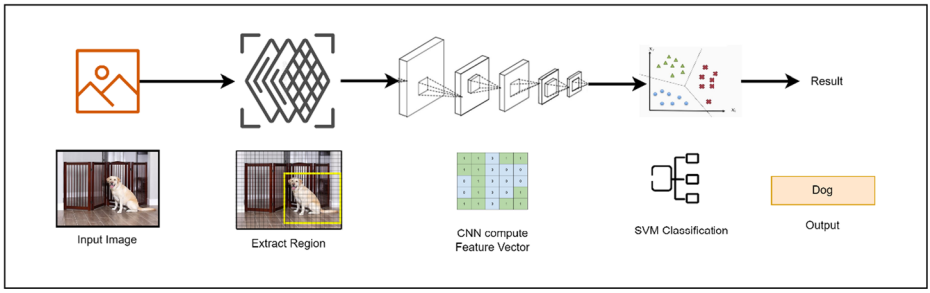

Fig. 4 Basic steps involved in R-CNN (Region-based Convolutional Neural Network)

Spatial Pyramid Pooling Networks To overcome the drawbacks of RCNN, K. He, X. Zhang, S. Ren et al. read and considered the theory of spatial pyramid matching (SPM) to introduce a new network structure SPP-Net in 2014 [57]. In RCNN, CNN is applied to a fixed size input image which makes this requirement is artificial. Some objects are cropped or distorted because of fixed size constraints. Their requirement is removed in SPP Net for improving the accuracy. The new network structure uses the SPP layers, which generate the predetermined length vector representation accepted by the fully connected convolution layer. The whole image is used only once in this detector to extract the features, unlike RCNN.

With the use of the deep convolutional network, a feature map is computed from the whole image. Further, the SPP layer used the feature map to extract fixed-length feature vectors. Feature maps are divided into three fixed-size grids, i.e., $M \times M$ by SPP layer as shown in Fig. 5, for the multiple values of M. Feature vector produced from each cell of the grids. After that, all the feature vectors are joined together and fed to the SVM (Soft Vector machine) classifiers and Bounding Box regressor. The accuracy of this technique is the same as RCNN, but the speed is 20 times faster than RCNN. It has some drawbacks like RCNN; the first is that it ignores all previous layers before the fully convolution layer, and the second is that training in SPPNet is multistage. Because it works on multi-scale and different aspect ratios, this technique works well without loss of information and undesirable distortion.

Fast Region-based Convolutional Neural Network Further improvement of RCNN and SPPNet is done by Girshic and introduced a new architecture shown in Fig. 6 named Fast RCNN. It takes input as a whole image and considered for producing the feature maps using the convolution layer [46].

It removes the multi-level pooling layer and uses only one single layer grid. The "Region of Interest (ROI)" pooling layer is used to a fixed-length feature vector is extracted and used with the specific case of the "Spatial Pyramid layer (SPP)" which used only one pyramid layer. Every feature vector is processed by the sequence of the "Fully Convolution (FC) layers" before the output layer, and the final FC output is considered into two sibling layers, i.e., Softmax and Bounding Box. The softmax layer has been produced probabilities for all objects' classes and one background class of the object. BBOx layer has generated the four real values for making the box around the predicted object. This technique does not perform separate training for the classifier and BBox regression. Fast RCNN is implemented using python and C++ (using Caffe) [46]. It used all the advantages of RCNN and SPP-Net, but its speed is a little bit slow because of proposal detection. But it saves extra storage space charges improves accuracy and efficiency.

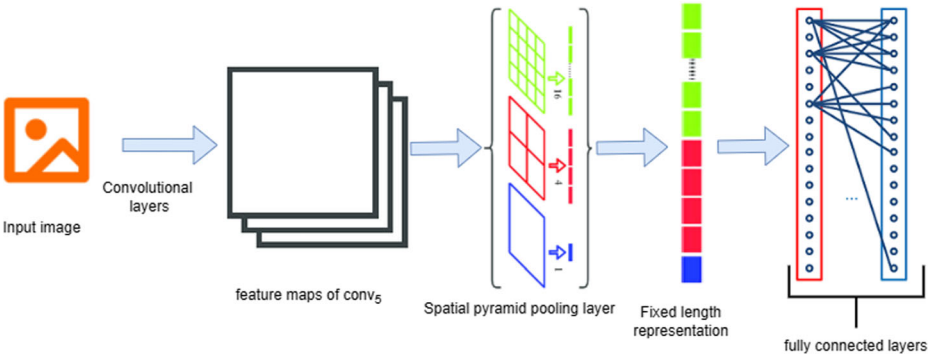

Fig. 5 Network structure of SPPNet

Faster Region-based Convolutional Network method Different methods are used to generate the candidate boxes like selective search, edge box, etc. However, these methods are not improving the efficiency of the object detection approach. Therefore in 2015, S. Ren, K. He, and R. Girshick et al. introduced a new method RPN in Faster RCNN, for generating the regions [142]. RPN uses the sliding window approach on the feature map for generating the bounding box on each object along with the score of the bounding box, as shown in Fig. 7.

This generated bounding box with a common aspect ratio is called an anchor box. After bringing down the proposals into fixed size, they are transferred to the full convolution layer, i.e., the Softmax and regression layer. For increasing the non-linearity in the output ($n \times n$)Convolution window, Relu is applied over the output window. With this new architecture end to end, training of object detection algorithm is achieved. By revising the architecture of these techniques, authors detect the objects [55].

Mask RCNN is a two-stage procedure based on a deep neural network, and the main objective is to solve the instance segmentation problem [59]. It is an extension of Faster RCNN detectors in which a mask network branch module is added for ROIs (Region of Interest) segment prediction. Both object detection as well as for instance segmentation, can be done by this technique simultaneously. This technique isolates the various objects from video and image. In simple words, it takes the image as input and provides the output with detected objects with masks, classes, and bounding boxes. This network has two stages; the first stage accepts the input image and generates the region proposals where the object might have existed. In the second stage, the class of objects is predicted, the bounding box refines and creates the mask of an object using pixel level. These two stages of mask RCNN are connected with FPN backbone architecture.

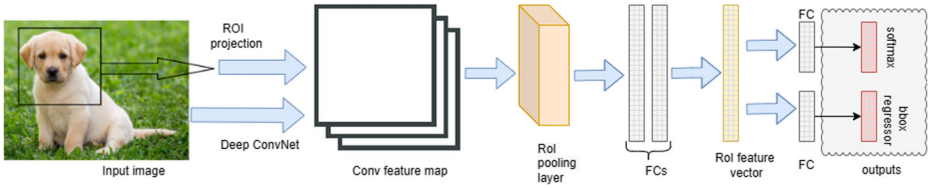

Fig. 6 Architecture of Fast RCNN

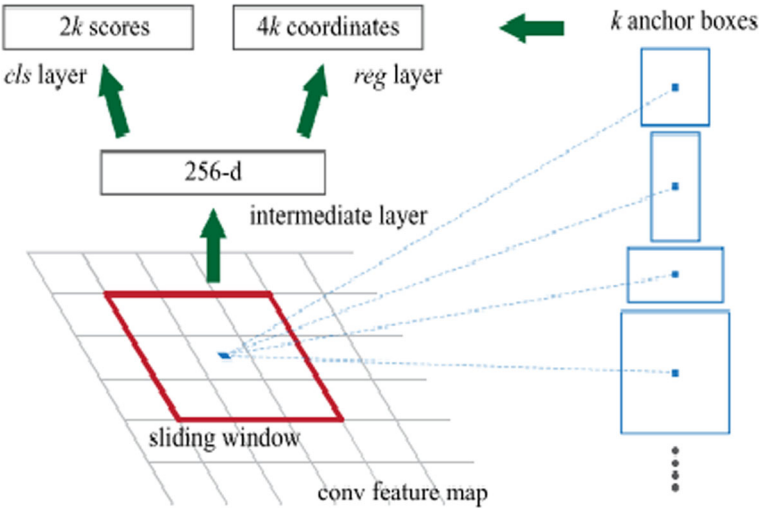

Fig. 7 Region Proposal Network (RPN) [142]

4.2.2 One-stage detector

No intermediate task is performed in one stage detector for taking the final output. Therefore, it leads the faster and simpler architecture. For example, YOLO, RetinaNet, SSD etc., are the one-stage detectors that are used to assign a bounding box to a specific position. So, detectors perform learning on certain object locations.

You Only Look Once In 2015, the first stage detector YOLO was introduced by the author J. Redmon, S. Divvala, R. Girshick et al. [141]. YOLO (You Only Look Once) follows the different methods for detection and verification. The architecture of YOLO is shown in Fig. 8. This technique uses the fixed grid detector, which makes the technique fast. A single neural network is applied to the whole image to detect objects in this technique. The whole image is divided into fixed regions, from each region, the probability and bounding box of the object is calculated. In Yolo, single CNN was used to predict the class probabilities on multiple bounding boxes. In this, training is performed on whole images. YOLO (You Only Look Once) predicts objects based on the convolutional layers that use the S × S grid system. These individual grids on the input image are responsible for detecting objects and predicting the boundaries of the object.

The various versions of YOLO are YOLOv2, YOLOv2 tiny network (ear detection) [205], tiny-YOLO-voc1(forest fire detection [191]), YOLOv3 [140], YOLOv4, YOLOv5. YOLOv1 has the limitation of small object detection and the worst accuracy if the input image size is

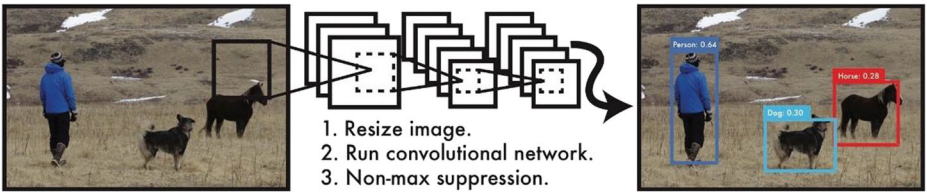

Fig. 8 Architecture of YOLO Detection System [141]

different from the training images size [141]. Therefore, YOLOv2 is introduced with Darknet-19 architecture with 19 convolution layers and 5 max pool layers, and one SoftMax layer [139]. Batch normalization reduced the overfitting by altering the scaling activation function. Image resolution increased from 224*224 to 448*448, and a feature map of 13 × 13 improved the accuracy to detect smaller objects in YOLOv2. In addition, the Region Proposal Network and Single Shot Multibox Detector (SSD) improved the performance in the YOLOv2 version. YOLOv3 is based on Darknet-53 architecture and improved the accuracy for smaller objects compared to earlier versions of YOLO [140].

In YOLOv3 logistic classifier is used instead of the SoftMax function that provides the benefits of a multi-label classifier in object detection. The key novelty of the YOLOv3 is that it makes object detection at three different scales. In YOLOv3, Feature Pyramid Network (FPN) allows the network to learn objects of different sizes. The EYOLOv3 is used to detect traffic objects from video [161]. With the advancement in YOLO, the updated version of YOLOv4 has been proven to be the fastest object detection and improved the mAP by 10% [9]. With the advancement in YOLO, the updated version of YOLOv4 has been proven to be the fastest object detection and improved the mAP by 10%. Yolov4 has CSPDarknet53 feature extraction architecture used to split the current layer into two distinct parts. One for passing every input through the convolution layer and the other without passing through the convolution layer. In the end, the results of both parts are aggregated. Spatial pyramid pooling is introduced to improve accuracy by separating out the most significant context feature in YOLOv4. The latest version of YOLOv5 is the translation of the Darknet framework to the PyTorch framework [81, 125]. Darknet is primarily written in C and offers much control over the network's operations. But PyTorch is written in the Python programming language and provides easy control of low-level operations with more configuration. Data augmentation, auto-learning bounding box anchors, light-weighted models, and fast performance are the key features that are introduced in YOLOv5. The release of YOLOv5 includes four different sized models, (i) YOLOv5s (small), (ii) YOLOv5m (medium), (iii) YOLOv5l (large) and (iv) YOLOv5x (extra-large).

Single Shot MultiBox Detector W. Liu, D. Anguelov, and D. Erhan et al. introduced a new technique, SSD (Single Shot MultiBox Detector*),* in 2015 [95]. It is a second technique of the one-stage detector in the deep learning era. For improving the accuracy of small object detection, SSD used multi-reference and multi-scale representation. The only difference between the SSD and the former object detection technique is that SSD runs only on the top layers, and the former run-on different layers of five different scales. SSD is also used for real-time fire detection with high detection accuracy [191]. SSD architecture is divided into two parts, i.e., the backbone model and the SSD head. The first part is the backbone model, a pre-trained classification network that acts as a feature map extractor. The second part is applied on top of the first part, in which several convolutional layers are stacked together. This part gives the output as the BB (bounding box) over the detected object. Various objects from the image are detected with these convolutional layers. It is an efficient and fast object detection model for the detection of multiple categories. With time, tiny SSD has been introduced to provide reliable performance compared to tiny Yolo on the VOC07 dataset [190]. The authors can use SSD to optimize the algorithm to detect objects such as vehicles and wheels using optimized SSD [41]. Deconvolutional Single Shot Detector (DSSD) has an extended version of faster RCNN in, which ResNet-101(Backbone) adopted.

Further, this technique adds two modules, i.e., the prediction and deconvolution module [42]. In the prediction module, the residual block is added with each prediction layer, and element-wise addition is performed on the output of the residual block and prediction layer. With the deconvolution model, the resolution of feature maps increased to strengthen features. The different kinds of objects of different sizes are predicted in the deconvolution layer. The ResNet-101 backbone architecture was selected for the model's training on the ILSVRC CLS-LOC dataset. The experiment is performed to show the effectiveness of the proposed model on the PASCAL VOC and MS COCO datasets. With deconvolution and predication module, 2.2% enhancement is brought on the VOC2007 dataset. Researchers have been trying to improve the detection accuracy in the real-time environment while applying some improvements to the existing detectors. Lu et al. have presented the AF-SSD (Attention and feature fusion SSD) to solve the problems caused by complex backgrounds, scale variation, and small objects [102]. This structure used the MRF (Multiscale Receptive Field) module of CNN that shows the region's size in which pixels of the feature map are recorded on the original image. The receptive field size affects the information of feature maps means a large size has more global information and small size has more detailed information on the feature map. Thus, this module has been designed to broaden the receptive field to capture multiscale features.

Retina net T. Y Lin, P. Goyal, R. Girshick, et al. have presented a new loss function in replacing the existing standard cross-entropy loss for handling the class imbalance issue during the model's training [93]. Compared to two-stage detectors, the accuracy of the one-stage detector is slower. Therefore, RetineNet has proposed a single-stage object detector that runs fast and provides accuracy on dense as well as small-scale objects. It is a single unified network with a backbone architecture and two subnetworks to perform object detection at multiple scales. The backbone architecture accepts any input image size and computes the convolutional feature map on input. Further, object classification is performed on the output of backbone architecture by the first subnetwork, and the second subnetwork performs bounding box regression. Two sub-networks are used for increasing the performance of a one-stage detector, as shown in Fig. 9.

RetinaNet has four main components, as shown in Fig. 9, i.e., classification subnetwork, regression subnetwork, top-down path, and Bottom-up path. The first one is the Bottom-up path in which ResNet is chosen as a backbone architecture to compute the feature maps at different image scales. FPN (Feature Pyramid Network) is used for lateral connections and top-down pathways in the second part. This helps to construct a multi-scale and rich feature pyramid from any size of a single resolution image. After that, the third part of RetinaNet is the classification subnetwork that predicts the probability of each detected object class and anchor box. The last part calculates the offset value of the BB from the anchor boxes. The Focal Loss (FL) function

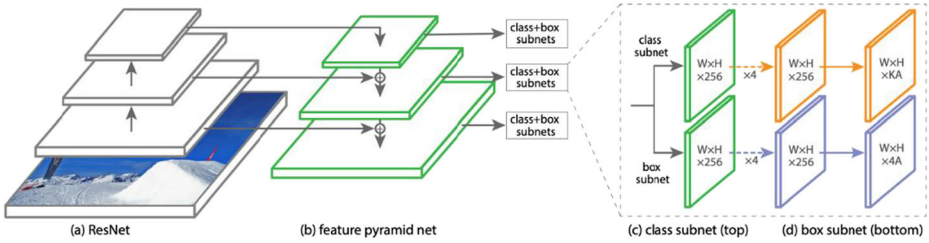

Fig. 9 Architecture of one stage detector RetinaNet with FPN [93]

has a high detection speed as well as maintains accuracy, just like two-stage detectors. The bigger object, the more stability is provided by RetinaNet [128]. FL improves the Cross-Entropy Loss proposed for handling the foreground-background class imbalance problem that occurred in one stage detector model. The proposed network have thousands of anchor boxes at each pyramid layer. It is widely applied to aerial and satellite images to detect objects. Sun et al. have developed a new end-to-end detector, R^4 ("Refined Single-stage detector with feature Recursion and Refinement for Rotating objects"), that uses RetinaNet as a base network to detect the objects from dense distribution, larger aspect ratio, and category imbalance [163].

