International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

An Evaluation of Deep Learning-Based Object Identification

[1]Johnson Kolluri and [2]Ranjita Das
[1]Research Scholar, Department of CSE, National Institute of Technology Mizoram, Aizwal, Mizoram, India.
[2]Assistant Professor, Department of CSE, National Institute of Technology Agartala, India.
Email Ids:[1]johnson.kolluri@gmail.com, [2]ranjita.nitm@gmail.com

Abstract
Identification of instances of semantic objects of a particular class, which has been heavily incorporated in people's lives through applications like autonomous driving and security monitoring, is one of the most crucial and challenging areas of computer vision. Recent developments in deep learning networks for detection have improved object detector accuracy. To provide a detailed review of the current state of object detection pipelines, we begin by analyzing the methodologies employed by classical detection models and providing the benchmark datasets used in this study. After that, we'll have a look at the one- and two-stage detectors in detail, before concluding with a summary of several object detection approaches. In addition, we provide a list of both old and new apps. It's not just a single branch of object detection that is examined. Finally, we look at how to utilize various object detection algorithms to create a system that is both efficient and effective. and identify a number of emerging patterns in order to better understand the using the most recent algorithms and doing more study.

Index terms: convolution neural networks, machine learning, deep learning, artificial intelligence, multi-layered neural networks.

I INTRODUCTION

Object identification and tracking has attracted a lot of attention in recent years owing to its wide range of applications and latest breakthrough research. It's common for object identification and tracking to be used together in both real-world and academic contexts[1]. security monitoring, Autonomous driving, robotic vision and transit surveillance are only a few examples of real-world applications[2]. Radar, LIDAR, and computer vision are just a few of the sensing technologies now accessible for monitoring and detecting moving objects. Image technology has advanced greatly in the last several years[3]. A new generation of cameras has emerged that is more affordable, smaller and more powerful than ever before. Computer processing power has also risen considerably during the same time period. Parallel computing technologies, such as and graphics processing units (GPU) and multi-core CPUs have been developed in recent years. Real-time Open CV object detection and tracking is now possible because of this technology's accessibility. Deep convolution neural networks (CNNs) and graphics processing units (GPUs) are two factors that have led to significant improvement in CV-based object recognition and tracking[3].

Deep learning (DL) and machine learning (ML) are both excellent subjects to study in this context, as are the characteristics that distinguish them. Pattern detection in data using examples or sample data is fundamental to machine learning (ML), an aspect of artificial intelligence (AI). In this case, the data is made available to the computer, allowing it to make inferences about the world around it[4]. Depending on the situation, the data (or samples) may be labelled, unlabeled, or a mix of both. As a result, learning may take place under supervised, unsupervised, or semi-supervised conditions[5]. Because of their ability to understand the relationship between input and output from examples, artificial neural networks (ANNs) are good candidates for machine learning applications. There are several other features of ANNs, including as adaptability, speed, robustness/ruggedness, and optimality[6]. Many developments in multi-layered neural networks (MLP) in the early 2000s laid the foundation for deep learning. DL refers to long-term, in-depth learning[7]. A subset of machine learning (ML), deep learning (DL) is a more extreme version of the former, taking it to new levels of complexity. Learning data representations distinguishes DL from task-specific methods. A popular deep architecture is the convolutional neural network (CNN), which is used for both image and video identification and learning.

To put it another way, object detection is a challenge of correctly categorizing individual objects and accurately anticipating their bounding boxes with a high degree of accuracy when seen via the DL framework[8]. The quantity of samples utilized in DL has an impact on the learning performance of the learner (or previous experiences). When the number is larger, it indicates that the performance is more accurate. The availability of vast data nowadays has resulted in DL being a relevant option. However, unlike traditional (shallow) learning, deep learning often requires

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

the use of hundreds or thousands of photos in order to get the greatest results. The phrase "shallow" refers to anything that is not very deep. Derivatives are thus computationally demanding and complex to design[9]. Very quick object identification and motion detection are only possible with a high-performance graphics processing unit (GPU).

Semantic items (such as people, buildings, and automobiles) may be detected in digital images and movies using an object detection algorithm[10]. Salient object detection[11], Edge detection[5], scene text detection[12], posture detection[13], pedestrian identification[14] and face detection[15] are some of the well-researched object detection domains. Security, military, transportation, healthcare, and many other aspects of daily life have all benefited from object detection. This list includes but is not limited to Caltech[16], KITTI[17], ImageNet[18], MS COCO[19] and a few more well-known object identification benchmarks, such as PASCAL VOC[20] and OpenImageV5[21]. Images and videos taken by a drone platform have been shared by the ECCV VisDrone 2018 [22]contest organisers.

The most common previous two-stage detector for picture object detection is Faster R-CNN[23]. One-stage detectors include YOLO [24] and SSD [25]. While one-stage detectors have the advantage of speedy inference, two-stage detectors have the advantage of excellent localization and object detection precision. Using the ROI pooling layer, two-stage detectors can be separated into two stages [23]. Faster R-CNN uses an RPN to identify regions where objects are most likely to be found. For classification and bounding-box regression tasks, the ROI Pool (ROI Pool) approach is used to collect information from each candidate box. It is possible to see in Fig.1 the basic design of a two-stage detector[26]. As a result, one-stage detectors [27]may be employed in real-time devices since they suggest predicted boxes straight from input pictures without requiring a region proposal phase. One-stage detectors' fundamental construction is shown in Fig.2.

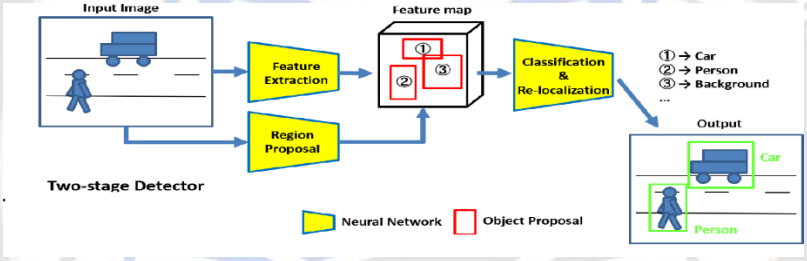

Figure 1: Two-stage detector architecture.

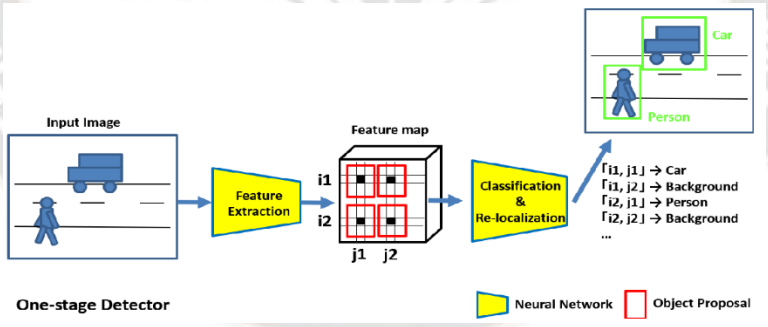

Figure 2: Architecture of single stage detectors.

This study's objective is to describe and evaluate a deep learning-based object detection task. In light of the quick advancements in computer vision research, previous surveys tend to focus on a limited number of domains, such as generic object identification, and may not include the most up-to-date approaches that provide some new answers and directions for these problems[28].

1. Readers will be able to see the cutting edge of the subject more clearly if they don't have to wade through the introductory material.
2. For the first time in a deep learning object detection survey, this study thoroughly and exhaustively analyses the most up-to-date detection solutions and a number of noteworthy research developments.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

3. To the best of our knowledge, some aspects of this research have been thoroughly examined and addressed for the first time.

We examine DL-based object recognition and tracking separately and in combination in the current review. Moreover, it examines which detector and tracker combinations are appropriate for different types of data. This review is the first of its type since it uses DL-based object recognition and tracking[29]. Because CV research is continually expanding, this article presents a systematic and complete examination of the features, functions, and performance of the many state-of-the-art approaches that are now available, as well as future directions in this area. To find the optimal detector-tracker combination models, we also wish to shed insight on how different DL models are used to broad object recognition, targeted item detection, and object tracking. As a result, it becomes easier to select deep models that work well for multi-object identification and tracking [30]. Following are applications for object detection and tracking, challenging research topics, and future prospects for DL. The last point is critical because it serves as a warning to those just getting started in deep learning and artificial intelligence research. Also included in this bibliography is a list of the most recent research studies on DL-based object recognition and tracking.

Here are the sections of this review study. Backbone network design is described in depth in Section 2, along with comparisons of several performance and parameter metrics. Section 3 provides an overview of the major object detection architectures that have been developed. Some of the most challenging issues in object detection are summarized in Section 4. A summary of the data and assessment criteria may be found in Section 5. The uses of object detection are summarized in Section 6. It is discussed in Section 7 where object detection will go in the future.

II BACKBONE NETWORKS IN OBJECT DETECTION

The backbone network acts as the main feature extractor when conducting an object detection task, collecting photos as input and outputting feature maps for each matching output image. With the exception of the final entirely linked layers, the backbone network used for detection is typically also the network used for classification [31]. In addition, there is a newer version of the fundamental categorization network. As an example, Zhao et al. [32] alter the number of levels or construct new layers entirely. To better meet particular requirements, certain studies [33] [34] leverage the newly created backbone for feature extraction.

People may select between highly linked backbones, such as ResNet [35], ResNeXt [36], AmoebaNet[37], and lighter backbones, such as SqueezeNet[38], MobileNet[39], ShuffleNet[40], MobileNetV2 and Xception [41], depending on their needs for accuracy vs. efficiency. Lightweight backbones can suit the needs of mobile devices. For real-time object detection, Wang and his colleagues [42] combine PeleeNet with SSD and optimize the architecture for performance. Complex backbones are required to fulfil the demands of high-precision and more precise applications. For real-time acquisitions such as video or webcam, it is necessary to have a backbone that has been thoughtfully developed in order to provide room for the detection architecture and find a balanced medium between speed and precision.

The shallower and sparsely connected predecessor is replaced with a deeper and more densely linked backbone to investigate more competitive detection accuracy. In Faster R-CNN [43], T.D.D. et al. [44] employ ResNet rather than VGG [45] to capture rich features due of its large capacity.

Object identification may be simplified and made more precise with the advent of new, high-performance classification networks. This is a great way to boost the network's overall speed since the backbone network also serves as a feature extractor. Feature quality is recognized to everybody, and hence it is a crucial stage that requires more study. For further information, see [46].

2.1 Architectures for object detections

The development of deep learning and increased computing power has both significantly contributed to the growth of object detection. There have been a number of significant developments since the first CNN-based object detector, R-CNN, that have significantly advanced the area of generic object detection. Each of these contributions has helped progress the field in its own unique way. For those who are just starting started, this section offers a concise introduction to the various object detection architectures that are available.

RCNN: The R-CNN algorithm is a CNN detector that focuses on regions. On PASCAL VOC datasets[20], Hafiz et al. [47] proved for the first time that a CNN may lead to considerably higher object identification performance than systems that depend on simpler HOG features. This was an important step forward in the field. The effectiveness and efficiency of deep learning as a technique for object detection have been shown. Four components make up the R-CNN detector. Category-independent region suggestions are generated by the first module. The second module takes each proposal for a region and generates a feature vector of a predetermined length. To categories the objects in a single picture, the third module uses a series of linear SVMs that are class-specific. A bounding-box regressor, the final

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

module, is used to precisely estimate the limits of a given object.