Lightweight and Adaptive Network for Multi-scale Object Detection In 2019, A light-weight and Adaptive Network for Multi-scale Object Detection (LADet) was proposed by J. Zhou et al. for handling the Scale variation challenges of object detection [216]. LADet has been performed object detection with the use of two modules such as AFPM and LCFM, i.e., Adaptive Feature Pyramid Module and Light-weight Classification Function Module, respectively (architecture of LADet has shown in Fig. 10). The proposed architecture uses the DenseNet-169 as a backbone architecture for extracting the multi-level feature from the input image. Further extracted features fed into SFPM in which complementary semantic information was generated from the feature maps. This module is further divided into two subparts, as shown in Fig. 10, i.e., FFM (Feature Fusion Module) and ACFR (Adaptive Channel-wise Feature Refinement). In the FFM, all the feature maps of pyramid levels are scale normalized to the same resolution. Further concatenation is performed on multiple outputs of scale normalization for fusing the feature maps.

Then, the ACFR model generates the complementary information on the output of FFM. Classification scores and dense bounding boxes are generated based on learned feature maps. The AFPM has used two convolution layers that contain structure-sparse kernels. This module also applied the permutation operation between these layers to estimate "high-rank kernel" or "original dense". The classification module predicts the probability of the presence of the detection object and class. This technique achieves better accuracy and speed in comparison to previous techniques. There are multi-level FM extracted from the backbone network fed into the AFPM for generating the pyramid FM with paired semantic information. After that, LCFM (classification subnet) and Box regression subnets are applied.

Over the past decades, the progress of object detection has been widely accepted in the real world to solve or help human beings. Therefore, Table 3 gives a brief overview of several object detection techniques and their strengths, metrics, and limitations.

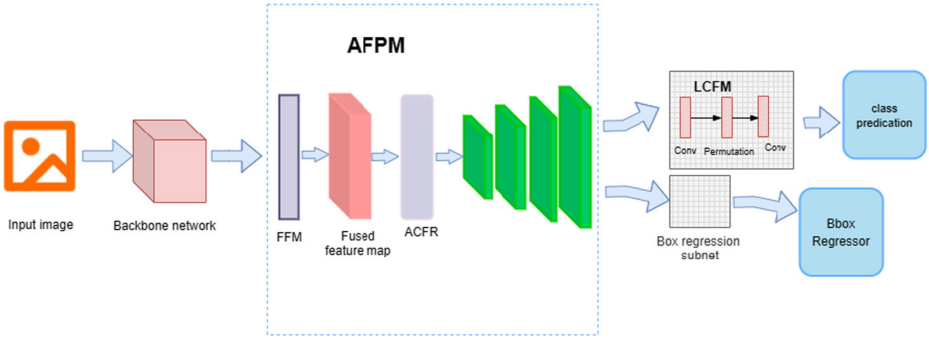

Fig. 10 Overview of LADet model

Table 3 Brief overview of object detection techniques

Technique	Strengths	Limitations	Input Size	Dataset	Results
Vj detector [181]	Use Integral image, Adaboost, and cascade classifiers for face detection	Work on the grey-scale image, and the detector fails on occluded faces.	384×288	MIT+CMU the test set with 130 images and 507 faces	Primarily focused on face detection, detect only frontal, an upright face which is 15 times faster than previous technique, achieves more than 90% detection rate on several false positives.
HOG [22]	On a large scale and small scale, it provides more global and fine-grained detail, respectively	Cannot detect different parts of the human body in a different pose, slow processing time and high memory consumption.	64×128	MIT pedestrian dataset with 509 training and 200 test images, INRIA contains 1805 images of 64X128	It used hybridization of SVM with HOG to detect the whole human body from images, achieved an 89% detection rate, developed a new pedestrian database named INRIA
DPM [37]	Detect the human with varying poses, partially visible objects, and deformable objects by processing the image in 2 s	Training and detection of multiple classes	–	PASCAL VOC 2006 and 2007 dataset	The system performs well on rigid objects such as cars and sofas, and deformable objects such as persons and horses, performed well on 10 categories from 20 of 2007 challenge, achieve.34 AP
RCNN [47]	Use the selective search, and the CNN architecture say AlexNet and SVM is used to classify the objects	Slow detection speed, fixed selective search algorithm so that learning is not happening at this stage, proposals need to be wrapped to be the fixed size; training pipeline is not end-to-end (CNN, SVM), only accept the size of image 227×227	227×227	PASCAL VOC 2007 and VOC2010	Take only interesting objects from different spatial locations within an image and with different aspect ratios; processing time on the image is 10S/ on GPU (NVIDIA Titan Black) and 53S/ on CPU; achieve mAP:62.9% (VOC 2010), 66.0% (VOC 2007), 31.4% on ILSVRC2013; average accuracy of segmentation is 47.9%
SPP Net [57]	ZF-5 backbone architecture is used, Flexible for handling input images of different scales, sizes, and aspect ratios.	Training is inefficient due to multistage pipeline, high memory consumption, compared with RCNN speed is better than but training is slow.	~1000×600	PASCAL VOC 2007	Variable size of the input image is accepted; FPS is less than 1, and mAP is 59.2% on VOC 2007.

Table 3 (continued)

Technique	Strengths	Limitations	Input Size	Dataset	Results
Fast RCNN [46]	VGG-16 backbone architecture is used, one fine-tuning stage, fast training and detection, scale invariance, and efficient backpropagation	To find ROI, it uses selective search, which is a slow and time-consuming process, for real-time application is too slow.	~1000×600	VOC 2012, VOC 2010, VOC 2007	Multi-task loss training is done in a single-stage; in feature caching, no disk storage is required; compared with SPPnet, it is much faster; a new pooling layer RoI was proposed, the first end-to-end trained detector. Achieve maP 70.0% on VOC2007, 68.8% on VOC2010 and 68.4% on VOC2012; FPS is less than 0.5
Faster RCNN [142]	Used RPNs that make the region proposals efficient and accurate; learned RPN also improves region proposal quality; VGG-16 (ResNet-101) used as the backbone	It takes much time in object proposals, and it requires many passes over an image to extract all the object's features from the image; still, speed is slow; due to RPN still running slow.	~1000×600	VOC2007, VOC2012	Use the Region Proposal Network (RPN) instead of selective search for regions. Use anchor box, FPS is 7; it is best is speed as well as quality use of RPN; achieve mAP 73.2% on VOC2007, and for VOC2012 is 70.4%
YOLO [141]	Used GoogLeNet as the backbone, and it takes the parts size of $s*s$ of the image, which has a high probability of containing the object; Significantly faster than previous detectors.	Difficult to detect small objects and objects which are close to each other; the aspect ratio is fixed.	448×448	PASCAL VOC 2007 and VOC 2012	The whole detection pipeline is a single network, both the classification and bounding box regression did simultaneously. As compared with other detectors, accuracy falls. Achieve mAP 63.4% on VOC2007 and 57.9% on VOC2012; FPS is 45.
SSD [95]	First unified network to perform well and generate accurate results; detector detects objects in multi-scale layers. Use VGG-16 as a backbone network, which is more accurate than YOLO	Detection of smaller objects has a worst performance than the larger objects	300×300	VOC2007, VOC2012, and COCO	By eliminating bounding box proposals and the subsequent pixel in the resampling stage that Improves its speed, FPS is 46. Achieve mAP on 77.2 VOC2007, 75.8 VOC2012 and 23.2 COCO
YOLOv2 [139]	Proposed backbone darkNet19	Not good for detection of small objects; 9000 object categories are detected	544×544	VOCO7 VOC12 COCO	Strategies used to improve accuracy and speed. Achieve mAP 76.8 on VOC 20O7, 73.4 on VOC 2012, 21.6 on COCO, and FPS is 67.

Table 3 (continued)

Technique	Strengths	Limitations	Input Size	Dataset	Results
RetinaNet [93]	Focal loss helps to deal with imbalance between foreground and background; Use two backbones (ResNet and FPN)	It cannot run in real-time for the detection of small objects. When the number of classes is large, using more anchors will increase more parameters in the anchor function	500×500	COCO	Achieve higher accuracy as compared to the previous technique, deal with imbalances and inconsistencies of the SSD; to tackle dense detection; FPS is 11 and achieves mAP 34.3%
DSSD [42]	Detect context-specific or small objects	Not fast as SSD because of more layers of Residual-101 network, extra layers added to predication and deconvolution layer etc.	321×321 or 513×513	COCO and PASCAL VOC 2007 and VOC 2012	Model is pre-trained on ILSVRC CLS-LOC dataset using Residual-101 backbone network; prediction and deconvolution module added to SSD model; achieve mAP 81.5% on VOC07 with input 513×513 and 89.0% on VOC2012 and 33.2% on COCO.
Tiny SSD [190]	Highly optimize network model for real-time embedded system and maintain object detection performance	Efficiency and performance is not matched with real-time detection	300×300	PASCAL VOC 2007	To optimize SqueezeNet no. of filters reduced and down sampling; SqueezeNet used as backbone architecture; achieve 61.1% mAP on VOC2007 and FPS is 2.3 MB
YOLOv3 [140]	DarkNet53 used a backbone architecture and multi-scale predication borrowed from RPN and SSD.	It still has the problem of small object detection	320×320	COCO	Both speed and accuracy are highly improved over YOLOv2. Achieve mAP 28.2 and FPS is 45.
MASK-RCNN [59]	Detect Not only the objects but also generate high-quality segmentation; used as ResNet-50-FPN backbone.	Works only on still images and failed to detect objects at low resolution and blurred images. Detection speed does not meet with real-time	1024×1024	COCO	Extends detector Faster RCNN to detect and instance segmentation; proposed RoI Align for pixel-to-pixel alignment; achieved AP 36.4% at 5 FPS.
LADet [216]	Used two modules (AFPM and LCFM) to provide better accuracy and efficiency when speed is important	Further improvement is required in the LCFM for the multi-scale object detection accuracy	512×512	COCO and PASCAL	To address scale invariance, use more anchors without increasing the anchor parameters. Achieve mAP 81.4% on PASCALVOC2007 and 33.6% on COCO
YOLOv4 [9]	A superior accurate and fastest detector in both speed and accuracy; CSPDarkNet53 is selected as the	Further improvement is required to achieve optimal accuracy in object detection.	416×416	COCO	An efficient and accurate model is trained on conventional GPU,

Table 3 (continued)

Technique	Strengths	Limitations	Input Size	Dataset	Results
	backbone after analysis; a lightweight and easy to use detector				PANet for feature aggregation, Bag of Freebies, Bag of Specials, CIOU, etc. is used to create the model. Achieve AP 43.5, and FPS is 65 on the coco dataset.
YOLOv5 [81]	Extremely fast and lightweight than YOLOv4; nearly 90% smaller than YOLOv4.	It accurately detects objects, but it has still need to improve to meet the real-time detection	640 × 640	COCO	Use PyTorch for implementation instead of DarkNet; 4 models are YOLOv5s, YOLOv5m, YOLOv5L and YOLOv5x; same as YOLOv4, but it used auto-learning BB anchors and Mosaic data augmentation; achieve mAP: 45

In this table, several techniques are reviewed, such as traditional as well as deep learning-based. In general, features pyramids and sliding approaches were used to detect the objects in traditional techniques, but these techniques only detect the objects with a fixed aspect ratio. So, a two-stage detector has been used to detect the objects which have detected the objects in two stages. In this, the first stage is used to generate the regions of interest from the input image, and then these regions are sent to the pipeline for bounding box regression and classification. One stage detectors such as YOLO, SSD etc., are ended to end detectors that detect the objects directly from input images by learning bounding box coordinates and class probabilities of specific classes

5 The performance metric for object detection

Different algorithms have been developed to detect the object from videos or images. Therefore, the performance of these algorithms can be evaluated using metrics. Many approaches have been assessed for the model's accuracy or an algorithm in the form of speed or accuracy. *Frame Per Second (FPS)* has been considered to evaluate the different approaches in terms of speed. It is the procedure to express how fast an algorithm or approach is. The fps of any approach is higher than the other approach, which means that approach can process many frames per second [115]. For example, the Fast YOLO has 155 FPS higher than YOLO, which has 45 FPS [214].

The most commonly used metric to evaluate object detection performance is Average Precision (AP). Before preceding the AP, here, this study should review some of the basic concepts that the AP uses, such as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). A TP is a correct detection of the ground truth (annotated bounding box), and FP provides the positive window on an incorrect or inappropriate place outside the ground truth bounding box. In FN, the detector is unsuccessful in generating the positive window, although an object's ground truth bounding box exists in the image. TN is a correct detection of the negative class. True negative is not used in object detection because many bounding boxes exist in the image that should not be detected [130].

In object detection, locating the different objects by localization means the bounding box is predicted around the objects. IoU (Intersection over Union) has been used to check how much-predicted bounding boxes are accurate. IoU provides the overlapping area of the ground-truth BB_g and predicated BB_p The bounding box is divided by predicted and ground-truth bounding boxes. Equation 1 shows how IOU works in the case of object detection and how it is important for object detection. The IoU is calculated by using equation number 1.

$$IoU(BB_g, BB_p) = \frac{area(BB_g \cap BB_p)}{area(BB_g \cup BB_p)}$$

The detection of an object is correct or incorrect, and this can be classified with the help of given threshold t and IOU overlap. For this, the threshold t and IOU are compared. Detection can be considered correct if the $IOU \geq t$; otherwise, it is incorrect. Several detectors overlap one ground-truth bounding box. Only one is considered a true positive from this detector [20]. Precision and recall concepts have been mostly used to assess object detection methods.

Precision is the capability of a model to recognize only significant objects. In other words, precision is a percentage of retrieved predictions that are relevant or correct. Precision P is calculated by using the TP (True Positive) and FP (False Positive). Precision is calculated by using the following equation number two:

$$\mathrm{P} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{2}$$

The recall is a fraction of relevant instances that are successfully retrieved. It is calculated using TP (True Positive) and FN (False Negative). Recall R is calculated by using the following equation number third.

$$\mathrm{R} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{3}$$

A detector identifies all relevant objects to find the ground–truth objects. A detector can be a good detector if the value of the precision or recall is still high while the confidence threshold value can be varied. Average precision (AP) is a metric that is calculated by using the average value of precision over recall intervals of 0 to 1. In other words, it is the area under the precision-recall curve (PRC). The value of AP is higher means the method's performance is good and vice versa. The mAP (mean Average Precision) is a metric used to measure the detector's accuracy in all the classes, and this is the actual metric to check the accuracy of detection [142]. The mAP is computed using the equation number fourth.

$$\mathrm{mAP} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=0}^{\mathrm{N}}\mathrm{AP_i} \tag{4}$$

Where N is the total number of classes and AP_i is the *ith* class of the AP. One more metric F1 score is a harmonic mean of Precision and Recall. Sensitivity (SEN) and Specificity (SPE) metrics are used for evaluation purposes and calculated by using the equations fifth and sixth [7], such as

$$\mathrm{SEN} = \text{True Positive}/\text{True Positive} + \text{False Negative} \tag{5}$$

$$\mathrm{SPE} = \text{True Negative}/\text{True Negative} + \text{False Positive} \tag{6}$$

LRP (Localization Recall Precision) Error has been proposed to deal with shortcomings of the AP (Average Precision) metric [209]. LRP is computed using three components: localization, FN, and FP.

As shown in Fig. 11, precision and recall metrics are used for comparing the various object detection. Some of the studies have used a combination of these metrics to evaluate the results. Recall and precision are the most widely used combination.

6 Applications of object detection

Object detection has been used widely in industries and areas. Various application areas of computer vision tasks are security, image retrieval, surveillance, machine inspection, automatic vehicle system, and many more. Current applications are discussed in the following section based on the latest publications from 2020 to 2022, as shown in Table 4.

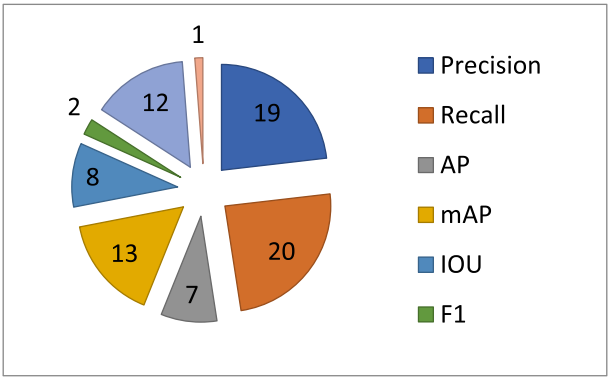

Fig. 11 Performance metrics used in different studies

6.1 Pedestrian detection

Object detection plays a significant role in identifying the pedestrian on the road. Many researchers have been used in many application areas such as video surveillance, autonomous driving, robotics [30, 45, 116] etc. video surveillance: In the intelligent video surveillance system, pedestrian detection plays an important task because this provides the semantic understanding about the video footage and helps to improve the safety system by detection of pedestrian movements [206]. Autonomous driving**:** object detection plays a significant role in the autonomous driving system. With the help of object detection, systems can identify nearby objects such as cars, traffic lights, road signs, pedestrians, motorcycles, etc. [112, 179, 189]. Therefore, the car uses object detection trained models to decide whether to apply break or turn according to identify the nearby objects. Visually impaired persons also use these autonomous driving systems to go anywhere in the world without any dependency on a human being [71]. Various researchers have developed different data sets such as MIT Pedestrian, INRIA, Catech, KITTI, CityPersons, EuroCity, nuSense, etc., as shown in Table 2. Some of the difficulties in pedestrian detection include small object detection, Dense and Occluded pedestrian, real-time detection, and weather conditions [112, 126].

Worldwide, researchers have worked on pedestrian detection. The social distancing monitor network was designed using PeeleeNet backbone architecture and yolov3 architecture on the human head (Merge-Head) dataset. Firstly, it will detect the pedestrian from the real-world images then calculate the distance between them [156]. The object detector's performance has been affected by label assignment. Therefore, for crowd scenarios, the label assignment strategy, i.e., LLA (Loss-aware Label Assignment), is designed for pedestrian detection in crowd areas. The architecture RetinaNet and FCOS is used with LLA to improve the MR by 9.53% and 5.47%, respectively, on the CrowdHuman and CityPersons dataset [44]. An investigation to improve the efficiency of experiments is performed on two-stage detectors, i.e., RCNN with different pre-trained networks such as AlexNet, VGG16, and VGG19. The Efficiency of RCNN is improved with VGG19 that achieves the highest AP, Recall, and F-measures [111]. The pedestrian segmentation and detection task are optimized with the use of the proposed RGBD pedestrian detection formwork, i.e., WSSN (Weak Segmentation Supervised Network). This framework does not use pixel-level segmentation during training. Despite this, weak segmentation masks are used, automatically made from depth images. These masks are further used for the classification task. The proposed method used the pre-

trained network VGG16 with the COCO dataset [50]. The augmentation process has been improved to increase the performance of the detectors. The "Shape Transformation-based Dataset Augmentation" (STDA) framework proposed to augment the dataset by performing a transformation on real pedestrians from the dataset into various shapes [19]. Pedestrian detection from a low-quality image is challenging for researchers. Therefore, jin et al. proposed a playground dataset that contains low-quality images with blurred scenes, heavy weather, etc., of different periods (day or night). Jin et al. also proposed an SRD (Super- Resolution Detection) network that enhances low-quality images' resolution [73]. The SRD is combined with some improvements in Faster RCNN to detect pedestrians from the low-quality videos. Gawande et al. have proposed SIRA (Scale, Illumination, Rotation, and artifacts) with Mask RCNN named SIRA Mask R-CNN detector that detects pedestrians in different conditions such as scale, illumination, rotation, etc. [43]. A brief overview of the recent publications is shown in Table 4. Most of the work has been done on CityPerson, Caltech, KITTI, and COCO datasets. Due to the advancement of deep learning, most studies have used the Faster RCNN and the different versions of YOLO for pedestrian detection techniques, as shown in Table 4.

6.2 Face detection and recognition

Face detection and recognition are widely studied in computer vision and are among the oldest computer vision applications [202]. Applications based on such techniques are most frequently used in daily routines, such as Facebook performing face recognition when the image is uploaded on that application [5]. Object detection is a process to identify the object location and define the class of the detected object. Face detection is a widely used computer technology in various applications to identify human faces from digital images, and further face recognition is used to verify a person detected from images or videos [131]. Face recognition is mostly used in biometric technology. All face detection algorithms have focused on detecting frontal human faces [77]. For detection, features of the images are compared with the features stored in the database. Nowadays, face recognition has been used in a lot of applications, such as used to unlock phones and other specific applications [52]. With the help of face recognition in banks, airports, retail stores, biometric surveillance, and other applications reduce crime or prevent violence. VJ detector has been widely used to detect the face from images [133]. In this, some difficulties include intra-class variation, multi-scale detection, occlusion, and real-time detection [202, 206]. Most of the applications consider facial expressions to analyze human behaviour. Smile detection is also part of facial expression used in many applications like patient monitoring, photo selection, etc. [127]. In Table 2, FDDB, AFLW, IJB, MALF, WiderFace, etc., widely used datasets are used by various researchers to solve their problems.

Face recognition has been used in the home security system. Ravishankar et al. have developed a smart home system that recognizes the face and sensor according to the match the face is activated. Three types of sensors are used in this system Fire sensor, turbidity sensor, gas sensor, and according to the sensor alert message is sent on the telegram bot [137]. From the security point of view, hung proposed a method with the use of HOG and CNN for face detection from images [67]. To detect the face from collaborating learning environment is also particularly challenging. Tran et al. have proposed a method that detects the face from collaborating environment along with a different variation of poses [174]. Researchers also identify the person from the live video by providing the label [153]. The face detection is also done based on the complexion of people using the hybrid approach in which RGB and YCbCr

Table 4 Summarization of target, method, metrics and results of latest object detection architectures

	Target and Ref.	year	Dataset	Method	Metric and Results
Pedestrian detection	Small-scale Pedestrian Detection [193]	2020	CityPersons and Caltech	Faster R-CNN, ResNet-18 or ResNet-50, FPN, SML	ResNet50, ResNet18 reduce miss rate by 2.6% and 2.0%
	Pedestrian detection [31]	2020	Own dataset	Hybrid Gaussian + HOG + SVM	false detection rate reduced to 4.3%
	Pedestrian Detection from Drone Images [68]	2020	Own, UAV123	Faster RCNN, Inception-v3	precision 98%, recall 99%, and F1 98%
	Pedestrian Detection [63]	2020	CVC-09, CVC-14	YOLOv3, Faster R-CNN	mAP: 84.78%
	Pedestrian Detection [170]	2020	Own dataset	Faster RCNN, transfer learning, ResNet18	AP: 60.88, P:70.42 R:69.17 T: 0.411
	Pedestrian detection with the camera [129]	2020	Own dataset	CNN, Max pooling	
	LLA used to assign a label for detecting dense pedestrians [44]	2021	CrowdHuman, CityPersons	Loss aware Label Assignment, ResNet50, RetinaNet, FCOS	Anchor based (AP: 84.69, recall: 89.90), Anchor free (AP:88.04, recall: 93.95)
	Proposed WSSN to detect pedestrians [50]	2021	KITTI, COCO, EPFL, RGBD people dataset	WSSN (version of Faster RCNN), VGG16	Achieve AP: 72.1% on EPFL and, 67.3% on RGBD, 80.9% on KITTI get in the average case.
	Augmentation framework based on shape transformation [19]	2021	Caltech, CityPersons	STD, base detector (ResNet50-based FPN)	Baseline detector improved 38% on benchmarks.
	Detect from low-quality images [73]	2021	Playground	Faster RCNN, VGG16 and SRGAN	AP: 99.10%
	Reconstruct the yolov3 to improve the speed [198]	2021	KITTI	YOLOv3 promote with use of G-module in depth wise convolution	Achieve 93.1% mAP and 25.5 FPS
	Three pre-trained networks investigate to refine the efficiency of RCNN [111]	2022	Penn-Fudan, KTH Football, CUHK01, CUHK02	AlexNet-RCNN, VGG16-RCNN, VGG29-RCNN	Achieve the highest AP, Recall, and F-measure is 53.65, 0.57, and 0.70, respectively, on the CUHK02 dataset
	To handle the scale, illumination, rotation, affine invariant in pedestrian detection [43]	2022	Caltech, INRIA, COCO, KITTI, ETH	SIRA Mask R-CNN with ResNeXt-101-FPN backbone architecture	The lowest miss rate on Caltech is 8.31%, the lowest log-average miss rate on INRIA is 7.31%, the miss rate on ETH dataset 32.63%, 79% accuracy on KITTI dataset and 8.68% miss rate on the proposed database.
	Pedestrian detection to check the distance between them due to covid19 [156]	2022	Merge head dataset	PeeleeNet (backbone), YOLOV3	AP: 92.22% FPS: 76

Table 4 (continued)

	Target and Ref.	year	Dataset	Method	Metric and Results
Face detection and recogni-tion	Face recognition from big data [218]	2020	ATR Jaffe	2DPCA, LBP	Accuracy >90%
	Face feature localization [159]	2020	Multi-PIE, BU-3DFE, SFEW	CNN	Accuracy:0.9168
	Develop face detection and recognition algorithm [106]	2021	LFW, ORL	OMTCNN, LCNN	Accuracy: 98.13%
	To enhance the accuracy of face detection, the use of a deep learning model [114]	2021	LFW and own created dataset	CNN, Max polling	Training accuracy: 95.72% and validation is 96.27%
	Face detection from surveillance videos [61]	2021	Own dataset	Hybrid approach using RGB and YCbCr color space	Average detection rate: 97.51%
	Face detection in occlusion and non-uniform illumination [80]	2021	AR, manual created, and Color FERET	The proposed technique uses HSV, YCbCr and L×a×b colour space model.	Achieve accuracy >96%, precision>95%, detection rate> 97% and false detection rate on AR anf FERET dataset, for manula creation precision is 94%, accuracy 93%, detection rate 94% approximate, detection rate is 6%
	Facial expression recognition [176]	2021	KDEF, GENKI-4 K, CK+	Designed a technique based on CNN and data augmentation	Achieve 97.69% on CK+, 94.67 on GENKI-4 K, 83.43 on the KDEF dataset
	Face detection and identification [67]	2021	FEI, LFW, UOF	HOG and SVM, CNN	Precision: 99.74, recall 99.74 and accuracy 95.26%
	Recognize face from the collaborative learning environment and occlusion handle [174]	2021	Manually created video dataset of 24 participants (11 boys and 13 girls)	Proposed face recognizer algorithm with k means	Accuracy is 86.2%
	Liveliness check for authentication [136]	2021	Manually created dataset, i.e., Liveliness dataset	MTCNN, FaceNet based on CNN, SVM	Accuracy: 99.46%
	Develop a face detection algorithm [51]	2021	Yale2B, Face94, ORL, M2VTS, and FERET	SURF, SIFT, K means clustering, Locality preserving projections (LPP)	Recognition rate on ORL 91.0%, Face94 87.3%, M2VTS 98.8%, FERT 76.8%, Yale2B 99.70
	Face recognition in the unconstrained environment [168]	2021	ORL, AR, LFW	HRPSM_CNN	Achieve 97% accuracy on ORL and AR, 96% on LFW
	Tracking of an employee with face recognition [69]	2022	LFW	FaceNet, KNN	Achieved 97.8% accuracy
	Criminal person identification by face [10]	2022	Manually inserted data of each criminal	Haar Cascade algorithm	94% positive response

Table 4 (continued)

	Target and Ref.	year	Dataset	Method	Metric and Results
	Cattle face identification [199]	2022	Manually created dataset	Proposed CattleFaceNet with the integration of RetinaFace_MobileNet and ArcFace	Accuracy is 91.3%, and FPS is 24
	Face recognition for security system [137]	2022	–	A haar-cascade algorithm, Local Binary Pattern Histogram algorithm	Accuracy: 87.5%
Remote Sensing Images	Detect objects in aerial image [163]	2020	DOTA and HRSC2016	R 4Det: RFP, RFC, FOCAL LOSS,	Dota (mAP): 75.8 4, HRSC2016 (mAP): 89.56
	object relationships integrate for OD [171]	2020	ImageNet,	DetNet-59, SGD, faster R-CNN	mAP: 0.8614
	Detect objects from the sea [152]	2020	LWIR	Yolov3, Faster RCNN, RetinaNet	mAP: 0.94
	SOD from remote sensing optical images [91]	2020	ORSSD	PDF-Net: VGG-16/ 19	P: 0:9144 R: 0:8027
	Detect multiple objects and micro-object [64]	2020	NWPU_VHR-10, DOTA, UCAS-AOD	SOSA-Net: AD-FCN, Resnet-101	mAP, Maxrecall: >99
	Multi-scale and rotation aware OD [40]	2020	DOTA	Faster-RCNN, FFA, RPN-O, ROI-O, RestNet101	FFA: 67.9% FPN: 66.7%
	Arbitrary-angle BB based location for OD [164]	2021	NWPU VHR-10 and TODRS-3	R-FRCNN: FEN and RPN	reduce the MAR and FAR
	multi scale object detection in high resolution image [215]	2021	NWPU VHR-10	OR-FS-SSD+CA	mAP: 94.74, FPS: 29.57
	Develop a saliency object detection framework with image-wise annotation [210]	2021	SPOT5 and GeoEye-1	progressively supervised learning (PSL)	GeoEye-1(Precision: 0.9409, recall: 0.7747, F measures: 0.8941), SPOT5(Precision: 0.8560, recall: 0.8261, F measures: 0.8448)
	To improve the detection accuracy, improved detection method YOLOv3 [194]	2021	TGRS-HRRSD and RSOD	Improve the DarkNet53 architecture, mish activation function, and YOLOV3	log average miss-rate: 0.0338, mAP: 99.78%
	Detect objects of different scales [188]	2021	DOTA and DIOR	FSoD-Net	DOTA (mAP: 75.33%), DIOR (mAP: 71.80%)

Table 4 (continued)

	Target and Ref.	year	Dataset	Method	Metric and Results
	Detect complex composite objects [165]	2021	Manually created Sewage treatment plant (STP) dataset, DIOR-composite dataset	PBNet with backbone VGG16	Recall: 90.02, precision: 72.44 AP: 85.39 and FPS: 15 on STP dataset. mAP 74.53 on DIOR-composite dataset
	Handle the Multiview angle issue [213]	2021	Pascal3D+ and SpaceNet MOVI	PSNet with ResNet-101 back-bone architecture	mAP is 84.0 on Pascal3D+, AP on SpaceNetMOVI 35.0
	Solve the problems that occur with small objects, complex backgrounds, and different scales [102]	2021	NWPU VHR-10, DOTA	AF-SSD	The mAP on DOTA is 52.6 and on NWPU VHR-10 is 69.8
	Target detection from multi scale and direction [208]	2022	DOTA	Modified YOLO	mAP: 77.68%
	Recognize the building rooftops [99]	2022	Own	MS-CNN	Precision: .8655, Recall: .8380, F1: .8516
	Objects detect from aerial images [92]	2022	VisDrona2019	GLE-Net	mAP: 23.1
	Extract buildings from remote sensing images [184]	2022	WHU and INRIA dataset	RU-Net	Recall: 96.98, Precision: 97.48, F1: 97.23, IOU: 94.61
	Crosswalk detection from remote sensing images [17]	2022	Crosswalk dataset	YOLOV3, Faster RCNN, DenseNet based on YOLOV3, U-Net, mixture classification strategy	Results use of YoloV3, Faster R-CNN, and YoloV3 Based on DenseNet are F1 score (0.8976, 0.7600, and 0.9109), Accuracy (86.24, 63.21,91.15), Recall (93.57, 95.29, 91.04), mAP(89.96, 92.19, 89.08) respectively
Text Detection	Multi scene TD [58]	2020	ICDAR2015, ICDAR2013, MSRA-TD500	SRPN: VGG-16, FCN, feature fusion, OHEM, Sigmoid	F: 85.40% at fps: 16.5
	TD of Arbitrary-shaped Scene [187]	2020	Total-Text, CTW1500, ICDAR2015	ContourNet: AdaptiveRPN, LOTM, Point Re-scoring Algorithm	R: 83.9%, P: 86.9% and F-measures: 85.4%
	Multi-oriented scene text image detection and recognition [123]	2020	ICDAR 2013, 2015, and 2019	ResNet-50, New.i.ReLU, new.i.inception layers, local word directional pattern (LWDP), SGD	Comparison of three datasets, ICDAR 2015 achieve S:0.8309 P:0.9347 F: 0.8797
		2021			F-measure of 76.85%