First, the authors use a selective search strategy to create region ideas. A convolutional neural network (CNN) is used in order to get a feature vector of 4096 dimensions for each potential location proposal. Because having input vectors of a certain length is necessary for a fully connected layer, the features of the area proposal should likewise be of the same length. Using a constant pixel size, the authors set CNN up with 227× 227 pixels. There are several factors that influence how large the region suggestions generated by the first module will be when used with different photos. A tight bounding box is drawn around the candidate area, and all pixels inside it are warped to the requisite size 227×227, regardless of its actual size or aspect ratio. The feature extraction network consists of a total of seven layers: five convolutional layers and two fully connected layers. The same CNN criteria apply to all subcategories as well. A SVM that does not share parameters with other SVMs may be trained for each category independently.

An efficient training technique for deep convolutional neural networks is to first train the network on a very broad dataset, then train it again on the particular dataset to fine-tune its performance. Initially the CNN was pre-trained on a big dataset by Hafiz et al (ImageNet classification dataset[18]). ImageNet-specific 1000-way categorization replaces the final fully-connected layer. Using stochastic gradient descent, the CNN's parameters that are applied to the warped proposal windows will be adjusted. A (N+1)-way object classification layer, where N is the total number of classes for the objects, and 1 is the background, is the last layer to be fully connected.

The writers put their examples into two categories: good instances and negative ones. There are a number of ways to fine-tune IoU (intersection over union) overlap thresholds. Region suggestions that fall below the cutoff point are said to be negative, while object proposals that rise above it are referred to as positive. The object suggestions allocated to this box must likewise have the highest IoU with a class that represents the ground truth. Setting the parameters is one of the options accessible while training an SVM. The "ground truth" boxes are used as models for the classes they belong to. Negative proposals for a class are those that overlap all of the ground-truth instances for that class by less than 0.3 IoU. These non-realistic alternative hypotheses with an overlap of 0.5 to 1 may be used to obtain about 30 extra examples of favorable outcomes. As a result, with such a large collection, overfitting may be avoided throughout the fine-tuning phase.

Fast RCNN: Fast RCNN [48] is an improved version of RCNN that fixes the runtime problem. Images are taken as input and feature maps are generated using the picture as an input. In the pooling-map, each feature is considered to be an area of interest (RoI).

After that, it is sent through three layers that are all completely connected together in order to classify the objects in the image and create a bounding box around each thing that is found. In comparison to RCNN, the calculation time may be greatly reduced since the locations of pooling features are employed for classification.

Additionally, RCNN has many stages of training, but Fast RCN has just one level of training. RCNN is more complex, while Fast RCN is simpler.

The RoI pooling-map is employed for classification instead of the input region recommendations, as previously stated. Several areas are shown on this map, each with its own set of notable characteristics. As a consequence of this, Fast RCNN does not need wrapping regions or reversing spatial features in order to provide recommendations for areas. By adjusting weight settings, a shortened single value decomposition (SVD) may be employed to speed up the detection process. On the PASCAL VOC 07[42] dataset, Fast RCNN achieved a mAP (mean average precision) of 66.7 percent. On the other hand, using RCNN yields a mAP of 66.0% on the same dataset. When compared to RCNN, Fast RCNN cuts the amount of time needed for training by a factor of nine. To comparison, the detection speed of RCNN is faster than that of fast RCNN trained on the shorter SVD datasets. This set of experiments makes use of an Nvidia K40 GPU. Fast RCNN outperforms RCNN in terms of detection performance measures, as shown by the aforementioned experiments. Fast RCNN, on the other hand, proposes its pooling map by using a selective search approach over the convolution feature map. This method causes it to operate more slowly than traditional RCNN.

Faster RCNN: Faster R-CNN [49] has been suggested three months after Fast R-CNN was first presented. The Fast R-CNN uses selective search, which is a slow procedure that takes about as long as the operation of the detection network, to suggest ROI. R-CNN is a competitor to RPN, or region proposal network, which anticipates region proposals successfully for a wide range of size and aspect ratio combinations. Full convolutional neural networks, or RPNs, are used. The detection network and RPN share the same set of convolutional layers and fully-image convolutional features, which allows RPN to generate region proposals more quickly. In addition, the use of multi-scale anchors as a reference provides a unique way for detecting objects of various sizes.

The usage of anchors allows for the generation of suggestions for regions of various sizes without the need for several scales of input images or characteristics. The center

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

of each feature window on the final shared convolutional layer's outputs (feature maps) slides a fixed-size window (3×3) relative to an input picture point that is the center of k (3×3) anchor boxes. This is done to make sure the final shared convolutional layer's outputs yield accurate findings. The authors believe that anchor boxes may be divided into three separate categories according to two of the criteria, namely size and aspect ratio. For the purpose of parameterizing the region proposal, a reference anchor box is employed.

Measure the distance between the anticipated box and the associated ground truth box, then go to step three. As a result, the anticipated box may be placed more precisely.

Using a quicker R-CNN enhanced both accuracy and detection efficiency, according to experiments. With shared convolutional calculations, Faster R-CNN obtained a mAP of 69.9 percent on the PASCAL VOC 2007 test set, while Fast R-CNN scored an overall mAP of 66.5 percent. Aside from that, the processing rate of Faster R-CNN was 5 frames per second, while the processing rate of Fast R-of CNN was 0.5 frames per second, with a total running period of 198 milliseconds. Both networks used the identical VGG [45] backbone.

R-FCN: One of the sub-networks in Faster RCNN is a shared, fully convolutional sub-network that is not reliant on ROI, while the other sub-network is an unshared, ROI-based network. Deep CNN models, such as AlexNet [50] and VGG16, enable quicker RCNN to provide more accurate results in a shorter amount of time. As an example of a fully convolutional network, ResNets [51] and GoogleNets [52] are already used for image categorization. To put it another way, the object detection networks built using ResNets and GoogleNets designs do not include a RoI network. With ResNets and GoogleNets, utilising Faster RCNN leads in worse performance. Due to the fact that image classification tasks are translational invariant, the object detection task cannot be translated but the image classification task may be translated. Images should be classified according to whether or not they include shifting objects, although any translation of an item in a bounding-box may be used for object identification. If the RoI pooling layer is manually introduced, there is a possibility that the translational invariance of the convolutional network will be affected. R-FCN was suggested as a remedy to this issue [53].

Position-sensitive score maps with a grid size of (gg) are produced for each item category by the R-final FCN's convolution layer. The answers from these score maps are then aggregated, and the final convolution layer is then added with a position-sensitive pooling layer. Following all of this, the g2 scores are averaged to produce a N+1-dimensional vector, which is then used to create the class-agnostic bounding boxes for each of the classes (N: number of object categories, 1: background). R-FCN passes the MS COCO and PASCAL VOC tests at a rate of 170 milliseconds per image.

Mask-RCNN: An example of this would be the segmentation method known as Mask R-CNN[54], which is an extension of the Faster R-CNN algorithm. Mask R-CNN may be considered as a more accurate object detector, regardless of the addition of a parallel mask branch. According to FPN[55], the features of interest to the RoI are retrieved using a backbone from various levels of the feature pyramid. This backbone offers outstanding accuracy and processing speeds.

The FPN is composed of bottom-up and top-down routes as well as lateral connections between them. Baseline ConvNets[56] calculate feature maps at two-scale scales in order to build up a hierarchical structure for the feature maps in the bottom-up pathway.

Higher pyramid levels are upsampled and their feature maps become physically coarser but semantically stronger using the top-down approach. The bottom-up route's last convolutional layer's output, which comes first, captures the top pyramid feature mappings.

Each of the top-down and bottom-up paths has a lateral connection that combines feature maps of the same spatial size. Convolutional layer 1×1 may alter the dimension even while feature maps have distinct dimensions. Predictions are produced for each new pyramid level that is formed after a lateral connection procedure. Due to the fact that lower-resolution feature maps are rich in semantic information and higher-resolution feature maps are necessary for detecting small objects, the feature pyramid network is in charge of compiling significant characteristics.

As illustrated in Fig. 1, a tiny feature map may be extracted from each RoI[30] using RoIAlign instead of pooling. Two phases of traditional RoI pooling are required to estimate the feature values for each bin using floating numbers. Applying quantization is the first stage in the process of deriving the coordinates of each RoI on feature maps from the input photos and the down sampling stride. These coordinates are derived from the input pictures. After the region of interest feature maps have been partitioned into bins, the feature maps are quantized such that they are all the same size. The RoI and the extracted features are out of alignment as a result of these two quantization procedures. Since this is the case, RoIAlign is able to steer clear of any quantization of the RoI boundaries or bins throughout these two steps. After computing the floating-number coordinates of the four regularly sampled sites, it next executes a bilinear interpolation operation in order to calculate the exact values of the features at those locations. This process is repeated

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

for each RoI feature map. The values for each bin may then be determined by aggregating the data using either the maximum or the average pooling technique. RoIAlign is shown in Fig. 1.

Experiments revealed that the accuracy was elevated as a result of the two previously mentioned enhancements. To better identify MS COCO, we used ResNet-FPN backbone and RoI Align operation, both of which enhanced box AP by 1.7% and 1.1%, respectively.

2.2 single stage detectors

In one-stage detectors, bounding boxes may be predicted over images without the need for a region proposal phase. This helps to speed up the detection process. Fig. 2 depicts the basic structure of a stage detector. Many one-stage detectors are available, including YOLOv2 [57], YOLO[33], YOLOv3[58], DSSD[59], SSD[60], RetinaNet[61], M2Det[62], DCN[63]and RefineDet[64]. The following sections go into detail about each of them:

YOLO: A one-stage object detector, YOLO (you only look once), was presented by Redmon et al. as an alternative to Faster RCNN. Real-time detection of complete photos and webcams is the most significant contribution. To begin, this pipeline is only capable of predicting less than 100 bounding boxes for each and every picture, but Fast R-selective CNN's search capability is capable of predicting 2000 region recommendations for each and every image. Second, YOLO frame identification is a regression problem; as a result, a unified architecture can directly extract characteristics from input photos to forecast bounding boxes and class probabilities. While Fast R-CNN and Faster R-CNN both operate at 0.5 frames per second and 7 frames per second, respectively, YOLO network operates at 45 frames per second on a Titan X GPU. YOLO network has a higher throughput than both of these models.

First, the input picture is divided into a S ×S grid using the YOLO pipeline. The identification of the object whose center a grid cell occupies is the responsibility of each individual grid cell. The confidence score, which indicates how accurate a box that contains that item really is, is calculated by multiplying the probability, denoted by the symbol "P(object)," by the IOU, which stands for "intersection over union." Each grid cell makes a prediction about the X, Y, W, and H border boxes (x, y, w, h), as well as their associated confidence scores and C-dimensional conditional class probabilities for C categories. The feature extraction network has two fully linked layers after the 24 convolutional layers. The authors use the top 20 convolutional layers, an average pooling layer, and a fully connected layer for pre-training on the ImageNet dataset. For improved detection, the entire network is utilized. In the detection step, you should increase the pre-training input resolution from 224×224 to 512×512 in order to capture more fine-grained visual data and to boost detection accuracy.