Table 4 (continued)

	Target and Ref.	year	Dataset	Method	Metric and Results
	TD using instance segmentation [217]		ICDAR2015, SCUT-CTW1500, RCTW-17	TextMountain: TS, TCBP and TCD, SGD, ResNet-50	
	Detect arbitrary text as a visual retainership detection problem to solve the text-line grouping problem [107]	2021	RCTW-17, Total-Text, MSRA-TD500, DAST1500 and CTW1500	ReLaText with GCN	RCTW-17(P: 75.9, R: 61.7 F: 68.1), MSRA-TD500 (P: 90.5, R: 83.2, F: 86.7), Total-Text (P: 84.8, R: 83.1, F: 84.0), CTW1500 (P: 86.2, R: 83.3, F: 84.8)
	Recognition and detection of scene texts in arbitrary shapes [200]	2021	CTW1500 and TotalText, ICDAR2019-Art, MSRA-TD500	Proposed technique with the use of "Mask-guided multi-task network"	CTW1500 (R: 77.7, P: 88.3, F: 82.7, FPS: 7.6), ICDAR2019-Art (R: 59.6, P: 77.8, F: 67.5, FPS 7.6), MSRA-TD500 (R: 76.7, P: 92.1, F: 83.7), TotalText (R: 74.7, P: 77.3, F: 76.0)
	Domain adaptive scene text detection using a self-training framework [18]	2021	MSRA-TD500, ICDAR2015,	MaskRCNN, TMM, balance loss and Gen-Loop	ICDAR2015 (P: 91.2, R: 85.4, F: 88.2), MSRA-TD500 (P: 87.9, R: 83.1, F: 85.4)
	Evaluate the performance of classifiers that recognize the Gurumukhi script newspaper [75]	2022	Own dataset	KNN, Random Forest, MLP, decision tree, diagonal classifier	With the use of diagonal feature extraction MLP classifier achieved 96.5% accuracy, and the random forest achieved 96.9% accuracy
	Detect text from the image to answer the questions [195]	2022	ST-VQA	TextVQA	Validation accuracy, 0.2842, testing accuracy: 0.289
Traffic sign detection and classification	Speed limit and vehicle entry prohibition signs [167]	2020	Dataset A and Dataset B	the average colour value of the filter, canny edge, generalized Hough transform	The success rate for Dataset A is 97.3%, and B is 98.2%
	Road damage detection [109]	2020	Road Damage Dataset 2019	PG-GAN, SSD with MobileNet, SSD with Resnet50	F measures: 0.60% (SSD with MobileNet) and 0.43% (SSD with Resnet50)
	Vehicle detection [146]	2020	PASCAL VOC 2007,2012 and MS COCO 2014	LittleYOLO-SPP Generalized IoU K-means	mAP: 77.44
	Recognize vehicles in foggy weather [90]	2020	GTI vehicle dataset	AITwo	Accuracy >97%

Table 4 (continued)

Target and Ref.	year	Dataset	Method	Metric and Results
Traffic sign detection [212]	2020	GTSDB, CCTSDB, Lisa	multiscale cascaded R-CNN ResNet50	Accuracy and recall rate on GTSDB 98.7% and 90.5%, on CCTSDB 99.7% and 83.62%, on LISA 98.9% and 85.6%
Traffic Sign Detection and Recognition [2]	2020	ITSD	Morphological shape filter, Nearest Neighbour matching-based recognition, SURF	Accuracy: 97.83%
License Plate detection for non-helmeted motorcyclist [72]	2021	Own dataset	Darknet-19 with YOLO	Overall detection rate: 98.52%.
License plate recognition [158]	2021	Chinese City Parking Data set with 100-nighttime image	Lite-LPNet	Detection 90% and recognition: 98.73%
Car license plate [135]	2021	CCPD, ALOP, PKU	The proposed method with backbone architecture is either Resnet18 or ResNet50	Achieve average accuracy is 97.8
multiple and mixed style LP recognition [66]	2021	HZM multi-style dataset (Own), compared with ALOP, PKU	ALPRNet ResNet-50 Ranger optimizer	Achieve 98.21 accuracy
Driver activity recognition [86]	2021	Own dataset	ST-GCLSTM, focal loss function, transfer learning, temporal exponentials mean filter	Achieve recall ratio: 88.80%

colour spaces are combined for face detection from surveillance videos [61]. With the use of a three-colour space model, Kumar et al. have proposed a technique to detect the faces from occluded and non-uniform illumination images [80]. Due to the advancement of deep learning, Lv et al. have combined the two deep learning methods such as OMTCNN and LCNN, for face detection and recognition. This system provides a good result and applicability for the embedded platform [106]. Javed et al. have also designed a system using a deep learning model that recognizes the face and achieves 95.72% accuracy [114]. Despite face detection and recognition, the different facial expressions of human beings are also recognized with the use of deep learning from the image [176]. Face liveliness is another area of face detection in which the person in front of the camera checks that the face of the person is live so that face spoofing attacks are prevented and ensured by biometric authentication. Rahman et al. have proposed an attendance system that detects a person's liveliness using heart-beat measurement and the deep learning model FaceNet [136]. This system helps for the detection of spoofing attacks in the authentication. According to Tamilselvi M. and Karthikeyan S., different methods have been used for face recognition, such as LBP (Linear Binary Pattern), Multi-SVM (Multi Support Vector Machine), CNN, DBC (Directed Binary Code) etc. However, satisfactory results have not been achieved yet due to occlusion and poor lighting in large databases. Therefore, they have proposed an HRPSM_CNN ("Hybrid Robust Point Set Matching Convolutional Neural Network), which achieved 97% accuracy [168]. This approach also provides better results for visually impaired assistive devices under different weather and lighting conditions. The identification of cattle is difficult due to food quality tracing, breeding association, fake insurance claims etc. Therefore, Xu et al. have presented a new framework named as CattleFaceNEt which is combination of RetinaFace-mobileNet and ArcFace (Additive Angular Margin Loss) [199]. Human interaction stopped during covid 19. Employees did their work from home. Therefore, the tracking system has presented face recognition [69]. The recent work of the recent publication is shown in Table 4. The table shows the proposed approaches to the different datasets and shows the results of these datasets.

6.3 Text detection

Most of the information has been preserved in the form of text in a large part of the world [70]. In this, researchers only determine whether the text is available in the image or not. If the text is available in the image, localization and recognition have been performed on the detected text [14, 74]. The ability to read the text in a natural, unconstrained environment is used by many real-world applications to help people [178]. The text is extracted from the images or video, whether the text is handwritten or printed, a scene photo, document photo etc. Object detection plays a significant role in this application so that information is electronically edited, stored more systematically or compactly, text displayed online, and used for machine processing [119]. For example, most researchers extract the text from images by using object detection and then convert the extracted text into speech. Visually impaired persons used this application to read the street signs or currency [7]. It has also been used to build digital maps by detecting and recognizing street signs and house numbers. Handwritten Chinese text is also detected using CNN [117]. Some of the difficulties that can be affected by text detection are broken and blurred characters, different fonts and language text rotations, and different fonts and language [204]. There are different data sets such as ICDAR, SVT, IIT5K, CocoText, Total-Text etc., as shown in Table 2.

Despite the English text, the Gurumukhi text is also recognized in the newspaper with the use of classifiers. Different classifiers and features techniques were analyzed that recognize the broken characters, heavily printed characters, etc. After analyzing, MLP (Multi-Layer perceptron) and Random forest classifier with diagonal feature extraction strategy achieve high accuracy [75]. The emerging research problem is VQA (Visual Question Answering), a combination of natural language processing and computer vision. This research topic is becoming popular due to applied in many application areas, such as assisting blind people and children's education. The proposed framework "TextVQA" used three modules: an attention mechanism, multimodal information fusion, and attention map loss to improve model accuracy. This LSTM provides the complex answer to predicated words [195]. Xue et al. have also developed a multi-task network technique to detect and rectify the text from arbitrary shapes [200]. The text line grouping problem is solved using a link relationship. The proposed approach "RelaText" divides the text detection problem into two parts: the first part is to find the text primitives with the use of an "anchor-free region proposal network" and the primitive graph is composed of the detected text primitives [107]. Then, Graph Convolution Network (GCN) approach is used to achieve better accuracy in small interline spacing and larger intercharacter. However, the object detectors suffer performance degradation when well-trained detectors are applied on the different target domains. In simple terms, a huge amount of data is required to train a detector for the target domain. Despite this, the annotation and data collection processes are very time-consuming and expensive processes. Therefore, Chen et al. have proposed a self-training framework that solves the "domain adaptation scene text detection problem" using unannotated images or videos [18]. The proposed framework used the text mining module and Gen-Loop (Image to video generation) method to train a network. All the brief information on the related work toward text detection and recognition is shown the Table 4. The table shows that text detection and recognition have been applied in many sub-application areas.

6.4 Traffic sign detection and classification

Traffic signs are very important for everyone to reach a destination safely. These are the rules or information to inform the drivers about the road conditions to be safe from accidents. Traffic signs contain a lot of information and are complex, so needing the constant attention of drivers towards the traffic signs is not an easy task [25]. In recent years, many object detection methods have been developed and improved several aspects of daily life, such as the health care system and cars with automatic control [118]. In the traditional methods, traffic signs have been detected by considering the properties of the signs, such as the colour and shape of signs [108]. After that, deep learning has been used for traffic signs or road sign detection and recognition [11, 39]. Some of the difficulties, such as illumination change, harsh weather, motion blur, and real-time detection, arise during the detection [76]. In this paper, some commonly used datasets, such as TLR, LISA, GTSDB, BelgainTT, MTSD, etc. (shown in Table 2), detect traffic signs and lights. Object detection also helps to track the objects from videos such as tracking a person in a video, during the football match ball is the track, tracking movements of a bat in cricket match [166]. Object tracking has been used in most of the area such as for motoring the traffic, for security or surveillance, robot vision, and animation [116].

Table 4 summarizes recent studies published in traffic sign detection and transportation detection. Other transportation detections such as car license plate (LP) detection, vehicle detection, and driver activity recognition are similar tasks performed by the machine. Recently,

automatic detection of road (traffic) signs and signals has gained much attention due to the rapid development of autonomous driving. Like other object detection, road signs and traffic light detection have challenges such as illuminations variations (day or night), motion blur, harsh weather, and real-time detection. Due to the quick advancement of deep learning, YOLO, SSD, RetinaNet, and Faster RCNN techniques were also adopted for training and testing on road signs/traffic light detection, road damage detection tasks [3, 6, 100, 148]. Some new approaches, such as adversarial training, different backbone architecture, and attention mechanism, have been used to increase speed or improve detection processes in challenging and complex environments [66, 72, 88, 100, 109, 135]. An accurate observation of the surrounding environment is required for an autonomous vehicle (AV) to help human beings to reach their destination safely. Most of the deep learning and machine architecture converts sensory data into semantic information that supports autonomous driving. The detection of objects is an important characteristic of this system of perception. In 3D OD, the three-dimension reveals more information about the objects, such as size and location. A different method has been used for detecting 3D objects, such as point-cloud monocular and fusion points. The government focuses on implementing safe and law-abiding actions in traffic to solve this problem. 3D car instance benchmark is released to understand autonomous driving [160]. Sensor fusion is utilized for attaining better features [4]. Lu et al. proposed the architecture consisting of RNN and 3D convolution to obtain the "centimeter-level localization" accuracy [101].

6.5 Remote sensing target detection

Object detection is fundamental as well as a challenging problem in optical remote sensing images. This application has received much attention in recent years and helps to improve the understanding of the earth [20]. Many applications such as disaster rescue, urban planning, geographic information system updating, and military investigation have used remote sensing target detection [89]. Some difficulties arise in the detection process, such as occluded targets. Many automatic target recognition (ATR) algorithms are used to recognize targets or other kinds of objects with the help of sensors. The various researchers presented different data sets such as TAS, OIRDS, DLR3K, DOTA, DIOR LEVIR, xView, etc. As shown in Table 2. ship detection is also a widely used application and challenge to detect the ship from remote sensing images [96]. Remote sensing target detection methods are categorized into four types such as knowledge-based methods, template-based methods, OBIA based methods and last, machine/deep learning-based methods [20]. In geospatial object detection, DIOR is the challenging benchmark. The various categories such as vehicle, harbour, overpass, bridge, etc., are not accurately detected by twelve representation methods due to the low quality and cluttered background of the aerial images compared with real environment images. Researchers suggest that the Snip and Snipper training schemes are applied to existing networks for better results [89]. The brief information on the recent work is shown in Table 4.

Detecting objects from the spatial and background distribution is too challenging in high-resolution remote sensing images. Zhao et al. proposed a multi-scale object detection approach using "rotation invariance deep features "determined by channel attention named OR-FS-SSD + CA [215]. This method detects objects with different orientations and scales. Due to the feature-capability of DNN (Deep Neural Networks), MS-CNN (multiscale convolutional neural network) is proposed to detect or recognize the building rooftop from the remote sensing image of high resolution [99]. The novel approach AF-SSD has developed to handle the problems that occur with complex background and scale variation [102]. In this method,

the first FFM method is introduced for the fusion of upper layers and shallow layers' features, and background noise DAM has been introduced. Moreover, the third model for this technique is designed to capture multiscale features and broaden the receptive field. Crosswalk detection plays a significant role in traffic safety. Therefore, researchers also work on crosswalk detection from remote sensing images [17]. The RU-Net methods have been proposed to extract the buildings from remote sensing images [184]. Zang et al. have developed a new salient object detection method named Progressively Supervised Learning (PSL), a combination of fully supervised and weakly supervised learning [210]. In this method, the "pseudo-label generation method" is proposed for the classification network and Grad-CAM ("gradient-weighted class activation mapping") for the computation of pseudo saliency maps. These methods minimize the demands of large-scale pixel annotations. Wu et al. have developed the target detection method in which DarkNet53 is improved based on YOLOV3 [194]. Due to the best performance of the YOLO architecture, Zakria et al. have also presented the approach with some modifications in YOLO-v4 for the detection of multiscale and direction targets from remote sensing images [208]. In this, they proposed the classification setting of the NMS threshold so that the method's accuracy is increased without affecting the speed. Due to extreme differences in object scale, Wang et al. have developed an FSOD-Net (Full-scale object detection network) which consists of a proposed MSE-Net (multiscale enhancement network) backbone architecture cascaded with SIRL (Scale-Invariant Regression Layer) [188]. GLE-Net (Global and local ensemble network) is proposed to detect the small, dense objects of aerial images taken by the drone. This network used the two base networks YOLOv5 and CenterNet for comparison and this network can act as plug-and-play network for improving the accuracy of any detection network [92]. Some of the objects such as golf courses, sewage treatment plants, and airports that are not present in fixed size or shape. Therefore, PBNet (Part-based convolutional neural network) has been designed for the detection of these rigid objects from remote sensing images [165]. Object detection from different view angles also is a challenging issue. Zhang et al. have presented an approach PSNet (Perspective-sensitive network) in which perspective-specific structural features are used instead of uniformed features [213]

6.6 Other application areas

Object detection has been used in other applications. *Object Recognition as Image Search: object* detection can also be used as an image search. For doing this, objects from the image are detected to get the label of objects in the first step, and these labels are passed to the URL to perform an image search. *Object counting:* Counting the number of objects from an image or real-time video can also be done with the help of object detection. During the festivals, People counting is used to estimate the crowd statistics. *Automatic Image Annotation:* The computer system automatically assigns the Meta-data to the digital image in a caption or keyword in the automatic annotation process. *Identity verification:* A person's identification and verification is used to authenticate a person by an organization or any other. The advancement in the various fields has required security for the organization, system etc. Therefore, biometric options are used to secure confidential information.

In *biometric recognition*, individual identification is based on the biological differentiation patterns of fingerprints, retina and iris patterns, or ear has been used [205]. Eye-blink information is used for sensing emergencies. Therefore, a method is used to detect eye blinks [85]. In image forgery detection, copy-move forgery is one of the easiest forms of detection in

which detection is performed on some part of the image that is copied and moved to another place within the same image. This kind of detection is used to compare the features of the original image with forgery images [27]. Extraction of an object: image segmentation and object extraction are closely related terms. In the segmentation process, the whole image is divided into sub-parts based on colour, intensity etc.

On the other hand, object extraction is used to represent the object in a more meaningful representation. Therefore, image segmentation is performed on the image to separate various parts for object extraction. After that user selects the background or foreground region with the help of a marker, and then the algorithm segments the foreground from the background of the image. Most photo editors have used this technology to change the image's background. In the future, by improving this technology, object extraction will be used to extract the object from the video.

7 Challenges

Object detection models have been trained on the huge amount of data to detect the object from the input image easily. While detection has been performed, several types of parameters affect the performance of detection algorithms. Some of them are discussed below:

7.1 Input source

Object detection is impossible without image, video, or any real-time scene, so a good quality input source is required. Some of the input source-related issues are discussed over there.

- **Camera problems**: In object detection challenges, issues related to the camera are always affected. An image has been captured with the static camera with limited colours information, without focus or low resolution in a real-time environment that gives a low quality of the image or blurred etc. [124]. The captured images are not detected perfectly because the trained models cannot be trained on low-quality or blurred images.
- **Illumination:** While capturing the image, different things are affected by the image or object that appears in the image, such as environment, physical location, lighting (indoor, outdoor, dawn, dusk, weather condition, backgrounds, viewing distance etc.) [76, 132]. The images captured from the outdoor environment are challenging because the environment is uncontrolled. An image is also captured indoors containing shadow or false-positive effects due to sudden changes in light. These types of conditions are produce results with a variety of objects such as pose, blur, clutter, scale etc. [98].

7.2 Object related issues

While objects are detected from an image or any input source, several challenges are occurred related to the objects. Some of the challenges are discussed below.

- **Size of the object and aspect ratio:** An object is present small or big in the image or videos. Different type's interesting items are available in different sizes and aspect ratios. After classifying an object or item in the image, localization of such items is too difficult

[48]. Further added that different items can be viewed in different shapes, for example, the aspect ratio of the person standing person or sitting person.

- **Intrinsic factors:** Many variations of instances appear in the image of particular object categories, such as they are different according to the texture, colour, material etc. Therefore, the detection of these object instances from images, videos, or real-time environments, has been affected the accuracy of object detection approaches [98]. Furthermore, the objects blended with the image background are also very hard to detect from images.
- **Occlusion:** Occlusion is also one of the challenges in object detection. The image contains several objects, and each object is different from the other objects. As a result, some objects are hidden behind the other objects either partially or fully in captured images, known as occlusion [76]. For example, feature extraction is difficult for the dog object behind the car object.

7.3 Efficiency

With time, technology has been changed to fulfill the requirements of people. For example, portable devices are used while people travel from one place to another, and the demand for wearable devices has been increasing day by day or more. Nevertheless, storage capacity and computational capabilities have been limited in wearable devices, so making efficient object detection is critical in real-time [16]. The efficiency challenges start from rising computational complexity due to the increasing number of object categories, locations, and scaling of objects within a single image or real-time environment.

7.4 Scalability

Scalability is a significant challenge for object detection [216]. High data rates, previously unknown or unseen objects, and situations that the detector should handle are difficult. In addition, the number of images and categories of the object has been increasing, making it impossible to annotate the images manually and forced to rely upon supervised methods. Viewpoint variations are also a challenge, i.e., objects look completely different when they capture from different angles.

7.5 Speed of real-time detection

Speed is an important metric that has been considered to measure the performance of object detection models. It is because object detection depends not only on the accurate classification and localization of objects but also on some other factors such as speed. Object detection has been performed in a real-time environment from the videos, so a high detection speed model is a big challenge. Nowadays, researchers have proposed different object detectors, but they still do not meet human speed.

8 Discussion

This study surveyed many research articles and categorized them into application areas. Seven research questions are framed to determine the current object detection status to carry out this

object detection research. This article also tried to find the research studies according to application areas such as pedestrian detection, remote sensing object detection, text detection, traffic sign detection, etc. Different performance metrics related to object detection are also discussed. Our attention is broad compared to the earlier surveys and includes the latest published research articles. This study also explored all the sub-application areas of object detection along with datasets and metrics. This section tried to answer all the questions defined in the review method.

RQ1. What is object detection? How objects are detected from an image?

Object detection is the process of locating the object instances of specific categories from real-world data. Many research papers have been reviewed for the understanding of this concept. It is one of the challenging areas of computer vision. Two terms, object localization and classification, are combined to detect the objects from an image. In simple terms, computer vision uses three different tasks particular objects are picked from the background images. Secondly, identify the object class in which it belongs, and the last is proposed object boundaries are defined. Firstly, an algorithm is applied to the input to identify regions of interest to detect the objects from videos and images. Each object has different features or properties that help to differentiate the objects from one another. Nowadays, various authors have introduced many approaches, as shown in Tables 3 and 4. All features are extracted from input data according to the category of the object. Further, these extracted features are used for the classification of that object. From the past decades and onwards, noticeable work has been done in this field of computer vision. However, all the implementation details of the proposed techniques vary from traditional to deep learning. All the object detection algorithms typically use machine learning or deep learning for generating results. Nowadays, most deep-learning or machine learning-based techniques have been used in different CNN architectures developed by various authors of the worldwide web. Different challenges have occurred during object detection. According to the challenges and problems, CNN architectures have been resented by authors. The different state-of-the-art methods are discussed in section 3, i.e., techniques for objects detection.

RQ2. Which type of challenges occurred during the detection of objects?

Object detection usually gives two answers: “What is the object?” and “where is the object”. Object classification and localization were challenging tasks a decade ago. Due to computer vision, digital devices can detect or identify the contents of the images. Despite the noteworthy improvements, object detection still faces challenges such as dual priorities, limited data, class imbalance, size, speed, environment, multiple scales, etc. Many authors have devoted their research to overcoming these difficulties and providing remarkable results during this review, but some challenges exist. In real-world images, many variations are included, such as lighting, occlusions, noise, camera distortions, background clutter etc. Most of the object detection techniques work on the detection of small objects from complex backgrounds. With the use of image pyramids, detection of small as well as large objects becomes simple and effective. Regardless of all recent advancements, accuracy on small objects is still challenging in object detection. Other factors also affect the quality of object detection, such as training strategies, backbone models, improving loss functions, handling imbalance between the positive and negative sample etc. Despite the many architectures that have been proposed to

handle the different challenges. But still, challenges affect the performance of the object detector to meet with real-time performance like human beings. Tremendous research works have been done in object detection to handle the challenges in different application areas as shown in the Table 4 [102, 107, 123, 164, 168, 188, 194, 200, 208, 213, 215]. For example, Li. et al. have proposed the AITwo approach to detect the vehicle from foggy weather, but it still needs improvements to meet human accuracy.

RQ3. Why is annotation required while training a deep learning model and which annotation tools are widely used?

Annotation is a process that has been used to identify the existing data in different formats. In deep learning and machine learning, supervised learning is performed therefore, labelling of data is required for models. The machines can clearly and accurately interpret the annotation data for training. To train a machine or deep learning model, the right methods and techniques must be used to annotate data correctly. When the annotation is performed on images, developers add the metadata to datasets for training. Different things have been used to annotate the input data, such as polyline, bounding box, masking, polygon, landmarking, etc. This study reveals some annotation tools for labelling the data, such as MakeSense.AI, LabelImg, VOTT, VIA, KAT, Photo Stuff, AktiveMedia, Caliph & Emir, LabelMe, LabelBox etc. Deep learning models cannot imagine without a training dataset. The annotation process is very time-consuming, and a lot of effort is required for manual annotations. Nowadays, authors think about the automatic annotation process for data training, i.e., self-supervised training learning. Despite the image annotation, video annotation can also be used to annotate the videos, frame by frame or as a stream. After reviewing many publications during this study, the most frequent annotation tools are LabelImg, CVAT, VOTT, and LabelMe.

RQ4. What are the different types of techniques mostly used for object detection?

In this systematic review, different techniques of object detection are revealed. There are traditional and deep learning-based techniques that many authors have used. In the starting era, object properties such as object components, shapes, edges, etc. were considered for detection. Based on object properties, performing detection on complex objects does not provide better results. Therefore, machine learning-based approaches were used to detect objects. Different methods, such as the statistical model of appearance, wavelet features, and gradient-based representation, were considered for performing detection. From these methods, wavelet feature representation was widely used in many object detection applications. In wavelet representation, learning is performed by transforming the pixels of the image into a set of wavelet coefficients, e.g., the Haar wavelets method, which many researchers most commonly used to detect the objects such as face detection, pedestrian detection, and general object detection etc. because of its high computational efficiency. After that CNN, has been used to detect objects from an image. With the use of a forward propagates network in CNN, the image features of any location, are extracted. Further, the VJ detector technique used the combination of feature pyramids and the sliding window, which helps to detect the faces from images. At that time, multiple detectors such as HOG, DPM etc., were built based on the feature pyramid and sliding window approach. VJ and HOG have been specially designed to detect objects of fixed aspect ratio. Therefore, “Object Proposals” based approaches have been used to detect the object of fixed aspect ratio.

All the object proposal-based approaches should provide a high recall rate, improve precision, high localization accuracy and reduce the processing time. Deep learning-based techniques have been divided into parts, i.e., two-stage detectors such as RCNN, Fast RCNN, Faster RCNN etc. are the object proposals-based techniques as in Fig. 3. The two-stage detectors provide high accuracy as compared to one-stage detectors. With the development of deep regression, the one-stage detection approaches have been used to resolve the problems of multi-scale detection. GPU has been used for detection in this modern era because of its high computing power. One-stage detectors are easy and simple to understand and predict the object boundaries directly of the object. These detectors provide accurate results on larger objects but are not good for localizing small objects. Further, multi reference-based approaches have become popular due to the set of anchors boxes of various locations of an image that are pre-defined in training time of different sizes and aspect ratios. Based on these references, predication of the detection box of performing. Nowadays, most object detection frameworks have been used Multi-resolution and multi-reference techniques for large and small object detection on images or videos. After reviewing numerous studies, we analyzed that two-stage detection techniques provide higher accuracy than the one-stage detector. On the other hand, one stage detector is much faster than a two-stage detector. Due to their fast processing speed, one-stage detectors have been applied in many real-time applications. In recent studies, one of the big challenges is how to combine the advantages of both one-stage and two-stage detectors. Nowadays, most techniques use state-of-the-art object detectors, as shown in Table 4. Some research studies have used state-of-the-art object detectors as a based detector to improve the accuracy of the techniques. The object detector approach is selected according to the application and the problem. A deep learning approach might be chosen if a huge amount of training data and powerful GPU are available. Otherwise, machine learning might be a good choice.

> **RQ5**. Which evaluation metrics have been used to evaluate the performance of object recognition techniques?

Although, there are several types of techniques that the various researchers have proposed according to their problem or domain. How did the authors identify which technique provides the optimal results from various techniques? The answer to this question is performance metric. All these techniques have been evaluated with the use of various kinds of metrics. Therefore, in this study, various types of metrics are used by various authors to evaluate their proposed techniques. As shown in section 4, the different performance metrics are Precision, Recall, mAP, IOU, F1, SEN, and SPE. The most commonly used metrics are Precision and recall, which are used to validate the results of different approaches. These metrics help evaluate the quality or quantity of results.

> **RQ6**. How the term "dataset" is important for deep learning? What are the different datasets available for object detection?

The term "dataset" is a collection of structurally arranged data (text, number, or images) related to a particular work. Data is mainly multimedia in object detection, such as images or videos. In deep learning, a huge amount of data is required to solve complex problems. Most deep learning models will not work well on small data sets. This study reveals multiple data sets according to their applications area. The various authors use all the datasets to detect several interesting objects in that dataset. In this study, the most popular generic object detection

datasets such as PASCAL VOC (20 class), COCO (80 class), ILSVRC (200 class), and OID (500 class) are described and defined as some additional datasets according to their specific application has shown in Table 2. Different researchers validate the multiple versions of the same dataset. All the researchers have been using the dataset according to the application area of object detection or interest, as shown in Tables 3 and 4. Most of the researchers have also released their benchmarks to handle the challenges of that application area.

RQ7. What are the important application areas of object detection?

Object detection has been used in many applications, such as personal security in all productive places. In this systematic review paper, this study reviewed some popular applications of object detection, such as pedestrian detection, text detection, face detection, traffic light, and sign detection, and remote sensing target detection. Significant challenges remain in object detection's specific application area, and different authors are doing research according to the interested application. Object detection and recognition have been applied to many subproblems in each application area. For example, face detection is applied in sub-area such as smile detection, facial expression detection (sad, angry, happy), mobile phone security, home security etc. As shown in Table 4, the past three years' research articles have been analyzed. The target and reference column of Table 4 shows the different sub-application areas of object detection. Despite these areas, there are still many areas in which object detection will be applied in the future.

9 Conclusion and future scope

Object detection has been received more attention in the area of computer vision. There are numerous opportunities for unseen object detection applications and optimizations of techniques. This review paper has presented the various closely linked concepts with object detection. The study analyzed around 200 studies related to object detection that are published in various digital libraries. We covered several categories where object detection is applicable, such as pedestrian detection, face detection and recognition, text detection etc. Many recent studies show that object detection applications can benefit the modern era. Object detection is also playing a significant role for physically impaired people. Some of the important and most commonly used datasets, annotation tools, and applications description are presented in this review paper. This paper also provides an overview of object detection, the techniques used for object detection, applications, and some important topics related to object detection.

We analyzed that the VJ detector is most commonly used for face detection. But it accepts only fixed-size input images. HOG is produced for removing this limitation and detecting objects of any size. In the two-stage detector, region-based CNN is produced for better accuracy results. Retina net uses two backbone architectures for increasing speed and accuracy. Some important performance metrics used to analyze the result of techniques are also presented. Dataset is required to create a new object detection model. So in this study, steps to create a new corpus and existing datases are highlighted. Further, different annotation tools are listed with their features and license type. This review paper just gives a general overview of object detection techniques and provides a basic understanding of object detection. Object detection techniques can be explored more deeply based on application areas and resource requirements. Besides many advancements, still, object detection has future directions.

- Detecting the objects from the video is more challenging than still images due to appearance variations in video frames such as defocus, motion blur, truncation, occlusion, and fast motion. Numerous research has been performed with video data, but it still needs improvements in detection. Despite the many studies that improve the accuracy, some more effective and efficient feature extraction and motion estimation networks [97]. Researchers worldwide can also concentrate on more dynamic targets and more complex data for future research.
- 3D sensors take additional depth information for better utilization of 2D images, and knowledge of images extends towards the real world. Due to object detection's rapid and constant development, performance has been increased, but some directions still need to be analyzed and explored. Some of them are depth estimation, temporal sequence, generalization, etc. are the relevant area in 3D object detection, which are cues for future work on 3D object detection [134].
- Small object detection: small object detection has been a challenge in a large or real-time environment. Some of the applications of OD, such as small vehicles from real-time CCTVV cameras, detecting some important targets state of the military, ship detects from remote sensing images etc., are the research direction. Some of the other research directions may include the design of lightweight networks and visual attention mechanisms.
- One-stage detectors have a fast processing speed because they are more suitable for real-time applications. But the accuracy of one-stage detectors is low for high precision applications. On the other hand, two-stage detectors have provided high accuracy but are inefficient and more time-consuming. These one-stage and two-stage detectors must combine to take advantage of both detectors. Here is a big challenge: "how to bring together the detectors".
- A specific method for a certain dataset (domain) always performs high detection. Hence, universal or multi-domain detectors must work on different domains without prior information about the other new domain. Without affecting the performance of the detector is too difficult to transfer the domain.
- The training process requires a well-annotated dataset in the supervised object detection methods, which is more time-consuming and inefficient. Human beings annotate each object manually in large datasets, which requires a lot of effort and time for large datasets. Therefore, automatic annotation approaches are required to eliminate manual annotation in supervised object detection tasks, which means an unsupervised object detection process is needed.
- The object detection performance is improved with feature fusion that aggregates the multiple levels of features. In addition, different tasks have been performed simultaneously, such as segmentation, semantic segmentation, multi-pose estimation, multi-object tracking, and object detection. Therefore, multi-task learning is a challenging task due to maintaining and improving accuracy as well as processing speed.
- Multi-source information is present in the real world due to the advancement of social media and big data technology. Many sources in the real-world environment provide textual information along with image data that can also be helpful in object detection tasks. For example, road signs with text information. It is also one of the future directions in which multidisciplinary information assists researchers with object detection.
- Inadequacy of benchmark Dataset: Different effective and efficient datasets are presented according to their particular application, but some limitations are also presented in these

datasets. For example, there are datasets with a front view of the objects but do not contain the other views of the objects. Even if they contain, the number of objects of a particular type is limited. Most of the datasets contain all the particular application type objects but lacking in number. Some of the datasets are biased according to the particular region; hence they would not be effective for testing the neutral data. As a result, the model trained on that type of dataset cannot be appropriately classified. Therefore, a proper dataset is required that contains a vast amount of objects in number with different orientations and is unbiased according to the particular region.