Localization mistake was shown to be the most significant contributor to prediction error in YOLO's experiments. While YOLO is three times faster than Fast R-CNN, it produces far more background false positives errors. On the PASCAL VOC dataset, Fast R-CNN and Faster R-CNN produced a mAP of 70.0 and 73.2 percent, respectively, but YOLO only managed a mAP of 63.4 percent at 45 frames per second (fps).

YOLOV2: One step up from YOLO, we have YOLOv2. It is decided that in order to increase both YOLO's speed and detection accuracy, the previous training task's decisions would be used in the new YOLOv2. The six tasks that make up YOLOv2 are as follows: batch normalization, high resolution classifier, convolution with anchor boxes, prediction of anchor box size and aspect ratio, convolution with fine-grained features, and multi-scale training. Batch normalization is one of the tasks. There's more information about this in the next section:

- Batch normalization: The SGD method is used to train YOLOv2. Minibatches are used in SGD's training process. The mean and variance of each mini-batch are calculated and utilised to activate. The last phase is normalising the activation of each mini-batch by comparing it to a mean of zero and a value of 1. Finally, the same distribution is used to sample all of the items in each of the minibatches. A batch normalisation[65] might be considered as the result of this procedure. It generates the same number of activations. By placing a batch normalising layer in front of each convolution layer in the YOLOv2 algorithm, it is possible to accomplish both convergence and regularisation. When compared to the first version of YOLO, the mAP is increased by 2% when BN is used in the new version.
- High resolution classifier: The YOLO backbone uses a (224x224) input resolution. This resolution is raised to (448 x 448 pixels) in YOLOv2's input. The increased resolution inputs need a network change for object detection. As a direct consequence of this, the classification network in YOLOv2 has been fine-tuned for an image with a resolution of 448 × 448 and 10 epochs. This results in a 4 percent increase to the mAP.
- Convolution with anchor boxes: In order to create region recommendations, Faster RCNN makes use of an anchor box. The bounding box is then predicted using these region suggestions parameterized with respect to the reference anchor box. The adoption of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

this prediction technique in YOLOv2 is advantageous. This technique includes calculating the class and object-ness scores for each anticipated bounding box. This method improves recollection by 7% while also lowering the mean absolute percentage by 0.3 percentage points.

- *Size and aspect ratio prediction of the anchor box:* YOLOv2 makes use of the k-means clustering technique on the training bounding boxes in order to get more accurate priors. The projected anchor box's Centre position is then defined using these priors. Cluster data is used to anticipate the dimensions of this anchor box's aspect ratio and size. As a result, the detection rate is increased.
- Fine grained features: YOLO was honed using (224 x224) photos, as previously mentioned. There are two versions of the Yolo v architecture: Yolov2 and Yolov1. YOLOv2 is re-trained using higher resolution photos (448x488) to locate tiny items. YOLOv2 employs both high- and low-resolution features in this retraining process by stacking neighboring characteristics into various channels. The detection mAP is increased by 1%.
- Multi-scale training: To make the network more robust to photos of varying sizes, it is advised that a new picture with a dimension size of between 320 and 608 pixels be used for each of every 10 batches. The picture should be picked at random. To put it another way, it suggests that a single network is capable of detecting at many degrees of precision. As an example, YOLOv2 earns a score of 78.4 percent mAP and 40 frames per second at a higher resolution, but YOLO accomplishes 63.44 percent mAP and 45 frames per second on VOC 07 at a lesser resolution. Despite its high speed and excellent detection accuracy, YOLOv2 can only recognize items with a high resolution and can only do so for a single category of item at a time.

YOLOV3: YOLOV3 [is the result of combining the greatest aspects of YOLOV1 and YOLOV2/9000. This allows users to enjoy the benefits of both systems. A mixture of the residual block[66], the feature pyramid network (FPN), and binary cross-entropy loss is used in the process of advancing YOLO to YOLOV3 status. Because of these updates, the detection network is now better equipped to deal with things that are more complicated (more categories, multi-size objects).

SSD: In a huge number of feature maps of various sizes, one stage detector that forecasts category scores and box offsets for a given set of default bounding boxes at each location. Additionally, this detector makes use of a standard set of default bounding boxes. Each feature map's bounding boxes have various sizes and aspect ratios from one another. Each feature map gains the capacity to be sensitive to the sizes of the objects it represents by deciding the scale of the default bounding boxes and keeping a constant distance between the top layer and the lowest layer. All object categories' offsets and confidences are projected for each default box. During training, the ground truth boxes were employed as negative examples as opposed to the usual bounding boxes. The authors employ "hard negative mining" to limit the ratio of negatives to positives to a maximum of 3:1 because there are so many default boxes that are negative. To do this, they select the default boxes with the greatest degree of confidence loss. The authors also employ data augmentation, which has been proven to significantly improve accuracy.

SSD512 [67]was determined to be on par with the VGG-16[68] backbone in terms of mean access time (mAP) and performance. SSD512 earned a mAP score of 85.4 percent on the PASCAL VOC 2007 test set and 85 percent on the PASCAL VOC 2012 test set, compared to Faster R-CNN (82.2 percent and 77.8 percent) and YOLO (82.2 percent and 79.3 percent) (VOC2012: 64.3 percent). SSD512 outperformed Faster R-CNN in all assessment criteria on the MS COCO DET dataset.

DSSD: A variant of SSD is known as the Single Shot Detector (SSD). In DDSD, SSD is employed as the backbone, while ResNet-101 is used in both the prediction and deconvolution modules. SSD is also used as the backbone in SSDD. To do element-by-element additions of the outputs of each prediction layer, a residual block is added to each prediction layer. The deconvolution module improves the feature map resolution to enable DSSD[59] to recognize small objects. By combining these two modules with the SSD, the DSSD is now capable of forecasting a specific group of objects of various sizes. The initial SSD model (ResNet-101) is developed using (513x513) images collected from the same dataset as ResNet-101 after the baseline network for DSSD has been built using ResNet-101[69]. The parameters of this learned SSD model are adjusted through training the de-convolution module. The DSSD513 model was proven to be accurate and useful in the testing using the PASCAL VOC dataset and the MS COCO dataset as test subjects. The mAP for the test dataset PASCAL VOC 07 is improved by 2.2% with the addition of the prediction and deconvolution modules to the SSD model.

RetinaNet: In July of 2021, Juan sales et al. presented RetinaNet[70] as a one-stage object detector that makes use of focused loss as a classification loss function. The R-CNN object detector has two stages. Initially, a few regions are

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

proposed, and then each site is categorized into one of a number of categories.

Consequently, object detectors with two stages have a higher degree of accuracy than detectors with one step, which imply a greater number of potential locations. There is a significant imbalance between the foreground and background classes when one-stage detectors are employed to train networks in order to attain convergence. As a result, the authors suggest the use of a loss function known as "focused loss," which may be used to reduce the weighting of losses applied to cases that are easily categorised. As a consequence of this, the detector is not bombarded with an excessively high number of straightforward negative data while it is being trained. Faster than prior one-stage detectors, RetinaNet also substantially alleviates the difficulty of training imbalanced positive and negative samples using one-stage detectors.

Comparatively, ResNet-101-FPN backbone with a RetinaNet obtained 44.2 percent AP on the MS COCO test-dev dataset, while DSSD513 only got 33.2 percent AP. ResNeXt-101-FPN surpassed DSSD513's 40.8 percent AP with ResNeXt-101. RetinaNet has increased the accuracy of detection of tiny and medium-sized objects.

M2Det: A multilayer feature pyramid network (MLFPN) is presented by Zhang et al.[71], which generates feature pyramids that are more successful in addressing a broad variety of scale differences among object instances. To get at final improved feature pyramids, the authors use a three-step process. As with FPN, the base feature is constructed by fusing multilevel characteristics gleaned from various levels of the backbone. After feeding the basis feature into a block of alternating joint Thinned U-shape Modules and Feature Fusion Modules in the second phase of the process, TUM decoder layers may be generated from the basis feature. By combining the decoder layers of comparable size, a feature pyramid with several levels of functionality is finally formed. Currently, development is being done on features that are both multi-scale and multi-level. Following the SSD architecture[72] is one of the last stages, which must be done so that comprehensive localization and classification of bounding boxes may be obtained.

RefineDet: Anchor refinement and object detection are the two components that make up RefineDet [73], which is comprised of two modules that are coupled to one another. These two modules are connected to one another by a transfer connection block so that the characteristics of one module may be transferred to the other module in order to enhance it. The training procedure is broken down into three stages: preprocessing, detection (which consists of two modules that are coupled to one another), and NMS.

For example, SSD, YOLO and RetinaNet all employ the same one-step regression procedure to get their final findings. When it comes to tiny items, the authors discover that using a two-step cascaded regression technique yields better predictions of hard identified objects and more precise object position data.

DCN: It's not possible to cover all of the object's pixels with a regular CNN's receptive field. The deformable kernel may be generated via deformable convolutional networks (DCNs) [74]. There are two types of DCN: DCNv2 and DCNv1[75], for example, Deformable convolution layers (DCNv2) are used instead of conventional convolution layers in DCNv2[76]. Using a learnable scalar value, all the deformable layers may be modified to improve the deformable effect and accuracy. DCNv2 performed much better than DCNv1 on the MS COCO test-dev dataset[77], achieving a mAP of 45.3 percent. Table 1 lists the benefits and drawbacks of single and two stage detectors[78].

Table 1: Summery of advantages and limitations of object detection architectures

Method	**Advantages**	**limitations**
RCNN	Extract picture characteristics using DCNNs; choose 2k area suggestions using a search method; Classify areas using SVM. Use the bounding box regressor to fine-tune your areas.	There is no end-to-end training accessible, and the speed of instruction is extremely slow. It also takes up a lot of space.
Fast RCNN	To extract the characteristics of the complete picture, use DCNNs. Extract 2,000 area ideas from the image using the selection search procedure, but map them to the feature maps instead; Utilizing the ROI Pooling layer to down sample the features of the area proposals will allow you to build maps with fixed-size feature areas. Make use of the multitasking capability to reduce loss.	Selective search is still sluggish in extracting area regions, and there is no end-to-end training.
Faster	In lieu of the selection search method, implement the	The performance is poor for both

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

RCNN	Region Proposal Network (RPN). The RPN collaborates with the backbone network to exchange feature maps. It is possible to get instruction from beginning to conclusion.	multi-scale and very small objects, and the detection speed is not fast enough to meet real-time requirements.
R-FCN	A feature fusion on several levels A Feature Pyramid Network is presented for multi-scale object identification as well as tiny object detection.	The detection speed is insufficient to fulfil the real-time needs.
Mask-RCNN	Instead of using the ROI pooling layer, use the ROI Align pooling layer, which enhances detection accuracy. To enhance detection accuracy, training in object recognition and segmentation should be combined. The ability to identify smaller targets is provided as a result.	In order to satisfy real-time requirements, the detection speed is not fast enough.
YOLO	It is proposed to use a unique single-stage detection network that can fulfil real-time criteria for detection speed.	The precision of detection is low, particularly when dealing with thick or tiny objects.
YOLO V2	To construct an anchor box, the k-means clustering technique is used to a new backbone network (DarkNet19).	Training that is difficult
YOLO V3	Fusion of many levels of features to improve accuracy across multiple scales; establishment of a new backbone network (DarkNet53).	Performance degrades as the IoU rises.
SSD	Anchoring mechanism operating on several scales and levels; detection technique using numerous layers.	Small object detection is difficult.
DSSD	Improve the accuracy of tiny item recognition using a multi-layer feature fusion process and up-sampling utilising deconvolution rather than conventional linear interpolation.	In comparison to SSD, detection speed slows down.
Retina Net	RetinaNet is a single, integrated network made up of a backbone network and two task-specific subnetworks. An off-the-shelf convolutional network serves as the backbone and is in charge of generating a convolutional feature map over an entire input image.	Working with each layer's feature map has the disadvantage of producing a lot of dense candidate frames, which leads to a lot of negative samples.
M2Det	When compared to previous approaches, M2Det benefits from the advantage of one-stage detection and suggested MLFPN structure, drawing a considerably superior speed-accuracy curve.	They have various restrictions as a result of the fact that they only build the feature pyramid according to the inherent multiscale and speed limitation.
RefineDet	High accuracy; high efficiency	Loss was high
DCN	DCN was created to improve the learning of explicit and bounded-degree cross features.	The majority of data in Web-scale applications is categorical, resulting in a huge and sparse feature space. In this situation, finding efficient feature crossings frequently needs manual feature engineering or a thorough search.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

2.3 CNN applications for specialized object detection

Object detection, one of the three core activities that comprise computer vision, may be applied to a broad variety of real-world scenarios. Depending on the unique demands, the implementation of object detection technologies in real-world application circumstances may appear fairly diverse. Important applications of object detection are discussed in this section, including salient item detection [79], face detection[80], image detection for remote sensing[81], pedestrian identification[82], and medical image detection[83].