Authors contribution Jaskirat Kaur and Dr. Williamjeet Singh performed material preparation, data collection, and analysis. Jaskirat Kaur writes the first draft of the manuscript.

Declarations

Conflict of interest The authors have no conflicts of interest to declare relevant to this article's content.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Afif M, Ayachi R, Pissaloux E, Said Y, Atri M (2020) Indoor objects detection and recognition for an ICT mobility assistance of visually impaired people. Multimed Tools Appl 79(41–42):31645–31662. https://doi.org/10.1007/s11042-020-09662-3
2. Alam A, Jaffery ZA (2020) Indian Traffic Sign Detection and Recognition. Int J Intell Transp Syst Res 18(1):98–112. https://doi.org/10.1007/s13177-019-00178-1
3. Bach M, Stumper D, Dietmayer K (2018) Deep Convolutional Traffic Light Recognition for Automated Driving, in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, vol. 2018-Novem, 851–858, https://doi.org/10.1109/ITSC.2018.8569522
4. Banerjee K, Notz D, Windelen J, Gavarraju S, He M (2018) Online Camera LiDAR Fusion and Object Detection on Hybrid Data for Autonomous Driving, in 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, vol. 2018-June, no. Iv, 1632–1638, https://doi.org/10.1109/IVS.2018.8500699
5. Becker BC, Ortiz EG (2008) Evaluation of face recognition techniques for application to facebook, in 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, pp. 1–6, https://doi.org/10.1109/AFGR.2008.4813471
6. Behrendt K, Novak L, Botros R (2017) A deep learning approach to traffic lights: Detection, tracking, and classification, in 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 1370–1377, https://doi.org/10.1109/ICRA.2017.7989163
7. Bhandari A, Prasad PWC, Alsadoon A, Maag A (2021) Object detection and recognition: using deep learning to assist the visually impaired, Disabil Rehabil Assist Technol, 1–9, 2019, Taylor & Francis, https://doi.org/10.1080/17483107.2019.1673834
8. Bhangale U, Patil S, Vishwanath V, Thakker P, Bansode A, Navandhar D (2020, Elsevier B.V.) Near real-time crowd counting using deep learning approach. Procedia Comput Sci 171(2019):770–779. https://doi.org/10.1016/j.procs.2020.04.084
9. Bochkovskiy A, Wang C, Liao HM (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection, [Online]. Available: http://arxiv.org/abs/2004.10934
10. Bouras C, Michos E (2022) An online real-time face recognition system for police purposes, in 2022 International Conference on Information Networking (ICOIN), IEEE, pp. 62–67, https://doi.org/10.1109/ICOIN53446.2022.9687212
11. Bouti A, Mahraz MA, Riffi J, Tairi H (2020, Springer Berlin Heidelberg) A robust system for road sign detection and classification using LeNet architecture based on convolutional neural network. Soft Comput 24(9):6721–6733. https://doi.org/10.1007/s00500-019-04307-6

12. Braun M, Krebs S, Flohr F, Gavrila DM (2019, IEEE) EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. IEEE Trans Pattern Anal Mach Intell 41(8):1844–1861. https://doi.org/10.1109/TPAMI.2019.2897684
13. Caesar H et al. (2020) nuScenes: A multimodal dataset for autonomous driving, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, no. March, pp 11618–11628, https://doi.org/10.1109/CVPR42600.2020.01164
14. Ch'ng CK, Chan CS (2017) Total-Text: a comprehensive dataset for scene text detection and recognition, in *2017* 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 935–942, https://doi.org/10.1109/ICDAR.2017.157
15. Chatterjee S, Zunjani FH, Nandi GC (2020) Real-time object detection and recognition on low-compute humanoid robots using deep learning, in 2020 6th International Conference on Control, Automation and Robotics (ICCAR), IEEE, pp. 202–208, https://doi.org/10.1109/ICCAR49639.2020.9108054.
16. Chen IK, Chi CY, Hsu SL, Chen LG (2014) A real-time system for object detection and location reminding with RGB-D camera, 2014 IEEE Int.Conf Consum. Electron., 412–413, https://doi.org/10.1109/ICCE.2014.6776063
17. Chen Z, Luo R, Li J, Du J, Wang C (2021, Taylor & Francis) U-Net based road area guidance for crosswalks detection from remote sensing images. Can J Remote Sens 47(1):83–99. https://doi.org/10.1080/07038992.2021.1894915
18. Chen Y, Wang W, Zhou Y, Yang F, Yang D, Wang W (2021) Self-training for domain adaptive scene text detection, in 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, pp. 850–857, https://doi.org/10.1109/ICPR48806.2021.9412558
19. Chen Z, Ouyang W, Liu T, Tao D (2021, Springer US) A shape transformation-based dataset augmentation framework for pedestrian detection. Int J Comput Vis 129(4):1121–1138. https://doi.org/10.1007/s11263-020-01412-0
20. Cheng G, Han J (2016) A survey on object detection in optical remote sensing images, ISPRS J Photogramm Remote Sens, 117, 11–28, Elsevier, https://doi.org/10.1016/j.isprsjprs.2016.03.014.
21. Cordts M et al. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding, in 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, 29(5), 3213–3223, https://doi.org/10.1109/CVPR.2016.350.
22. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection, in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE, pp. 886–893, https://doi.org/10.1109/CVPR.2005.177
23. Dam GC, Management A (2019) U. S. Geological survey grand canyon monitoring fiscal year 2019 Annual Project Report to the Glen Canyon Dam Adaptive Management
24. Dasiopoulou S, Giannakidou E, Litos G, Malasioti P, Kompatsiaris Y (2011) A survey of semantic image and video annotation tools, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. LNAI 6050, Springer, Springer, pp. 196–239
25. de Charette R, Nashashibi F (2009) Traffic light recognition using image processing compared to learning processes, in 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp. 333–338, https://doi.org/10.1109/IROS.2009.5353941
26. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database, in 2009 IEEE conference on computer vision and pattern recognition, IEEE, 248–255, https://doi.org/10.1109/CVPR.2009.5206848
27. Dhivya S, Sangeetha J, Sudhakar B (2020, Springer Berlin Heidelberg) Copy-move forgery detection using SURF feature extraction and SVM supervised learning technique. Soft Comput 24(19):14429–14440. https://doi.org/10.1007/s00500-020-04795-x
28. Dollar P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: A benchmark, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 304–311, https://doi.org/10.1109/CVPR.2009.5206631
29. Dollar P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art, in IEEE transactions on pattern analysis and machine intelligence 34(4), 743–761, https://doi.org/10.1109/TPAMI.2011.155
30. Dominguez-Sanchez A, Orts-Escolano S, Garcia-Rodriguez J, Cazorla M (2018) A New Dataset and Performance Evaluation of a Region-based CNN for Urban Object Detection, in 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1–8, https://doi.org/10.1109/IJCNN.2018.8489478.
31. Du F, Wang WL, Zhang Z (2020) Pedestrian detection based on a hybrid Gaussian model and support vector machine, Enterp Inf Syst, 1–12, Taylor & Francis, https://doi.org/10.1080/17517575.2020.1791363.

32. Dutta A, Zisserman A (2019) The VIA Annotation Software for Images, Audio and Video, in Proceedings of the 27th ACM International Conference on Multimedia, ACM, pp. 2276–2279, https://doi.org/10.1145/3343031.3350535.
33. Ertler C, Mislej J, Ollmann T, Porzi L, Neuhold G, Kuang Y (2019) The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale, Comput Vis Pattern Recognit, 1–17, [Online]. Available: http://arxiv.org/abs/1909.04422
34. Everingham M et al. (2006) The 2005 PASCAL Visual Object Classes Challenge, in Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. MLCW 2005. Lecture Notes in Computer Science, Springer Berlin Heidelberg, vol. 3944 LNAI, pp. 117–176, https://doi.org/10.1007/11736790_8.
35. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The Pascal visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4
36. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2015, Springer) The pascal visual object classes challenge: A Retrospective. Int J Comput Vis 111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5
37. Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model, in 2008 IEEE conference on computer vision and pattern recognition, IEEE, 1–8, https://doi.org/10.1109/CVPR.2008.4587597
38. Fregin A, Muller J, Krebel U, Dietmayer K (2018) The DriveU Traffic Light Dataset: Introduction and Comparison with Existing Datasets, in 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 3376–3383, https://doi.org/10.1109/ICRA.2018.8460737
39. Fu M, Huang Y (2010) A survey of traffic sign recognition, in 2010 International Conference on Wavelet Analysis and Pattern Recognition, IEEE, pp. 119–124, https://doi.org/10.1109/ICWAPR.2010.5576425
40. Fu K, Chang Z, Zhang Y, Xu G, Zhang K, Sun X (2020, Elsevier) Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS J Photogramm Remote Sens 161(January):294–308. https://doi.org/10.1016/j.isprsjprs.2020.01.025
41. Fu J, Zhao C, Xia Y, Liu W (2020) Vehicle and wheel detection: a novel SSD-based approach and associated large-scale benchmark dataset. Multimed Tools Appl 79(17–18):12615–12634. https://doi.org/10.1007/s11042-019-08523-y
42. Fu C, Liu W, Ranga A, Tyagi A, Berg AC (n.d.) DSSD : Deconvolutional Single Shot Detector
43. Gawande U, Hajari K, Golhar Y (2022) SIRA: scale illumination rotation affine invariant mask R-CNN for pedestrian detection. Appl Intell, no. 0123456789, Springer US, https://doi.org/10.1007/s10489-021-03073-z.
44. Ge Z, Wang J, Huang X, Liu S, Yoshie O (2021, Elsevier) LLA: loss-aware label assignment for dense pedestrian detection. Neurocomputing 462:272–281. https://doi.org/10.1016/j.neucom.2021.07.094
45. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite", in 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp. 3354–3361, https://doi.org/10.1109/CVPR.2012.6248074
46. Girshick R (2015) Fast R-CNN, in 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 1440–1448, https://doi.org/10.1109/ICCV.2015.169
47. Girshick R, Donahue J, Darrell T, Malik J (2015) Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell 38(1):142–158. https://doi.org/10.1109/TPAMI.2015.2437384
48. Godinho De Oliveira BA, Ferreira FMF, Martins CAPDS (2018) Fast and lightweight object detection network: detection and recognition on resource constrained devices. IEEE Access 101(1):8714–8724. https://doi.org/10.1109/ACCESS.2018.2801813
49. Grosicki E, El-Abed H (2011) ICDAR 2011 - French Handwriting Recognition Competition, in 2011 International Conference on Document Analysis and Recognition, IEEE, pp. 1459–1463, https://doi.org/10.1109/ICDAR.2011.290
50. Guo Z, Liao W, Xiao Y, Veelaert P, Philips W (2021, Elsevier) Weak segmentation supervised deep neural networks for pedestrian detection. Pattern Recognit 119:108063. https://doi.org/10.1016/j.patcog.2021.108063
51. Gupta S, Thakur K, Kumar M (2021, Springer) 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. Vis Comput 37(3):447–456. https://doi.org/10.1007/s00371-020-01814-8
52. Hadid A, Heikkila JY, Silven O, Pietikainen M (2007) Face and Eye Detection for Person Authentication in Mobile Phones, in 2007 First ACM/IEEE International Conference on Distributed Smart Cameras, IEEE, pp. 101–108, https://doi.org/10.1109/ICDSC.2007.4357512
53. Halaschek-Wiener C, Golbeck J, Schain A, Grove M, Parsia B, Hendler J (2005) Photostuff - an image annotation tool for the semantic web, 4th Int. Semant. Web Conf. Poster Pap., pp. 2–4

54. Han J, Zhang D, Cheng G, Liu N, Xu D (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process Mag 35(1):84–100. https://doi.org/10.1109/MSP.2017.2749125
55. Han C, Gao G, Zhang Y (2019) Real-time small traffic sign detection with revised faster-RCNN. Multimed Tools Appl 78(10):13263–13278. https://doi.org/10.1007/s11042-018-6428-0
56. Harzallah H, Jurie F, Schmid C (2009) Combining efficient object localization and image classification, in 2009 IEEE 12th International Conference on Computer Vision, IEEE, pp. 237–244, https://doi.org/10.1109/ICCV.2009.5459257
57. He K, Zhang X, Ren S, Sun J (Sep. 2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
58. He W, Zhang X-Y, Yin F, Luo Z, Ogier J-M, Liu C-L (2020, Elsevier Ltd) Realtime multi-scale scene text detection with scale-based region proposal network. Pattern Recognit 98:1–14. https://doi.org/10.1016/j.patcog.2019.107026
59. He K, Gkioxari G, Dollar P, Girshick R (2020) Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 42(2):386–397, IEEE. https://doi.org/10.1109/TPAMI.2018.284
60. Heitz G, Koller D (2008) Learning spatial context: using stuff to find things, in European conference on computer vision, Springer, Berlin, Heidelberg, 30–43, https://doi.org/10.1007/978-3-540-88682-2_4.
61. Hosni Mahmoud HA, Mengash HA (2021, springer) A novel technique for automated concealed face detection in surveillance videos. Pers Ubiquitous Comput 25(1):129–140. https://doi.org/10.1007/s00779-020-01419-x
62. Houben S, Stallkamp J, Salmen J, Schlipsing M, Igel C (2013) Detection of traffic signs in real-world images: The German traffic sign detection benchmark, in The 2013 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1–8, https://doi.org/10.1109/IJCNN.2013.6706807
63. Hu J, Zhao Y, Zhang X (2020) Application of transfer learning in infrared pedestrian detection, in 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), IEEE, pp. 1–4, https://doi.org/10.1109/ICIVC50857.2020.9177438
64. Hua X, Wang X, Rui T, Zhang H, Wang D (2020, Elsevier B.V.) A fast self-attention cascaded network for object detection in large scene remote sensing images. Appl Soft Comput 94:106495. https://doi.org/10.1016/j.asoc.2020.106495
65. Huang Z et al. (2019) ICDAR2019 competition on scanned receipt OCR and information extraction, Proc Int Conf Doc Anal. Recognition, ICDAR, pp. 1516–1520, https://doi.org/10.1109/ICDAR.2019.00244.
66. Huang Q, Cai Z, Lan T (2021, IEEE) A single neural network for mixed style license plate detection and recognition. IEEE Access 9:21777–21785. https://doi.org/10.1109/ACCESS.2021.3055243
67. Hung BT (2021) Face recognition using hybrid HOG-CNN approach, in International Journal of Image and Graphics, 1254, 715–723
68. Hung GL, Bin Sahimi MS, Samma H, Almohamad TA, Lahasan B (2020, Springer) Faster R-CNN deep learning model for pedestrian detection from drone images. SN Comput Sci 1(2):116. https://doi.org/10.1007/s42979-020-00125-y
69. Irbaz MS, Al Nasim MA, Ferdous RE (2022) Real-time face recognition system for remote employee tracking. Lecture Notes on Data Engineering and Communications Technologies 95:153–163
70. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2014) Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition, pp. 1–10, [Online]. Available: http://arxiv.org/abs/1406.2227
71. Jakob J, Tick J (2020) Camera-based on-road detections for the visually impaired. Acta Polytech Hungarica 17(3):125–146. https://doi.org/10.12700/APH.17.3.2020.3.7
72. Jamtsho Y, Riyamongkol P, Waranusast R (2021, Elsevier B.V.) Real-time license plate detection for non-helmeted motorcyclist using YOLO. ICT Express 7(1):104–109. https://doi.org/10.1016/j.icte.2020.07.008
73. Jin Y, Zhang Y, Cen Y, Li Y, Mladenovic V, Voronin V (2021, Elsevier Ltd) Pedestrian detection with super-resolution reconstruction for low-quality image. Pattern Recognit 115:107846. https://doi.org/10.1016/j.patcog.2021.107846
74. Karatzas D, Mestre SR, Mas J, Nourbakhsh F, Roy PP (2011) ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email), in 2011 International Conference on Document Analysis and Recognition, IEEE, pp. 1485–1490, https://doi.org/10.1109/ICDAR.2011.295.
75. Kaur RP, Kumar M, Jindal MK (2022) Performance evaluation of different features and classifiers for Gurumukhi newspaper text recognition. J Ambient Intell Humaniz Comput no. 0123456789, Springer, https://doi.org/10.1007/s12652-021-03687-8
76. Khurana K, Awasthi R (2013) Techniques for object recognition in images and multi-object detection. Int J Adv Res Comput Eng Technol 2(4):1383–1388