Face detection: The most important application area for object detection is face detection, which also forms the basis for sentiment analysis, face alignment, gender identification, and face recognition. Other applications that rely heavily on face detection include face alignment and gender identification. In the actual world, there are several factors that might make facial identification a challenging detection task. These factors include variations in facial traits, illumination, motion, and occlusion.

The goal of face detection is to determine whether or not images contain faces and, if they do, to pinpoint those faces. The handcrafted feature extractor and the sliding window are the cornerstones of traditional face recognition. In order to determine the position of the face, sliding matching with the detected picture feature is also done using the face template feature. The VJ detection method, created by Viola and Jones in 2001[84], is one popular method. The detector's use of cascaded AdaBoost classifiers and Haar features [85,86] allowed for a significant improvement in both detection speed and accuracy. In addition, both ACF [87] and DPM [88] contribute to an improvement in the efficiency of face detection. On the other hand, traditional face recognition algorithms continue to have a number of shortcomings. Since the beginning of the age of deep learning, face recognition technology that is based on deep learning has shown exceptional performance.

The general object recognition capabilities of deep learning-based object identification algorithms have led to the development of numerous face detection approaches. Deep learning-based object identification algorithms are also quite successful in face detection. Rosa Andrie Asmara and colleagues introduced Cascade CNN [89] as a solution to the issue of sensitivity to light and angle in practical applications. Multiple distinct cascaded DCNN classifiers make up cascade CNN. The multi-task face detection system MTCNN [90] was created by Zhang et al. and uses a cascaded design very similar to the Cascade RCNN.

It utilises a three-part framework that incorporates face detection as well as face key detection. Faceness-Net[91], a method for coarse-to-fine detection that employs a number of DCNN-based network classifiers in order to recognize faces, is similar to Cascade CNN in this regard. Li hai et al. [92] suggested substituting a Face RCNN for a Faster RCNN and adding a centre loss that was computed using softmax to improve the classifier's overall performance. In order to overcome the difficulties presented by small objects and various scales in face recognition, baburoglu et al. devised a hybrid-resolution model[93] that analyzes photo pyramids in a way that is size-independent and utilizes a scaled hybrid detector. An SSH was created by agghey and colleagues [94] in order to enable multi-scale face recognition by performing detection on feature maps of varying sizes. Anchor approaches such as FaceBoxes [95], S3FD[96], and ScaleFace [97] are also able to handle the detection of tiny objects and multi-scale faces.

Salient object detection: Bringing attention to the principal object areas in an image, which are often referred to as the salient regions, is the objective of the technique of salient object detection. Detecting salient objects is an important application of object detection and computer vision that finds widespread use in a variety of contexts, including the interpretation of images and videos, the creation of computer graphics, and the guidance of robots.

During the time before deep learning. Itti et al.[98]provided the first saliency model for recognising spatially discontinuous elements in an image that was based on center-surround processes. Saliency object identification was helped along by the implementation of an approach proposed by Wang et al. [99] that included the substitution of binary segmentation for saliency detection.

Because DCNNs are capable of providing a decent representation of features, their use in the process of salient object detection is becoming more widespread. An MCDL architecture was proposed by Li and colleagues [100] that makes use of multi-layer perceptron (MLP) to first extract local and global contexts and then classify the foreground and background of an image. All of these methods are based on multi-layer perceptrons, including MAP, Super CNN, MDF and LEGS. Although multi-layer perceptron-based techniques improve performance, they are slow and insensitive to spatial information, and they need a lot of processing power. Complete convolutional networks are the foundation of the most powerful prominent object detection systems that are presently available. Recurrent fully convolutional networks (RFCN) were introduced by valipour and colleagues[101] in order to improve detection performance even more. Gomez et. al created a deep network set with the purpose of constructing a compact and uniform saliency map [102]. This map differentiates pixels from the object boundary. Using several pixel-supervised heuristic saliency algorithms, Feng et al.[103] proposed a

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

Deeplab-based DUS for the purpose of learning probable saliency and noise patterns. It is also possible to consult the review [104] for a more comprehensive summary of the situation.

Pedestrian detection: Detecting pedestrians is an essential component of intelligent surveillance systems, autonomous vehicles, and robotic navigation systems. The difficulties associated with detecting pedestrians are far more sophisticated than those associated with detecting generic objects. Face recognition is more difficult than pedestrian recognition because pedestrian objects contain the properties of both flexible and rigid objects. This makes pedestrian objects more subject to the impact of posture, thick occlusion, light, and viewing angle, which makes pedestrian recognition more difficult. The conventional method of identifying pedestrians, similar to the old method of detecting faces, relies on manually crafted feature extractors. For instance, Luo et al. released the HOG + SVM pedestrian detection approach at CVPR2005[105].

Remote sensing image detection: Image detection via the use of remote sensing is most often used in the context of military surveillance, urban planning and traffic navigation, land and resource assessment, and urban exploration. Aircraft, ships, automobiles, highways, airports, ports, and other structures are among the things. The following are the primary challenges in remote sensing image detection:

- The wide field of view of remote sensing photos results in great image quality, putting a premium on object recognition speed.
- Because the vast perspective leads in reduced item sizes relative to the picture, tiny object recognition is problematic for remote sensing images as well.
- The majority of the items in nature photographs are horizontal. The rotation invariance of the object is an essential problem when taking remote sensing photos from above.
- The context of the remote sensing picture is fairly intricate.

Deep learning techniques applied to remote sensing picture recognition are now being investigated as potential solutions for these issues. Wentong Li proposed YOLT [106], which is based on high-speed YOLOv2 and accelerates high-resolution remote sensing image detection by using two detections. In order to get around the difficulty of detecting incredibly minute objects, the feature map's resolution is raised concurrently. This was done to enhance the ability to find really minute details. Li and colleagues suggested using an R2-CNN to improve the detection of small objects in remote sensing images [107]. By including attention processes in the network, this was made possible. Multi-angle anchors were incorporated into the RPN by Bynum et al. [108] to address the issue of rotational invariance. Gong Cheng and his colleagues created a rotation invariant layer in order to address the challenges posed by rotation invariance[109]. This will be done in order to maximize accuracy.

Medical field: In the field of medicine, supplementary medical treatments include things like cancer detection, medical image detection, skin disease detection, sickness detection, and healthcare monitoring, amongst other things.

Computer-Aided Diagnosis (CAD) technologies might be of use to physicians in the process of categorizing the many types of cancer. After an acceptable picture has been captured, the important processes that must be completed by a CAD framework are image segmentation, feature extraction, classification, and object identification. These stages are defined by the following terms:

There is often a mismatch in the distribution of data between the source domain and the target domain. This is typically the case because there are significant individual variances, data scarcity, and privacy issues. A domain adaptation framework[110] is necessary in order to perform medical picture identification.

Life field: Intelligent homes, commodity detection, event detection, pattern detection, photo caption creation, rain/shadow detection, species identification, and other applications are the most often used ones in the life sector. Other applications in this sector include:

The purpose of event detection is to locate real-world events on the Internet, such as elections, festivals, speeches, and protests. Other types of events that may be detected include natural disasters and demonstrations. The variety of data kinds is more than it has ever been previously as a direct result of the widespread usage of social media and the new characters it has introduced.

Multi-domain event detection (MED) produces extensive event descriptions. In their paper[111], Yang et al. provide a framework for the event detection of dealing with data from several domains. Using the construction of affinity graphs, Shi et al.[112] combine aspects of online social interaction with activities involving event detection. Xiao et al. [113]develop a multimodal graph-based system for identifying events in a dataset consisting of one hundred million still images and moving movies. Please refer to the results of the poll [114] for any more information.The recognition of patterns is always impeded by a variety of variables including scene occlusion, changes in position, variations in light, and sensor noise. In order to improve recurrent pattern or periodic structure detection, researchers develop resilient baselines in both two-dimensional photographs and three-dimensional point clouds.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

The production of a caption for an image using a computer is referred to as "picture caption generation," and it is a process known as "caption generation." The most important step is to parse out the semantic information included inside photographs and then transform that data into natural languages. The creation of image captions requires the complex melding of natural language processing (NLP) and computer vision technology, which is a challenging endeavour in and of itself. Common approaches to solving this issue include encoder-decoder frameworks, attention mechanisms, multimodal embedding, and reinforcement learning. Combining Long Short-Term Memory architecture with Graph Convolutional Networks, which they refer to as GCN-LSTM, is the unique approach that Hou and colleagues [115]take to the problem of investigating object interactions.

Endalie et al. [116] provide a novel rain model in addition to a deep learning architecture for the purpose of determining whether or not a single photograph contains precipitation. Yun Zhou et al. [117] build a one-of-a-kind deep neural network to identify shadows by conducting an analysis of the spatial visual environment in a manner that is sensitive to the direction of light. An accurate identification of species is the cornerstone of taxonomic research, and a recent study [118] recommends using a deep learning technique to accomplish the task of species identification.

III BENCHMARK DATASETS

Static photos are needed for detection, whereas videos are needed for tracking. A variety of datasets, including ImageNet, MS COCO and PASCAL VOC are used for the identification of generic objects. Tracking is done using MOT2015 and MOT2016. In the next sections, we'll go through all of these datasets, as well as their performance indicators.

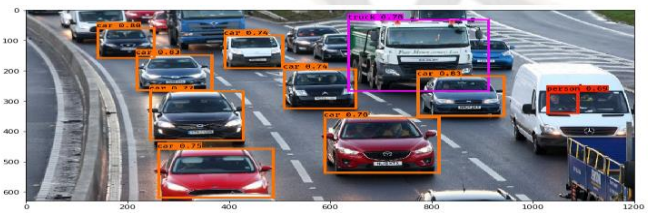

Figure 3. Object detection boundary boxes.

Detecting an item requires identifying the object's class and pinpointing its exact location inside the picture. Fig. 3 shows an example of a common bounding box for an object's location. It's important to utilize difficult data sets as benchmarks since they allow researchers to compare various methods and establish objectives for their answers. Face identification was the focus of early algorithms, which utilized a variety of ad hoc datasets. Face recognition datasets that were more realistic and difficult to use were later developed. Detecting pedestrians is another common problem, for which a number of datasets have been developed. The Caltech Pedestrian Dataset includes samples that are labelled and have bounding boxes associated with them. MS COCO, ImageNet and PASCAL VOC are some of the common benchmarks that are used for object identification. There are several more benchmarks. The performance of detectors that make use of the same dataset is evaluated, for the most part, via the use of the official metrics.

3.1 PASCAL VOC

Series 07 through 12 of the PASCAL VOC [76] may be found. There are 5 thousand training and 5 thousand test pictures in PASCAL VOC 07 PASCAL VOC 12, on the other hand, has 5.7K training and 5.7K test pictures. At least a half-dozen different types of things are included in each collection: automobiles and people; bicycles and people; buses and locomotives; cats and birds; horses and kites; lambs and boats; sofas and televisions. There are four primary branches to these 20 categories: cars, people, animals, and domestic things. A total of 27,000 items are tagged as bounding boxes in PASCAL VOC datasets. In Fig. 3, you can see several annotated photos. The VOC2007 dataset has unbalanced datasets, but the class human is clearly the largest, with a size approximately 20 times greater than the training set's smallest class sheep. How can detectors effectively address this issue in the context of the surrounding environment? Detectors must also be able to distinguish between multiple views (front, back, left/right, and undefined).

3.2MS COCO

There are at least 5,000 labelled instances for more than half of the 91 common item categories that are included in the Microsoft Common Objects in Context (MS COCO) dataset. The PASCAL VOC dataset has 20 categories, and these categories cover all of them. In all, there are 2,500,000 occurrences of each tagged in 328,000 photos in this dataset. All items in the MS COCO dataset are located in real contexts, which provides us with a wealth of contextual information. Figure 4 shows the examples from the PASCAL VOC dataset

While ImageNet[18] contains more categories, there are less occurrences per category in COCO. The dataset also includes a greater average number of instances per category than the PASCAL VOC datasets and the ImageNet object identification dataset (1k). This is because the dataset has 27k examples on average for each category. MS COCO has 7.7 object instances for every image, which is much more than PASCAL VOC and ImageNet (3.0). The PASCAL dataset includes 1.4 categories per image, whereas the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

ImageNet dataset has 1.7 categories per picture. The average number of categories in a photo in the MS COCO dataset is 3.5. While just 10% of the images in MS COCO fall under a single category, more than 60% of the images in ImageNet and PASCAL VOC are made up of just one sort of object. Small items require a higher level of contextual reasoning to recognize, as is well known. There's a lot of useful contexts packed within the images in the MS COCO dataset. Nearly 800,000 people and 600 hair drivers make up the largest and smallest classes in this dataset respectively. Nearly 800 people fall within the category of hair brush. Only 20 classes have many or few instances, while the rest 71 classes have about the same number of occurrences. The first two lines of Fig. 5 illustrate three typical picture categories from the MS COCO dataset.

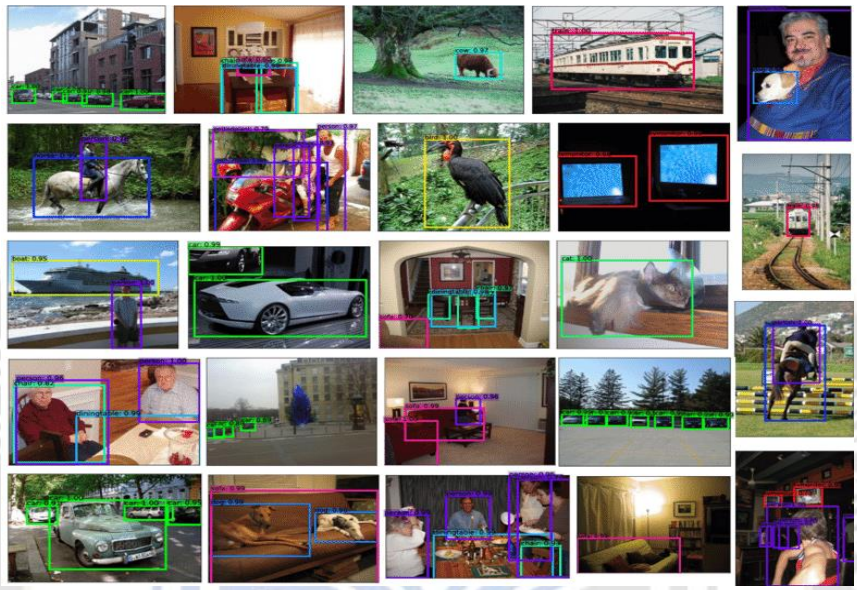

Figure 4: Examples from the PASCAL VOC dataset.

Figure 5: Three categories of MS COCO dataset.

3.3ImageNet dataset

Vision tasks and practical applications may be pushed to new heights by challenging datasets. Datasets like ImageNet[51] are useful for large-scale benchmarks[120]. A test in the ILSVRC on object detection measures an algorithm's ability to identify and locate all occurrences of a given target item in an image. Nearly 450 thousand training photos, 20 thousand validation images and 40 thousand test images are included in ILSVRC2014's 200 object classes. Figure 6 depicts some of the examples of ImageNet dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

Figure 6: ImageNet one sample image from each category.

3.4 MOT

It is frequently utilized in cutting-edge research since it contains 11 films, each of which has either one or two object types, either a human or an automobile. The MOT is nearing MOT is divided into two parts: MOT2015 and MOT2016. Each scene in MOT has distinct distributions of pedestrians that may be detected by MOT [119]. There are a total of 22 and 16 video clips included in this year's MOT2015 and MOT2016 [120], respectively. One half of each of these scenes from each film is used for instructional reasons. The remainder are just for testing purposes. Most of these films are shot at various frame rates, from very low to very high, on both mobile and stationary platforms. These movies are shot with consideration for a variety of other factors such as lighting, occlusion, and even the weather.

3.5VisDrone2018

The large-scale tracking benchmark dataset and visual object identification known as VisDrone2018[121] was made available to researchers late in the previous year. It is made up of photographs and videos obtained by drones. The purpose of this dataset is to improve drone-based visual tasks.

Figure 7 : An picture taken from a drone, complete with a bounding box and item category labels. the VisDrone 2018 dataset provided this image.

The photographs and video clips in the benchmark were taken in 14 cities in China, from the north to the south, spanning the whole country. To be more exact, VisDrone2018 comprises 10,209 pictures and 263 video clips, none of which overlap with one another, all of which include extensive annotations such as bounding boxes, item

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

categories, occlusion ratios, and so on. There are about 2.5 million annotated photos and video frames in this benchmark. The benchmark is the largest dataset of its kind ever released, allowing for broad examination and exploration of drone-based visual analysis techniques. There are a high number of little things, such as dense automobiles, walkers and bicyclists, that will make it difficult to identify specific groups. More than 80 percent of the training photos feature more than 20 items per image, with an average of 54 objects per image in the 6471 training images. As illustrated in Fig. 7, tiny and dense objects are difficult to recognize in this dataset because of the reduced brightness of these pictures than in daytime. The MS COCO metric [122] is used in this dataset.

3.6OpenImage V5

Annotations may be made with Open Image for object bounding boxes, image-level labels, visual linkages and object segmentation masks among other things [20]. A total of 16M bounding boxes for 600 item types are included in Open Images V5, making it the biggest collection containing object position annotations to date. To begin with, the boxes in this dataset were drawn by expert annotators (Google employees) to guarantee correctness and uniformity. Secondly, the photographs in it are very diversified, with an average of 8.3 items per image. Visual association annotations (e.g. "lady playing guitar" and "beer on the table") are also included in this dataset. In all, there are 329 triplets in the database, and there are 391,073 samples to choose from. An object's outline is marked with segmentation masks, which help to define its spatial extent to a greater degree. Finally, 36.5 million image-level labels covering 19,969 classes have been added to the dataset. Figure 8 shows examples of OpenImage V5 dataset

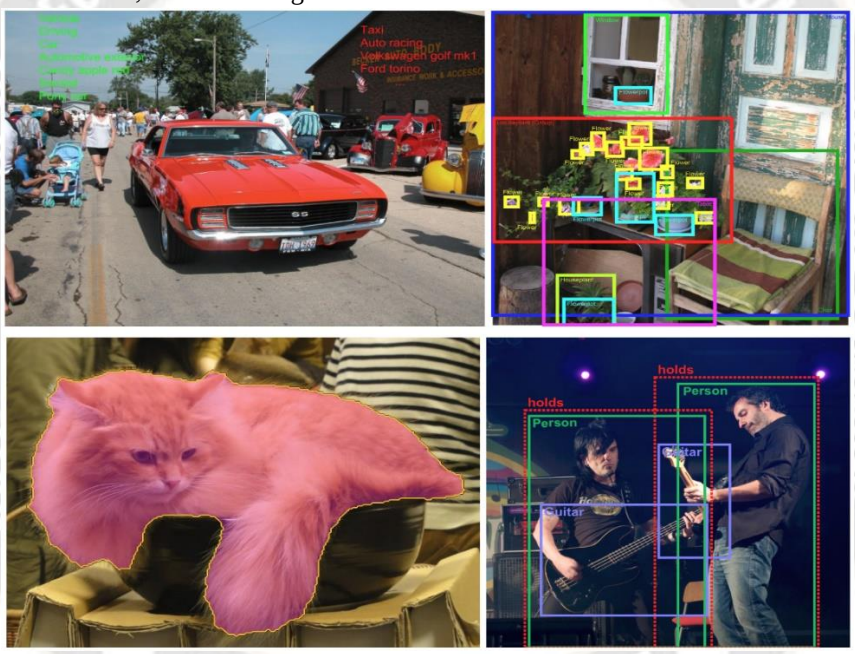

Figure 8: examples of OpenImage V5 dataset.

Detection results of various datasets were shown in table 2 and table 3.

Table 2: Detection results of various datasets over MS COCO

Method	**Backbone network**	**Dataset**	**mAP(%)**
RCNN	VGG-16	train	19.7
Fast RCNN	VGG-16	trainval	21.9
Faster RCNN	VGG-16	trainval35k	22.6
R-FCN	VGG-16	trainval35k	22.6
Mask-RCNN	ResNetXT-101	trainval35k	39.8
YOLO	DarkNet-53	trainval35k	32
YOLO V2	DarkNet-19	trainval35k	33
YOLO V3	DarkNet-19	trainval35k	21.5
SSD	ResNet-101	trainval35k	31.2

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

DSSD	ResNet-101	trainval35k	28
Retina Net	ResNet-101-FPN	trainval35k	34.4
M2Det	ResNet-101	trainval35k	38.8
RefineDet	ResNet-101	trainval35k	41.8
DCN	RetinaNet	trainval35k	32

mAP= mean Average Precision.