77. Klare BF et al. (2015) Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A, in 2015 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp. 1931–1939, https://doi.org/10.1109/CVPR.2015.7298803.
78. Kostinger M, Wohlhart P, Roth PM, Bischof H (2011) Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization, in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, pp. 2144–2151, https://doi.org/10.1109/ICCVW.2011.6130513
79. Kumar R, Kumar S, Chand P, Lal S (2014) Object detection and recognition for a pick and place robot, in IEEE Asia-Pacific world congress on computer science and Engineering, 2014, 2–9, https://doi.org/10.13140/2.1.4379.2165
80. Kumar A, Kumar M, Kaur A (2021, Springer) Face detection in still images under occlusion and non-uniform illumination. Multimed Tools Appl 80(10):14565–14590. https://doi.org/10.1007/s11042-020-10457-9
81. Kuznetsova A, Maleva T, Soloviev V (2020) Detecting Apples in Orchards Using YOLOv3 and YOLOv5 in General and Close-Up Images, in Neurocomputing, 149, no. Part A, 233–243
82. Kuznetsova A et al (2020, Springer) The open images dataset V4. Int J Comput Vis 128(7):1956–1981. https://doi.org/10.1007/s11263-020-01316-z
83. LabelBox (2018) https://github.com/Labelbox/Labelbox/blob/master/README.md.
84. Lam D et al. (2018) xView: Objects in Context in Overhead Imagery, [Online]. Available: http://arxiv.org/abs/1802.07856
85. Lamba PS, Virmani D, Castillo O (2020, Springer Berlin Heidelberg) Multimodal human eye blink recognition method using feature level fusion for exigency detection. Soft Comput 24(22):16829–16845. https://doi.org/10.1007/s00500-020-04979-5
86. Laroca R, Zanlorensi LA, Gonçalves GR, Todt E, Schwartz WR, Menotti D (2021, wiley) An efficient and layout-independent automatic license plate recognition system based on the YOLO detector. IET Intell Transp Syst 15(4):1–21. https://doi.org/10.1049/itr2.12030
87. Learned-Miller E, Jain V (2010) FDDB : a benchmark for face detection in unconstrained settings
88. Li J, Liang X, Wei Y, Xu T, Feng J, Yan S (2017) Perceptual generative adversarial networks for small object detection, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, 1951–1959, https://doi.org/10.1109/CVPR.2017.211
89. Li K, Wan G, Cheng G, Meng L, Han J (2020, Elsevier) Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogramm Remote Sens 159(2019):296–307. https://doi.org/10.1016/j.isprsjprs.2019.11.023
90. Li F, Luo Z, Huang J, Wang L, Cai J, Huang Y (2020) AlTwo: Vehicle Recognition in foggy weather based on two-step recognition algorithm, in Neurocomputing 149, no. Part A, Springer, Springer, pp. 130–141.
91. Li C et al (2020, Elsevier B.V.) A parallel down-up fusion network for salient object detection in optical remote sensing images. Neurocomputing 415:411–420. https://doi.org/10.1016/j.neucom.2020.05.108
92. Liao J, Liu Y, Piao Y, Su J, Cai G, Wu Y (2022, Springer) GLE-Net: A global and local ensemble network for aerial object detection. Int J Comput Intell Syst 15(1):2. https://doi.org/10.1007/s44196-021-00056-3
93. Lin T-Y, Goyal P, Girshick R, He K, Dollar P (2017) Focal loss for dense object detection, in 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2999–3007, https://doi.org/10.1109/ICCV.2017.324.
94. Liu K, Mattyus G (2015) Fast multiclass vehicle detection on aerial images. IEEE Geosci Remote Sens Lett 12(9):1938–1942. https://doi.org/10.1109/LGRS.2015.2439517
95. Liu W et al. (2016) SSD: Single Shot MultiBox Detector, in European conference on computer vision, Springer, Springer, 21–37
96. Liu Z, Wang H, Weng L, Yang Y (2016) Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds, IEEE Geosci Remote Sens Lett vol. 13, no. 8, pp. 1074–1078, IEEE, https://doi.org/10.1109/LGRS.2016.2565705
97. Liu D, Cui Y, Chen Y, Zhang J, Fan B (2020, Elsevier B.V.) Video object detection for autonomous driving: motion-aid feature calibration. Neurocomputing 409:1–11. https://doi.org/10.1016/j.neucom.2020.05.027
98. Liu L et al (2020, Springer US) Deep Learning for Generic Object Detection: A Survey. Int J Comput Vis 128(2):261–318. https://doi.org/10.1007/s11263-019-01247-4
99. Liu Y, Liu J, Ning X, Li J (2022, Taylor & Francis) MS-CNN: multiscale recognition of building rooftops from high spatial resolution remote sensing imagery. Int J Remote Sens 43(1):270–298. https://doi.org/10.1080/01431161.2021.2018146
100. Lu Y, Lu J, Zhang S, Hall P (2018, Springer) Traffic signal detection and classification in street views using an attention model. Comput Vis Media 4(3):253–266. https://doi.org/10.1007/s41095-018-0116-x

101. Lu W, Zhou Y, Wan G, Hou S, Song S (2019) L3-Net: Towards Learning Based LiDAR Localization for Autonomous Driving, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, vol. 2019-June, 6382–6391, https://doi.org/10.1109/CVPR.2019.00655
102. Lu X, Ji J, Xing Z, Miao Q (2021) Attention and feature fusion SSD for remote sensing object detection. IEEE Trans Instrum Meas 70, https://doi.org/10.1109/TIM.2021.3052575
103. Lucas SM (2005) ICDAR 2005 text locating competition results, in Eighth International Conference on Document Analysis and Recognition (ICDAR'05), IEEE, vol. 2005, pp. 80–84 Vol. 1, https://doi.org/10.1109/ICDAR.2005.231.
104. Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R (2003) ICDAR 2003 robust reading competitions, in Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., IEEE, vol. 1, 682–687, https://doi.org/10.1109/ICDAR.2003.1227749
105. Lux M (2009) Caliph & Emir: MPEG-7 photo annotation and retrieval, MM'09 - Proc. 2009 ACM Multimed. Conf. with Co-located Work. Symp 925–926, https://doi.org/10.1145/1631272.1631456
106. Lv X, Su M, Wang Z (2021) Application of face recognition method under deep learning algorithm in embedded systems. Microprocess. Microsyst, 104034, Elsevier B.V., https://doi.org/10.1016/j.micpro.2021.104034
107. Ma C, Sun L, Zhong Z, Huo Q (2021) ReLaText: exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks. Pattern Recogn 111:107684. https://doi.org/10.1016/j.patcog.2020.107684
108. Madani M, Bagheri M, Sahba R, Sahba A (2011) Real time object detection using a novel adaptive color thresholding method, MM'11 - Proc. 2011 ACM Multimed. Conf. Co-Located Work. - Ubi-MUI 2011 Work. Ubi-MUI'11, pp. 13–16, https://doi.org/10.1145/2072652.2072656
109. Maeda H, Kashiyama T, Sekimoto Y, Seto T, Omata H (2021, wiley) Generative adversarial network for road damage detection. Comput Civ Infrastruct Eng 36(1):1–14. https://doi.org/10.1111/mice.12561
110. Manikandan NS, Ganesan K (2019) Deep learning based automatic video annotation tool for self-driving car, [Online]. Available: http://arxiv.org/abs/1904.12618
111. Masita KL, Hasan AN, Shongwe T (2022) Refining the efficiency of R-CNN in Pedestrian Detection. Lecture Notes in Networks and Systems 216:1–14
112. Mathias M, Timofte R, Benenson R, Van Gool L (2013) Traffic sign recognition - how far are we from the solution?, Proc Int Jt Conf Neural Networks, https://doi.org/10.1109/IJCNN.2013.6707049
113. Maze B et al. (2018) IARPA Janus Benchmark - C: Face Dataset and Protocol, in 2018 International Conference on Biometrics (ICB), IEEE, pp. 158–165, https://doi.org/10.1109/ICB2018.2018.00033
114. Mehedi Shamrat FMJ, Al Jubair M, Billah MM, Chakraborty S, Alauddin M, Ranjan R (2021) A Deep Learning Approach for Face Detection using Max Pooling, in 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, no. June, pp 760–764, https://doi.org/10.1109/ICOEI51242.2021.9452896
115. Mehta R, Ozturk C (2019) Object Detection at 200 Frames per Second, in Lecture Notes in Computer Science, 11133 LNCS, Springer, Springer, 659–675
116. Mei X, Hong Z, Prokhorov D, Tao D (2015, IEEE) Robust multitask multiview tracking in videos. IEEE Trans Neural Networks Learn Syst 26(11):2874–2890. https://doi.org/10.1109/TNNLS.2015.2399233
117. Melnyk P, You Z, Li K (2020, Springer Berlin Heidelberg) A high-performance CNN method for offline handwritten Chinese character recognition and visualization. Soft Comput 24(11):7977–7987. https://doi.org/10.1007/s00500-019-04083-3
118. Merkulova IY, Shavetov SV, Borisov OI, Gromov VS (2019, Elsevier Ltd) Object detection and tracking basics: Student education. IFAC-PapersOnLine 52(9):79–84. https://doi.org/10.1016/j.ifacol.2019.08.128
119. Mishra A, Alahari K, Jawahar C (2012) Scene Text Recognition using Higher Order Language Priors, in Procedings of the British Machine Vision Conference 2012, British Machine Vision Association, pp. 127.1–127.11, https://doi.org/10.5244/C.26.127
120. Mogelmose A, Trivedi MM, Moeslund TB (2012) Vision-based traffic sign detection and analysis for intelligent driver assistance systems: perspectives and survey. IEEE Trans Intell Transp Syst 13(4):1484–1497. https://doi.org/10.1109/TITS.2012.2209421
121. Murdock M, Reid S, Hamilton B, Reese J (2015) ICDAR 2015 competition on text line detection in historical documents, in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 1171–1175, https://doi.org/10.1109/ICDAR.2015.7333945
122. Nada H, Sindagi VA, Zhang H, Patel VM (2018) Pushing the Limits of Unconstrained Face Detection: a Challenge Dataset and Baseline Results, in 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, pp. 1–10, https://doi.org/10.1109/BTAS.2018.8698561
123. Naiemi F, Ghods V, Khalesi H (2021, Elsevier Ltd) A novel pipeline framework for multi oriented scene text image detection and recognition. Expert Syst Appl 170(2020):114549. https://doi.org/10.1016/j.eswa.2020.114549

124. Nayagam M, Ramar K (2015) A survey on real time object detection and tracking algorithms. International Journal of Applied Engineering Research 10(9):8290–8297
125. Nepal U, Eslamiat H (2022) Comparing YOLOv3, YOLOv4 and YOLOv5 for Autonomous Landing Spot Detection in Faulty UAVs. Sensors 22(2):464. https://doi.org/10.3390/s22020464
126. Neumann L et al. (2019) NightOwls: A Pedestrians at Night Dataset", in Computer Vision – ACCV 2018, vol. 11361, H. Li, G. Mori, and K. Schindler, Eds. Springer International Publishing, Springer International Publishing, pp. 691–705
127. Nguyen CC, Tran GS, Nghiem TP, Burie J-C, Luong CM (2019) Real-time smile detection using deep learning. J Comput Sci Cybern 35(2):135–145. https://doi.org/10.15625/1813-9663/35/2/13315
128. Nguyen N-D, Do T, Ngo TD, Le D-D (2020) An evaluation of deep learning methods for small object detection. J Electr Comput Eng 2020:1–18. https://doi.org/10.1155/2020/3189691
129. Ogura R, Nagasaki T, Matsubara H (2020) Improving the visibility of nighttime images for pedestrian recognition using in-vehicle camera. Electron Commun Japan 103(10):35–43. https://doi.org/10.1002/ecj.12268
130. Padilla R, Netto SL, da Silva EABB (2020) A Survey on Performance Metrics for Object-Detection Algorithms", in 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, vol. 2020-July, 237–242, https://doi.org/10.1109/IWSSIP48289.2020.9145130
131. Papageorgiou C, Poggio T (2000, Springer) Trainable system for object detection. Int J Comput Vis 38(1):15–33. https://doi.org/10.1023/A:1008162616689
132. Pattewar T, Chaudhari A, Marathe M, Bhol M (2019) Real-time object detection : a survey. Int Res J Eng Technol 06(04):231–237
133. Paul V, Michael J (2001) Robust real-time object detection. Int J Comput Vis 57:1–25
134. Qian R, Lai X, Li X (2021) 3D object detection for autonomous driving: A Survey 14(8), 1–24, [Online]. Available: http://arxiv.org/abs/2106.10823
135. Qin S, Liu S (2021) Towards end-to-end car license plate location and recognition in unconstrained scenarios. Neural Comput Appl, pp. 1–11, Springer, https://doi.org/10.1007/s00521-021-06147-8
136. Rahman MM, Al Mamun S, Kaiser MS, Islam MS, Rahman MA (2021) Cascade classification of face liveliness detection using heart beat measurement, in Advances in Intelligent Systems and Computing, vol. 1309, Springer, Springer, pp. 581–590
137. Ravishankar V, Vinod V, Kumar T, Bhalla K (2022) Sensor integration and facial recognition deployment in a smart home system, Springer, Springer, pp. 759–771
138. Razakarivony S, Jurie F (2016) Vehicle detection in aerial imagery : a small target detection benchmark. J Vis Commun Image Represent 34, 187–203, Elsevier, https://doi.org/10.1016/j.jvcir.2015.11.002
139. Redmon J, Farhadi A (2017) YOLO9000: Better, Faster, Stronger, in 2017 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.
140. Redmon J, Farhadi A (2018) YOLOv3: An Incremental Improvement, Comput Vis Pattern Recognit, 1–6, arXiv preprint arXiv:1804.02767, [Online]. Available: http://arxiv.org/abs/1804.02767
141. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, Real-Time Object Detection, in 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, 779–788, https://doi.org/10.1109/CVPR.2016.91
142. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
143. Risnumawan A, Shivakumara P, Chan CS, Tan CL (2014) A robust arbitrary text detection system for natural scene images. Expert Syst Appl 41(18):8027–8048, Elsevier. https://doi.org/10.1016/j.eswa.2014.07.008
144. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
145. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 77(1–3):157–173. https://doi.org/10.1007/s11263-007-0090-8
146. S SJ, P ER (2021, Elsevier GmbH) LittleYOLO-SPP: A delicate real-time vehicle detection algorithm. Optik (Stuttg) 225:165818. https://doi.org/10.1016/j.ijleo.2020.165818
147. Saathoff C, Schenk S, Scherb A (2008) KAT : the K-space annotation tool. Proccedings SAMT, 1–2
148. Sai Srinath NGS, Joseph AZ, Umamaheswaran S, Priyanka CL, Malavika Nair M, Sankaran P (2020, Elsevier BV) NITCAD - Developing an object detection, classification and stereo vision dataset for autonomous navigation in Indian roads. Procedia Comput Sci 171(2019):207–216. https://doi.org/10.1016/j.procs.2020.04.022

149. Sanchez JA, Toselli AH, Romero V, Vidal E (2015) ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset, in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 1166–1170, https://doi.org/10.1109/ICDAR.2015.7333944.
150. Sanchez JA, Romero V, Toselli AH, Villegas M, Vidal E (2017) ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset, in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 1383–1388, https://doi.org/10.1109/ICDAR.2017.226.
151. Santra S, Roy S, Sardar P, Deyasi A (2019) Real-time vehicle detection from captured images, 2019 Int. Conf. Opto-electronics. Appl Opt Optronix 2019, 1–4, IEEE, https://doi.org/10.1109/OPTRONIX.2019.8862323
152. Schöller FET, Plenge-Feidenhans'L MK, Stets JD, Blanke M (2019) Assessing deep-learning methods for object detection at sea from LWIR images, in IFAC-PapersOnLine, Elsevier Ltd, 52(21), 64–71, https://doi.org/10.1016/j.ifacol.2019.12.284
153. Setta S, Sinha S, Mishra M, Choudhury P (2022) Real-time facial recognition using SURF-FAST. Lecture Notes on Data Engineering and Communications Technologies 71:505–522
154. Shahab A, Shafait F, Dengel A (2011) ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images, in 2011 International Conference on Document Analysis and Recognition, IEEE, pp. 1491–1496, https://doi.org/10.1109/ICDAR.2011.296
155. Shao S et al. (2019) Objects365: A Large-Scale, High-Quality Dataset for Object Detection, in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, pp. 8429–8438, https://doi.org/10.1109/ICCV.2019.00852
156. Shao Z, Cheng G, Ma J, Wang Z, Wang J, Li D (2021) Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. IEEE Trans Multimed, pp. 1–1, https://doi.org/10.1109/TMM.2021.3075566.
157. Sharma N, Mandal R, Sharma R, Pal U, Blumenstein M (2015) ICDAR2015 Competition on Video Script Identification (CVSI 2015), in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 1196–1200, https://doi.org/10.1109/ICDAR.2015.7333950
158. Shashirangana J et al (2021, wiley) License plate recognition using neural architecture search for edge devices. Int J Intell Syst:1–38. https://doi.org/10.1002/int.22471
159. Shi Y, Zhang Z, Huang K, Ma W, Tu S (2020, Elsevier Inc) Human-computer interaction based on face feature localization. J vis Commun Image represent 70:1–6. https://doi.org/10.1016/j.jvcir.2019.102740
160. Song X et al. (2019) APOLLOCAR3D: a large 3D car instance understanding benchmark for autonomous driving, Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit vol. 2019-June, pp. 5447–5457, IEEE, https://doi.org/10.1109/CVPR.2019.00560
161. Sudha D, Priyadarshini J (2020, Springer Berlin Heidelberg) An intelligent multiple vehicle detection and tracking using modified vibe algorithm and deep learning algorithm. Soft Comput 24(22):17417–17429. https://doi.org/10.1007/s00500-020-05042-z
162. Sun Y et al. (2019) ICDAR 2019 Competition on Large-Scale Street View Text with Partial Labeling - RRC-LSVT, in 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 1557–1562, https://doi.org/10.1109/ICDAR.2019.00250
163. Sun P, Zheng Y, Zhou Z, Xu W, Ren Q (2020, Elsevier B.V) R4 Det: refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. Image Vis Comput 103:1–26. https://doi.org/10.1016/j.imavis.2020.104036
164. Sun F, Li H, Liu Z, Li X, Wu Z (2021, Taylor & Francis) Arbitrary-angle bounding box based location for object detection in remote sensing image. Eur J Remote Sens 54(1):102–116. https://doi.org/10.1080/22797254.2021.1880975
165. Sun X, Wang P, Wang C, Liu Y, Fu K (2021, Elsevier) PBNet: part-based convolutional neural network for complex composite object detection in remote sensing imagery. ISPRS J Photogramm Remote Sens 173:50–65. https://doi.org/10.1016/j.isprsjprs.2020.12.015
166. Susanto ER, Analia R, Sutopo PD, Soebakti H (2017) The deep learning development for real-time ball and goal detection of barelang-FC, in 2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA), IEEE, pp. 146–151, https://doi.org/10.1109/ELECSYM.2017.8240393.
167. Suzuki T, Kageyama Y, Ishizawa C (2020, wiley) Recognition method for speed limit signs and its applicability in recognition of vehicle entry prohibition signs at night. IEEJ Trans Electr Electron Eng 15(10):1–9. https://doi.org/10.1002/tee.23215
168. Tamilselvi M, Karthikeyan S (2022, Elsevier) An ingenious face recognition system based on HRPSM_CNN under unrestrained environmental condition. Alexandria Eng J 61(6):4307–4321. https://doi.org/10.1016/j.aej.2021.09.043

169. Tanner F et al. (2009) Overhead imagery research data set — an annotated data library & tools to aid in the development of computer vision algorithms, in 2009 IEEE Applied Imagery Pattern Recognition Workshop (AIPR 2009), IEEE, 1–8, https://doi.org/10.1109/AIPR.2009.5466304
170. Tarchoun B, Jegham I, Ben Khalifa A, Alouani I, Mahjoub MA (2020) Deep CNN-based Pedestrian Detection for Intelligent Infrastructure, in 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, pp. 1–6, https://doi.org/10.1109/ATSIP49331.2020.9231712
171. Tian Z, Zhan R, Wang W, He Z, Zhang J, Zhuang Z (2020, Taylor & Francis) Object detection in optical remote sensing images by integrating object-to-object relationships. Remote Sens Lett 11(5):416–425. https://doi.org/10.1080/2150704X.2020.1722330
172. Timofte R, Zimmermann K, Van Gool L (2014) Multi-view traffic sign detection, recognition, and 3D localisation, in Machine Vision and Applications, Springer, 25(3), 633–647, https://doi.org/10.1007/s00138-011-0391-3
173. Tousch A-M, Herbin S, Audibert J-Y (2012) Semantic hierarchies for image annotation: A survey, in pattern recognition, Elsevier, 45(1), 333–345, https://doi.org/10.1016/j.patcog.2011.05.017
174. Tran P, Pattichis M, Celedón-Pattichis S, LópezLeiva C (2021) Facial recognition in collaborative learning videos, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13053, no. 1613637, Springer, Springer, pp. 252–261
175. Tzutalin (2015) Labelimg, https://github.com/tzutalin/label.
176. Umer S, Rout RK, Pero C, Nappi M (2022, Springer) Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Humaniz Comput 13(2):721–735. https://doi.org/10.1007/s12652-020-02845-8
177. Varma S, Sreeraj M (2013) Object detection and classification in surveillance system, in 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS), IEEE, 299–303, https://doi.org/10.1109/RAICS.2013.6745491
178. Veit A, Matera T, Neumann L, Matas J, Belongie S (2016) COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images, [Online]. Available: http://arxiv.org/abs/1601.07140.
179. Vennelakanti A, Shreya S, Rajendran R, Sarkar D, Muddegowda D, Hanagal P (2019) Traffic Sign Detection and Recognition using a CNN Ensemble, in 2019 IEEE International Conference on Consumer Electronics (ICCE), IEEE, pp. 1–4, https://doi.org/10.1109/ICCE.2019.8662019
180. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features, in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, IEEE Comput. Soc, pp I-511-I–518, https://doi.org/10.1109/CVPR.2001.990517
181. Viola P, Jones MJ (2003, Springer) Robust real-time face detection. Int J Comput Vis 57(2):137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb
182. VoTT: Vott (visual object tagging tool) (2019) https://github.com/microsoft/VoTT/blob/master/README.md.
183. Wang K, Belongie S (2010) Word Spotting in the Wild, in 11th European Conference on Computer Vision, Springer, Springer, 591–604
184. Wang H, Miao F (2022, Taylor & Francis) Building extraction from remote sensing images using deep residual U-Net. Eur J Remote Sens 55(1):71–85. https://doi.org/10.1080/22797254.2021.2018944
185. Wang W, Shen J, Yang R, Porikli F (2018, IEEE) A unified spatiotemporal prior based on geodesic distance for video object segmentation. IEEE Trans Pattern Anal Mach Intell 40(1):20–33. https://doi.org/10.1109/TPAMI.2017.2662005
186. Wang J, Jiang S, Song W, Yang Y (2019) A Comparative Study of Small Object Detection Algorithms, in 2019 Chinese Control Conference (CCC), IEEE, vol. 2019-July, pp. 8507–8512, https://doi.org/10.23919/ChiCC.2019.8865157
187. Wang Y, Xie H, Zha Z, Xing M, Fu Z, Zhang Y (2020) ContourNet: Taking a Further Step Toward Accurate Arbitrary-Shaped Scene Text Detection, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 11753–11762, https://doi.org/10.1109/CVPR42600.2020.01177
188. Wang G, Zhuang Y, Chen H, Liu X, Zhang T, Li L, Dong S, Sang Q (2022) FSoD-net: full-scale object detection from optical remote sensing imagery. IEEE Trans Geosci Remote Sens 60(c):1–18. https://doi.org/10.1109/TGRS.2021.3064599
189. Wei X, Zhang H, Liu S, Lu Y (2020, Elsevier Ltd) Pedestrian detection in underground mines via parallel feature transfer network. Pattern Recognit 103:107195. https://doi.org/10.1016/j.patcog.2020.107195
190. Womg A, Shafiee MJ, Li F, Chwyl B (2018) Tiny SSD: A Tiny Single-Shot Detection Deep Convolutional Neural Network for Real-Time Embedded Object Detection, in 2018 15th conference on computer and robot vision (CRV), IEEE, 95–101, https://doi.org/10.1109/CRV.2018.00023.

191. Wu S, Zhang L (2018) Using popular object detection methods for real time forest fire detection, in 2018 11th International Symposium on Computational Intelligence and Design (ISCID), IEEE, pp. 280–284, https://doi.org/10.1109/ISCID.2018.00070
192. Wu X, Sahoo D, Hoi SCH (2020, Elsevier B.V.) Recent advances in deep learning for object detection. Neurocomputing 396:39–64. https://doi.org/10.1016/j.neucom.2020.01.085
193. Wu J, Zhou C, Zhang Q, Yang M, Yuan J (2020) Self-mimic learning for small-scale pedestrian detection, in Proceedings of the 28th ACM International Conference on Multimedia, ACM, pp. 1–9, https://doi.org/10.1145/3394171.3413634
194. Wu K, Bai C, Wang D, Liu Z, Huang T, Zheng H (2021, IEEE) Improved object detection algorithm of YOLOv3 remote sensing image. IEEE Access 9:113889–113900. https://doi.org/10.1109/ACCESS.2021.3103522
195. Wu J et al (2022, Elsevier) A multimodal attention fusion network with a dynamic vocabulary for TextVQA. Pattern Recognit 122(108214):1–10. https://doi.org/10.1016/j.patcog.2021.108214
196. Xia GS et al. (2018) DOTA: a large-scale dataset for object detection in aerial images, Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 3974–3983, https://doi.org/10.1109/CVPR.2018.00418
197. Xiao Y et al (2020) A review of object detection based on deep learning. Multimed. Tools Appl. 79(33–34):23729–23791. https://doi.org/10.1007/s11042-020-08976-6
198. Xu H, Guo M, Nedjah N, Zhang J, Li P (2022) Vehicle and pedestrian detection algorithm based on lightweight YOLOv3-promote and semi-precision acceleration. IEEE Trans Intell Transp Syst, 1–12, https://doi.org/10.1109/TITS.2021.3137253
199. Xu B et al (2022, Elsevier) CattleFaceNet: a cattle face identification approach based on RetinaFace and ArcFace loss. Comput. Electron Agric. 193:106675. https://doi.org/10.1016/j.compag.2021.106675
200. Xue C, Lu S, Hoi S (2022, Elsevier) Detection and rectification of arbitrary shaped scene texts by using text keypoints and links. Pattern Recognit 124:1–31. https://doi.org/10.1016/j.patcog.2021.108494
201. Yang B, Yan J, Lei Z, Li SZ (2015) Fine-grained evaluation on face detection in the wild, in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 1–7, https://doi.org/10.1109/FG.2015.7163158
202. Yang S, Luo P, Loy CC, Tang X (2016) WIDER FACE: A Face Detection Benchmark, in 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, 5525–5533, https://doi.org/10.1109/CVPR.2016.596.
203. Yao C, Bai X, Liu W, Ma Y, Zhuowen Tu (2012) Detecting texts of arbitrary orientations in natural images, in 2012 IEEE conference on computer vision and pattern recognition, IEEE, 1083–1090, https://doi.org/10.1109/CVPR.2012.6247787.
204. Ye Q, Doermann D (Jul. 2015) Text detection and recognition in imagery: a survey. IEEE Trans Pattern Anal Mach Intell 37(7):1480–1500. https://doi.org/10.1109/TPAMI.2014.2366765
205. Yuan L, Lu F (2018) Real-time ear detection based on embedded systems, in 2018 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE, 115–120, https://doi.org/10.1109/ICMLC.2018.8526987
206. Yucel MK, Bilge YC, Oguz O, Ikizler-Cinbis N, Duygulu P, Cinbis RG (2018) Wildest Faces: Face Detection and Recognition in Violent Settings, [Online]. Available: http://arxiv.org/abs/1805.07566
207. Yuliang L, Lianwen J, Shuaitao Z, Sheng Z (2017) Detecting curve text in the wild: new dataset and new solution, [Online]. Available: http://arxiv.org/abs/1712.02170.
208. Zakria Z, Deng J, Kumar R, Khokhar MS, Cai J, Kumar J (2022) Multiscale and direction target detecting in remote sensing images via modified YOLO-v4. IEEE J Sel Top Appl Earth Obs Remote Sens 15:1039–1048. https://doi.org/10.1109/JSTARS.2022.3140776
209. Zhang H, Hong X (2019) Recent progresses on object detection : a brief review, in Multimedia Tools and Applications, Multimedia Tools and Applications 78, no. June, 27809–27847, https://doi.org/10.1007/s11042-019-07898-2.
210. Zhang L, Ma J (2021) Salient object detection based on progressively supervised learning for remote sensing images. IEEE Trans Geosci Remote Sens 59(11):9682–9696. https://doi.org/10.1109/TGRS.2020.3045708
211. Zhang S, Benenson R, Schiele B (2017) CityPersons: a diverse dataset for pedestrian detection, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 4457–4465, https://doi.org/10.1109/CVPR.2017.474
212. Zhang J, Xie Z, Sun J, Zou X, Wang J (2020, IEEE) A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection. IEEE access 8:29742–29754. https://doi.org/10.1109/ACCESS.2020.2972338
213. Zhang X, Liu Y, Huo C, Xu N, Wang L, Pan C (2022) PSNet: perspective-sensitive convolutional network for object detection. Neurocomputing 468:384–395. https://doi.org/10.1016/j.neucom.2021.10.068

214. Zhao Z-QQ, Zheng P, Xu S-TT, Wu X (2019, IEEE) Object detection with deep learning: A Review. IEEE Trans. Neural Networks Learn. Syst. 30(11):3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865
215. Zhao X, Zhang J, Tian J, Zhuo L, Zhang J (2021, Taylor & Francis) Multiscale object detection in high-resolution remote sensing images via rotation invariant deep features driven by channel attention. Int J Remote Sens 42(15):5764–5783. https://doi.org/10.1080/01431161.2021.1931537
216. Zhou J, Yuqiao T, Li W, Wang R, Luan Z, Qian D (2019) LADet : A Light-weight and Adaptive Network for Multi-scale Object Detection, in Proceedings of The Eleventh Asian Conference on Machine Learning, 912–923.
217. Zhu Y, Du J (2021, Elsevier) TextMountain: accurate scene text detection via instance segmentation. Pattern Recognit 110:107336. https://doi.org/10.1016/j.patcog.2020.107336
218. Zhu Y, Jiang Y (2020, Elsevier BV) Optimization of face recognition algorithm based on deep learning multi feature fusion driven by big data. Image Vis Comput 104:104023. https://doi.org/10.1016/j.imavis.2020.104023
219. Zhu H, Chen X, Dai W, Fu K, Ye Q, Jiao J (2015) Orientation robust object detection in aerial images using deep convolutional neural network, in 2015 IEEE International Conference on Image Processing (ICIP), IEEE, pp. 3735–3739, https://doi.org/10.1109/ICIP.2015.7351502.
220. Zhu Z, Liang D, Zhang S, Huang X, Li B, Hu S (2016) Traffic-sign detection and classification in the wild, in 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, 2110–2118, https://doi.org/10.1109/CVPR.2016.232
221. Zou Z, Shi Z (2018) Random access memories: a new paradigm for target detection in high resolution aerial remote sensing images. IEEE Trans Image Process 27(3):1100–1111. https://doi.org/10.1109/TIP.2017.2773199