Table 3: Detection results of various datasets over PASCAL VOC

Method	**Training data**	**Test data**	**Region**	**Backbone**	**mAP(%)**
RCNN	VOC 07	VOC 07	SS	AlexNet	58.5
Fast RCNN	VOC 07	VOC 07	SS	VGG16	64
Faster RCNN	VOC 07 +VOC 12	VOC 12	SS	VGG16	65
R-FCN	VOC 07 +VOC 12	VOC 12	SS	VGG16	73
Mask-RCNN	VOC 07 +VOC 12	VOC 12		DarkNet-53	34.4
YOLO	VOC 07 +VOC 12	VOC 12	-	VGG16	56.7
YOLO V2	VOC 07 +VOC 12+ MSCOCO	VOC 12	-	VGG16	54.5
YOLO V3	VOC 07 +VOC 12	VOC 12	-	VGG16	65.4
SSD	VOC 07 +VOC 12	VOC 12	RPN	VGG16	76.5
DSSD	VOC 07 +VOC 12+ MSCOCO	VOC 12	FRPN	DarkNet-53	43.4
Retina Net	VOC 07 +VOC 12+ MSCOCO	VOC 12	RPN	ResNet-101	45.4
M2Det	VOC 07 +VOC 12+ MSCOCO	VOC 12	RPN	ResNet-101	65.4
RefineDet	VOC 07 +VOC 12	VOC 12	RPN	ResNet-101	75.4
DCN	VOC 07 +VOC 12	VOC 12	RPN	G-AlexNet	43.4

IV PERFORMANCE METRICS

Object detection and tracking tasks use the following performance metrics:

Mean Average Precision (mAP)

AP, also known as "Average precision," is a typical statistic that is used for the purpose of evaluating the accuracy of object detectors such as Faster R-CNN and SSD. The average precision value is calculated on the average of 0 to 1. With an example, we'll show that it's really rather easy. Prior doing that, though, let's take a brief look back at accuracy, recall, and IoU.

Precision: Precision is a metric for gauging the accuracy of your forecasts. is measured by how many of your hypotheses came true.

Recall: Recall is a metric that assesses how well you remember all of the positives. In our top K forecasts, for example, we can locate 80% of the potential positive situations.

Mathematically, these are their definitions:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2.\frac{precison.recall}{precision + recall}$$

Where,

$$TP = True\ Positive$$
$$TN = True\ Negetive$$
$$FP = False\ Positive$$
$$FN = False\ Negetive$$

Multi object tracking accuracy (MOTA)

It assesses the tracker's and detection's overall accuracy. It takes care of both tracker and detection output. When evaluating a tracker, three mistakes should be taken into account.

Miss Detection (m): When an object is present in the ground truth but is not identified by the detection method, it is called a miss detection.

False-positive (fp): Object that is not existent in the ground truth but is detected as such by the detection method (False detections).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

Mismatch Error (mme): Due to erroneous tracking, an item in the ground truth is incorrectly associated to another object.

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$$

Identity Switches (IDS)

The number of times that two trajectories switch IDs is referred to as the number of identity changes. Each trajectory has its own unique ID.

4.3Multi Object Tracking Precision (MOTP)

Ground-truth and bounding box alignment are predicted by this proportion. To compute MOTP, use this formula.

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$

Where, d_t=The distance between the ground truth object localization and the detection output

c_t= Total number of matches between the detection result and the ground truth

Mostly tracked targets (MT)

It refers to the proportion of ground-truth trajectories that can be explained for at least 80% of their whole existence by a track hypothesis.

Mostly lost targets (ML)

It refers to the percentage of ground-truth trajectories that have at least 20 percent of their lifespan explained by a track hypothesis.

Frames per second (fps):

It is significant in the detection and tracking processes and refers to the quantity of frames processed per second. IDS, MOTA, MT, MOTP, ML, and Speed are utilized in the tracking phase, whereas mAP and Speed are employed in the detecting process.

4.1 Simulation Modules

Adaptive Weighetd Model for Moving Object Detection

This section discuss proposed multiple object tracking with weighted adaptive structural network in real-time surveillance system. The adaptive neural network is involved in the processing of information with desired adjustments to achieve the target in the system. The evaluation is based on an examination of the large complex information processing in the artificial intelligence system. Initially, the mask is created to capture the local target structure. Through the implementation of the weighted fusion, local filtering is performed. The designed model is implemented over the adaptive structure with a regression tracker to classify the multiple objects in the surveillance system. The proposed architecture is given in figure 9.

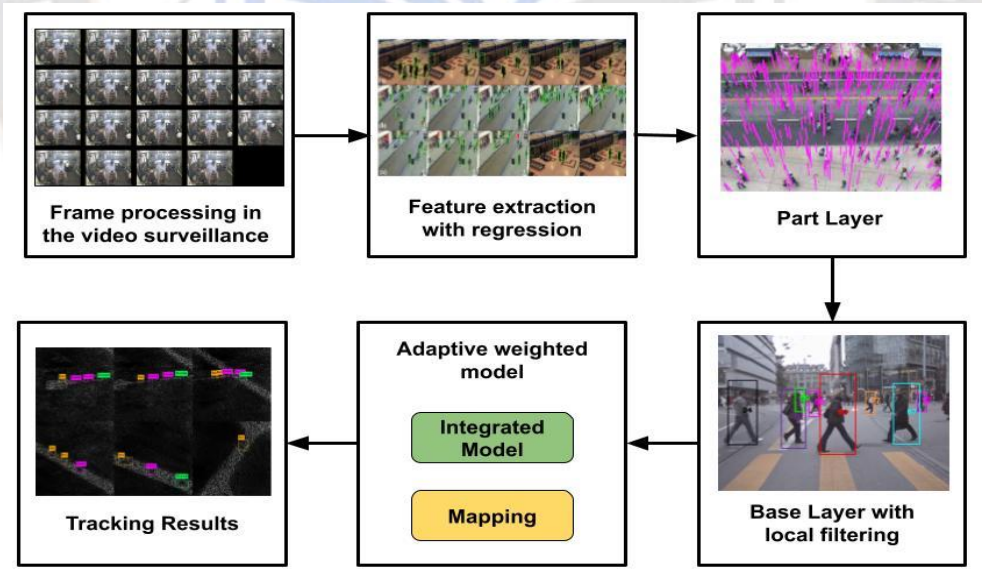

Figure 9: Weighted Adaptive Architecture Model

Filtering

The pre-processing in input images is processed with the random noise for the wide exists of the infrared images those interferes in the detection process. The targets are estimated based on the clutters structure with the detection steps with the employed guided filtering with the detailed structure in the original images for the denoising. Through Binary Mask Generation the binary mask are detected in the occluded facial images in the object mask. The stage – 1 binary mask generated with the G1 generator for the occluded images lc and generate the binary mask lpre_mask. With the apparent estimation the l2 are norms are modified wih the response filter those are represented in the equation

$$\boldsymbol{T} = \sum_{k=1}^{K} \|\boldsymbol{q}_k\|^2 \boldsymbol{M}_k$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

where qk are the responses from single scale loglets, Mk are the filter orientation tensors

In the examination of the operation associated with the denoising and smoothing the images are evaluated for the original and masked images. The image extraction perform the extraction of the skeletons in the image sources those are masked with the optimized image constraints. With the implementation of the angle-based filtering process the boundries of the images are smoothened with the scale-invariant intrinsic vales in the edge pixels with the independent filtering variables. The intrinsic values are computed based on the reference quantity of the spatial movement and expansion of the image edges with the vector unit based on the corner direction with consideration of the curvature degree for the global edge vector to reflect the local image plane smoothness. The edge vector angle subjected to the filtering process represented as the bilateral process based on the consideration of the spatial distance between the edge vectors. The edge vector subjected to the bilateral filtering process through spatial distance computed between edge vector. The edge vector angles are computed based on the intrinsic expansion computed in the corener values. With the discrete values the variable curve are expressed as the follows in equation:

$$\Omega = \{(r_i, \theta_i)\}_{i=1}^{n-1}$$
$$\begin{cases} r_i = l_i / l_{i-1} = \|\overline{v_{\iota+1} v_\iota}\| / \| \overline{v_{\iota-1} v_\iota} \| \\ \theta_i \in (-\pi, \pi] \end{cases}$$

In the above equation, the two-edge vector $\overline{v_{\iota+1} v_\iota}, \overline{v_{\iota-1} v_\iota}$ represents the modulus length of r_i and common vertex is denoted as v_i and the direction of angle of rotation is denoted as θ_i for the adjacent edge vectors. The computation is based on the consideration of the counterclockwise direction. The discrete angle vertex is computed based on the bilaternal filtering scheme computed as in equation

$$\theta_i^* = \begin{cases} \dfrac{\theta_1 + c(v_2, v_1) s(\theta_2, \theta_1) \theta_2}{1 + c(v_2, v_1) s(\theta_2, \theta_1)}, i = 1 \\ \dfrac{c(v_{j-1}, v_i) s(\theta_{i-1}, \theta_i) \theta_{i-1} + \theta_i + c(v_{i+1}, v_j) s(\theta_{i+1}, \theta_j) \theta_{i+1}}{c(v_{i-1}, v_i) s(\theta_{i-1}, \theta_i) + 1 + c(v_{i+1}, v_j) s(\theta_{i+1}, \theta_i)}, i = 2,3, \\ \dfrac{c(v_{n-2}, v_{n-1}) s(\theta_{n-2}, \theta_{n-1}) \theta_{n-2} + \theta_{n-1}}{1 + c(v_{n-2}, v_{n-1}) s(\theta_{n-2}, \theta_{n-1})}, i = n-1 \end{cases}$$

where $\theta * i$ representes the vertex vi cirectional corner with the curve for the bilaternal filtering process is denoted as $c(v_j, v_i) = \exp\left(-\frac{1}{2}\left(\frac{d(v_j, v_i)}{\sigma_d}\right)^2\right)$ with the Euclidean distance between the variables are computed based on the vertices and (v_j, v_i). The vertices corner is denoted as the $s(\theta_j, \theta_i) = \exp\left(-\frac{1}{2}\left(\frac{\delta(\theta_j, \theta_i)}{\sigma_r}\right)^2\right)$ and $\delta(\theta_j, \theta_i) = |\theta_j - \theta_i|$. The weighted factor size computed based on the σ_r with the specific situation and the size distance is measured with the σ_r for the higher distance size. The larger the value of σ_r the higher the value of the angle different effect. With the local noise estimation the thoughts are computed based on the residual estimation using the estimation theory for the available same signal those are weighted together with use of weights corresponding to the variance noise. The local certainty in the background is computed based on the value c is computed using the equation

$$c = \frac{\gamma(E_{\text{tot}} - E_{\text{res}})}{E_{\text{res}} + \gamma E_{\text{tot}}}$$

where Etot represents the filter set for the total signal, residue of the energy is denoted as Eres, the deviation from the actual filter is represented as the ideal shape of filter. The estimated certainty is ranges from 0 to 1 those are adjusted with the value γ.

Adaptive weighting fusion strategy

With baseline fusion model the features are integrated with the spatial features and spectral density with the HIS features through the feature addition and concatenation of the channel. With the effective method the operation is evaluated based on the consideration fo the feature extractors Xspc and Xspa. The features Xspc and Xspa involved in the estimation of the aligned weights those are directly fused. Instead the baseline fusion based tensor alignment model is developed based on the input features Xspc and Xspa for the output features Gspc and Gspa respectively. The feature dimensions are computed in the intermediate layer based on the computed input and output layer. Specifically, the features of the convolutional operation (Conv2D) the tensor features are aligned based on the kernel size of 1 × 1. The each batch in the convolutional layer comprises of the rectified linear unit those are normalized with the Gspc and Gspa for the training process. Finally, the fusion is performed for features Fss with the formulated Gspc and Gspa based on the channel dimensions. In figure two basic fusion model for the developed architecture model is presented as follows:

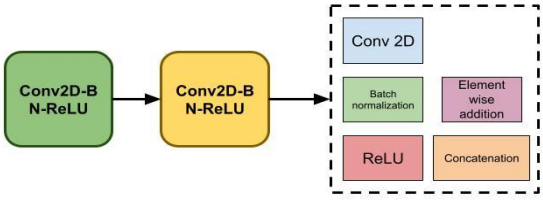

Figure 10. Elementwise addition and feature extraction

With adaptive feature weights strategy the Multiview fusion is adopted for the 3D object mechanism for the image monitoring and interpretation. In this scenario, the fundamental machine learning schemes are adopted based on the consideration of the spectral and spatial features with the hyperspectral interpretation in the point of 3D

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

perspectives. Through adaptive weighted features the multibranch neural network are computed for the fully connected ReLU block to map the distributed feature representation with the carse fusion in the sample space labels. The weights are computed based on each pixel features for the each branch of network with the feature map in the each element comprises of the different weighted matrix with computation of the parameter variation in the FC operator layer. Figure 10 shows elementwise addition and feature extraction

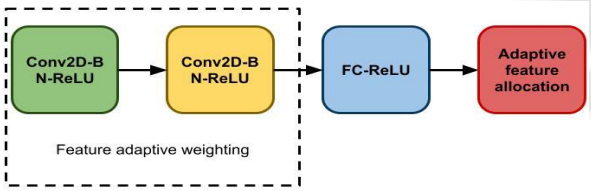

Figure 11: Adaptive weighting fusion models with (a) element-wise addition and (b) feature concatenation.

Consider the unknown quantity of the different N sensors with the quantity Y for every observation value in each sensor is represented as Yj (j = 1, 2, . . ., N) with the jth sensor observed value as I equation

$\hat{Y} = \sum_{j=1}^{N} W_j Y_j$

where Wj is the satisfied weights in the network as presented in equation

$\sum_{j=1}^{N} W_j = 1$

The estimated variance of the network is presented as in equation

$\sigma^2 = \sum_{j=1}^{N} W_j^2 \sigma_j^2$

Where, σ_j^2 denoted the different sensor variance the variance at time is represented as w_j^2 with the minimal value based on the computed auxiliary function as presented in equation

$f(w_1, w_2, \dots, w_N, \lambda) = \sum_{j=1}^{N} w_j^2 \sigma_j^2 + \lambda\left(\sum_{j=1}^{N} w_j - 1\right)$

In above equation $w_1, w_2, \dots, w_N, \lambda$ the partial derivatives are computed based on the constraints those are represented as in equation

$$\begin{cases} \frac{\partial f}{\partial w_1} = 2w_1\sigma_1^2 + \lambda = 0, \\ \cdots \\ \frac{\partial f}{\partial w_N} = 2w_N\sigma_N^2 + \lambda = 0, \\ \sum_{j=1}^{N} w_j - 1 = 0. \end{cases}$$

After organizing, the following is obtained in equation

$$\begin{cases} w_1 + w_2 + \cdots + w_N = 1, \\ w_j = \frac{u}{\sigma_j^2},\ j = 1,2,3, \dots; \end{cases}$$

Then, the computed extreme value for the $\sigma_{\min}^2$

$\mu = \frac{\lambda}{2}\ \sigma_{\min}^2 = \mu$

Substituting the Substituting the simultaneous equations a

$w_j = \frac{1}{\sigma_j^2 \sum_{i=1}^{N} 1/\sigma_i^2},\ j = 1,2,3, \dots, N.$

In the above equation the estimation is based on the computation of the true value $\bar{Y}_j(k)$ for the measured data in the single data measurement. Through the measured objects the relative constant is evaluated for the certain time period to increases the accuracy with the computed kth estimated true value $\bar{Y}_j(k)$. The average value for the estimated data measurement for time k is denoted as

$\bar{Y}_j(k) = \frac{1}{k}\sum_{i=1}^{k} Y_j(i),\ j = 1,2,3, \dots, N$

The estimation is based on the computation as follows:

$\hat{\bar{Y}} = \sum_{j=1}^{N} w_j \bar{Y}_j(k)$

With the unbiased estimation $\bar{Y}_j$ with the estimated in the unbiased state is denoted as Y with the total variance estimation time as in equation

$$\bar{\sigma}^2 = E\left[Y - \hat{Y}^2\right] = E\left[\sum_{j=1}^{N} w_j^2 (Y - Y_2(k))\right] = \frac{1}{k}\sum_{j=1}^{N} w_j \sigma_j^2.$$

+e same as the solution process of σ2 min, the conditional extreme values ar

$\bar{\sigma}_{\min}^2 = \frac{1}{k(\sum_{i=1}^{N} 1/\sigma_i^2)} = \frac{\sigma_{\min}^2}{k}$

$\min_{\mathbf{v}^{(i)}, \mathbf{Z}^{(i)}} \sum_{i=1}^{m} \|\mathbf{V}^{(i)} - \mathbf{V}^{(i)}\mathbf{Z}^{(i)}\|_F^2$

The intra-structure similarity is computed based on the subspace structure with the property of the self-representation in the latent dimension of k × ni for the constructed graph denoted as $\mathbf{V}^{(i)}$. Figure 11 shows adaptive weighting fusion models

4.2 Experimental Analysis

In final process, to increases the overall performance of the multi-object detection and tracking this phase uses the adaptive weighted model for the object detection and classification. The Proposed model involved in the tracking as well as classification of the objects in the video frame. With part and base layer, the tracking and classification of the objects are performed. The proposed model comprises of the processing of the video sequences involved in the estimation of the frames in the video sequences. Upon the estimation of the frames in the video sequences the objects in the frame are evaluated based in the masking of the features in the video sequences. With the weighted adaptive model the sequences are processed and computed for the tracking of the objects in the video frame. Finally, the classification is performed in the developed adaptive weighted model for the object detection and tracking of the features in the frame sequences. In figure 12 presented the objects classified in the frame is presented.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

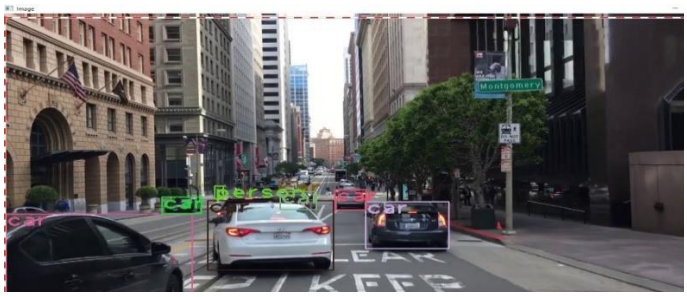

Figure 12: MOT tracker tracks 1 objects

The proposed MOT technique comprises the YOLO-based ANN for tracking as well as classification of objects in video frames. The proposed technique comprises the part layer for the processing of the video frames. The effectiveness of the suggested model is assessed and looked at while taking the various parameters into account. The suggested model is assessed taking into account many factors as accuracy, precision, recall, and MAP. In figure 13 the objects tracked in the frame are presented in the single frame.

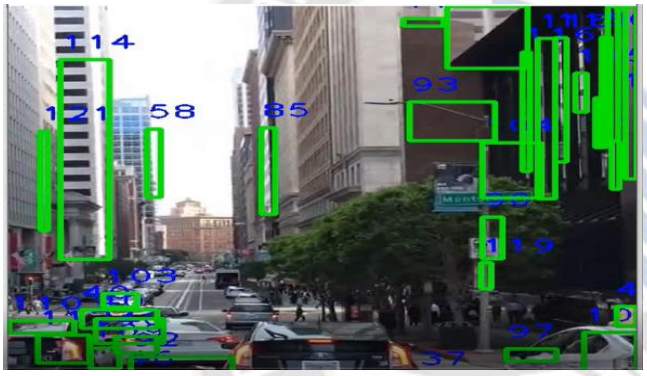

Figure 13: MOT tracker for object computation in the frame

With the estimation of the frame object in the video sequence the masking is performed for the extraction and processing of the objects in the video frame. The masked region in the video frames are computed or the extraction of the objects in the video sequences. The figure 14 provides the masked region in the video frames is presented.

Figure 14: Masked objects in the frame

Through the extraction of the MOT in the video frame the objects in the system were classified for the detection and classification of the objects in the video frame sequences. In the figure 14 the objects in the frame 1 are processed and computed are presented. Based on the estimation of the objects in the frame the making is performed in the video process and involved in the computation of the objects tracking in the video sequences. In figure 15 the estimated MOT features in the video sequences for the object tracking and classification is presented.

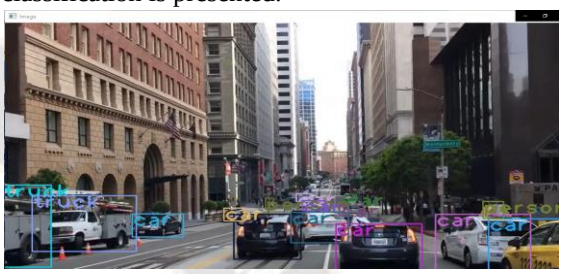

Figure 15: MOT Detected in a frame

Through the estimation of the image features in the video frames the objects are computed for tracking and classification with the YOLO architecture in the video sequences.

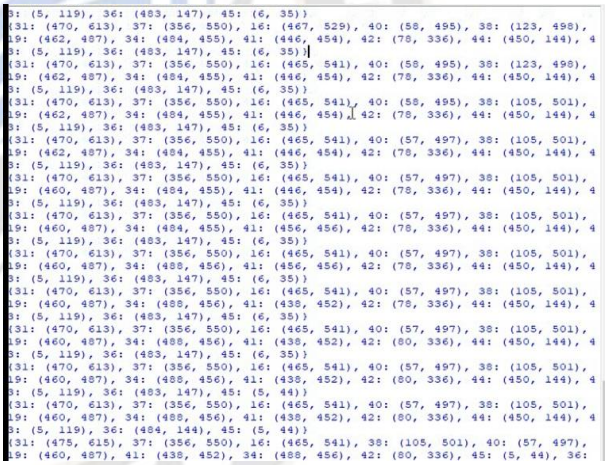

Figure 16: Time instances in the frame for the MOT

The time stamp computed for the every frame in the video are computed for the detected masked objects and presented in the figure 16. In table 4 presented overall summary of object tracked with consideration of the different parameters are presented.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

Table 4: Comparison of Parameters

MOT	Accuracy (%)	Precision (%)	Recall (%)	MAP
MOT 1	0.8598912071204204	0.862553477105986	0.8189770844301077	0.9174968119988154
MOT 2	0.8911950064212537	0.8610385397460701	0.8377800741441823	0.9425294867179055
MOT 3	0.9147656855347406	0.9106508484035426	0.8198797516885679	0 .9029634971800119
MOT 4	0.9318344982184051	0.9339121915452149	0.8358461051615493	0.8900679007347325
MOT 5	0.9114070889686633	0.915636184289186	0.8121277572916239	0.9108590043419386
MOT 6	0.9026920814021585	0.8929694540570153	0.8469507353489647	0.9270644190252084
MOT 7	0.9233247034898774	0.9295061381066072	0.8210888017487482	0.9206433027829412
MOT 8	0.9106182699663652	0.911341062255566	0.8102017051596151	0 .9290090620719028
MOT 9	0.9231911412861663	0.9277057017285966	0.8254246642291113	0.926756308348119
MOT 10	0.9468262385505021	0.9366674939672878	0.8254734315885994	0.9295708323235157

In table 14 shows the proposed analysis of multiple object tracking with the weighted adaptive structural network in a real-time surveillance system. Here the parameters compared are accuracy, precision, recall, and MAP. The comparative analysis is carried out for the MOT benchmark dataset for both proposed and existing techniques. here the proposed technique obtained accuracy up to 94%, precision up to 93%, recall up to 82%, and MAP up to 92%. The other features obtained by the proposed technique are the correlation of 95%, the entropy of 5.49%, homogeneity of 5.49%, the threshold of 23.041%, the contrast of 57%, and energy of 46%. The existing technique obtained accuracy up to 82%, precision up to 72%, recall up to 63%, and MAP up to 91%. From this above analysis, the proposed technique obtained optimal results in multiple objects tracking systems. Through the object tracking and detection, the features in the video frames are evaluated with the average computation of the features in the frame. The image frame features considered are correlation, entropy, Homogeneity, threshold, contrast, and energy. In table 5 the features and characteristics estimated and derived for the proposed model are presented.

Table 5: Estimation of Features

Features	Values
Correlation	0.9567387170965557
Entropy	5.4986739701865925
Homogeneity	5.4386739701865925
Threshold	23.041746697633947
Contrast	0.5783156764509081
Energy	0.46194221566790106

The suggested model gives the computed frame's feature vector value for the video sequences. The component layer in the tracking method is used in the proposed model to estimate the correlation value of 0.95. The values are obtained as 5.49 and 5.43, respectively, for similarity and homogeneity. The technique's computed threshold, contrast, and energy level are displayed as 23.0417, 0.578, and 0.462, respectively, for the suggested model. The suggested technique's accuracy estimation is contrasted with that of the current model, which is based on the masking of characteristics in the video frame.

V CONCLUSION AND FUTURE TRENDS

As more powerful computing equipment has been available, the object identification technology that is based on deep learning has seen rapid advancement. There is an increasing need for high-precision real-time systems to be deployed on

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

applications that require a greater degree of accuracy. Researchers have developed a number of approaches because the ultimate goal of this task is to achieve high accuracy and efficiency detectors. These strategies involve starting from scratch, merging one-stage and two-stage detectors to achieve good results, creating new architecture, enhancing processing performance, utilizing good representations, anchor-free methods, and handling complex scene challenges (tiny objects, occluded objects). The development of increasingly sophisticated object detectors has resulted in an increase in the number of applications for object detection across a variety of industries, including defense, transportation, medicine, and the life sciences, because the ultimate goal of this task is to achieve high accuracy and efficiency detectors. The detection domain has a number of branches as well. Although this domain has recently experienced success, there is still much room for development.

- Using one-stage and two-stage detectors together: On the one hand, two-stage detectors make use of a tightly tailing approach that is both time demanding and inefficient in order to collect the maximum number of reference boxes feasible. Researchers need to decrease the amount of repetition they do while maintaining a high level of accuracy in order to solve this challenge. These advantages make them an attractive option. Even while it is happening rather quickly, the decline in accuracy nevertheless presents a challenge for tasks that need great precision. Combining the benefits of detectors with one stage with detectors with two stages remains a difficult problem.
- Video object detection: Video defocus, Motion blur, intense target movements, motion target ambiguity, occlusion, truncation and small targets, and a number of other problems combine to make it difficult to carry out this task successfully in real-world and remote sensing contexts. The examination of altering goals and increasingly complicated source material, such as video, is going to be one of the primaries focuses of research in the years to come.
- Effective post-processing techniques: When using a methodology with three phases (for one stage detectors) or four stages, post-processing occurs early in the process for producing the final findings (for two stage detectors). Only the prediction result with the highest accuracy for a single item may be provided to the metric program to be used in the accuracy score calculation in the majority of detection metrics. The accuracy of the measurement may suffer as a result of post-processing operations such as NMS and its updates, which may exclude things that are conveniently available yet have a high confidence level in their classification. Utilizing post-processing techniques that are both more effective and precise is an alternative route for the object detection domain.
- Weakly supervised object identification techniques: It is far more effective and simpler to use a significant portion of images that have just been labelled with the object class and not the object bounding box in order to train the network. Weakly supervised object detection (WSOD) is a technique that uses only a small number of "completely annotated" images to identify a large number of "partially annotated" images (supervision). As a direct result of this, the development of WSOD approaches is an important topic that needs further research.
- Multi-domain object detection: When applied to the supplied dataset, domain-specific detectors always generate good detection performance. This would result in a universal detector that is capable of operating in a number of different image domains. Transferring a domain is a challenging undertaking that calls for more investigation.
- 3D object detection: Research into three-dimensional object identification is becoming more popular as a result of the development of 3D sensors and the proliferation of applications that use 3D knowledge. LiDAR point clouds, as opposed to detection based on two-dimensional images, provide accurate information on depth, which may be used to characterize the shapes and locate objects they take. The use of LiDAR enables accurate three-dimensional object localization. Object recognition algorithms that are based on LiDAR often perform better than their 2D counterparts.
- Salient object detection (SOD): The purpose of salient object detection, often known as SOD, is to draw attention to significant aspects in photographic images. The process of identifying and localizing objects of interest within a continuous scene is referred to as "object identification" in the medium of video. SOD is both driven by and used in a wide variety of object-level applications across a number of different industries. Accurate object identification in motion pictures may be facilitated by the provision of key areas of interest on significant objects in each frame. As a

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

consequence of this, the detection of highlighted targets is an essential part of the preparatory phase for challenging detection tasks and high-level identification tasks.

- Unsupervised object detection: The supervised method training approach is time demanding and inefficient, and these methods need a dataset that has been well annotated. When working with massive datasets, it is prohibitively expensive, time-consuming, and impractical to annotate a bounding box for every object. A potential trend in unsupervised object identification is the development of automated annotation technologies to save up human annotation time. For intelligent detection missions, unsupervised object detection is a potential research path.
- Multi-task learning: The aggregation of multi-level backbone network characteristics is a critical strategy that may be utilized to increase detection performance. Furthermore, due to the increased availability of information, performing multiple computer vision tasks simultaneously, such as semantic segmentation, object identification, edge detection, highlight detection and instance segmentation has the potential to significantly improve the performance of individual tasks. Although multitask learning is an effective method for aggregating many tasks in a network, it poses significant hurdles for academics in terms of maintaining processing speed while enhancing accuracy.
- Multi-source information assistance: As a result of the proliferation of social media and the further development of multi-source information and big data technologies is becoming easier to acquire. A great number of social networking sites may provide not only images but also textual descriptions of the individuals in question, which may be of assistance in the process of finding them. Fusing information from several sources is becoming an increasingly important topic of research as a result of the proliferation of various technologies.
- Building a terminal object detection system: By bringing artificial intelligence from the cloud to the terminal, individuals will be able to deal with large amounts of data and solve issues more effectively and quickly. Terminal detectors have evolved into more efficient and dependable devices with a wide range of application situations since the introduction of lightweight networks. Real-time applications will be available thanks to the chip detecting network based on FPGA.
- Medical imaging and diagnosis: The Food and Drug Administration (FDA) strongly endorses the use of medical devices that are powered by AI. In April of 2018, the Food and Drug Administration (FDA) gave its first authorization to a software called IDx-DR that identifies diabetic retinopathy with more than 87.4 percent accuracy using artificial intelligence. Clients may find their mobile phones to be a very helpful asset in the diagnosis process of their families if photo recognition technology is integrated into these devices. This strategy comes with a significant number of prerequisites and challenges to overcome.
- Advanced medical biometrics: Deep neural networks were used by academics to begin examining and evaluating previously unidentified risk factors. Physicians may be able to forecast a patient's chance of developing heart disease by analyzing speech patterns and retinal images using neural networks. In the not-too-distant future, passive monitoring will use medical biometrics.
- Airborne remote sensing and real-time detection: Accurate processing of remote sensing pictures is required in both military and agricultural applications. These industries will see tremendous growth thanks to automated hardware and integrated software. Real-time high-altitude detection is achieved by loading a deep learning-based object identification algorithm onto a SoC (System on Chip).
- Deep learning-based detector: Deep learning-based systems often need a large quantity of data during the training process; however, the Generative Adversarial Network provides an effective framework for fabricating bogus images. The object detector becomes more robust and has greater generalization power when real-world scenes are combined with GAN-generated simulated data. How much you require and how much it is capable of creating by mixing real-world scenes with GAN-generated simulated data.
- Automatic material classification: Simulation technologies are used in a variety of fields, including aerospace technology, military training, disaster response, and municipal planning among others, in order to properly mimic and depict genuine environments. The simulation system's operational interface must convey an atmosphere that is true to the scenario being simulated.

Landscape reconstruction relies heavily on the creation of three-dimensional terrain photographs, a time-consuming and labor-intensive process that relies heavily on-site altitude, viewing angle, day and night illumination, and weather. Reconstructing landscape and topography necessitate this technique, which involves creating a terrain model from pictures provided by visual systems. In spite of the fact that it is an invention of the mind, the material is authentic, and the physical setting is laid out in a logical manner on the Internet.

REFERENCES

[1] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: state of the art," *Appl. Intell.*, vol. 51, no. 9, pp. 6400–6429, Sep. 2021, doi: 10.1007/s10489-021-02293-7.

[2] S. K. Pal, D. Bhoumik, and D. Bhunia Chakraborty, "Granulated deep learning and Z-numbers in motion detection and object recognition," *Neural Comput. Appl.*, vol. 32, no. 21, pp. 16533–16548, Nov. 2020, doi: 10.1007/s00521-019-04200-1.

[3] L. Jiao *et al.*, "A Survey of Deep Learning-Based Object Detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019, doi: 10.1109/ACCESS.2019.2939201.

[4] Y. Xiao *et al.*, "A review of object detection based on deep learning," *Multimed. Tools Appl.*, vol. 79, no. 33–34, pp. 23729–23791, Sep. 2020, doi: 10.1007/s11042-020-08976-6.

[5] C. Hui, B. Xingcan, and L. Mingqi, "Research on Image Edge Detection Method Based on Multi-sensor Data Fusion," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Jun. 2020, pp. 789–792. doi: 10.1109/ICAICA50127.2020.9182548.

[6] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 6, p. 420, Nov. 2021, doi: 10.1007/s42979-021-00815-1.

[7] W. Fuhl and E. Kasneci, "Multi Layer Neural Networks as Replacement for Pooling Operations," Jun. 2020, doi: https://doi.org/10.48550/arXiv.2006.06969.

[8] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, Jun. 2019, doi: 10.1007/s10462-018-09679-z.

[9] B. Playe and V. Stoven, "Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity," *J. Cheminform.*, vol. 12, no. 1, p. 11, Dec. 2020, doi: 10.1186/s13321-020-0413-0.

[10] M. Naderipour, M. Zarandi, and S. Bastani, "A Multi-layer General Type-2 Fuzzy Community Detection Model in Large-scale Social Networks," *IEEE Trans. Fuzzy Syst.*, pp. 1–1, 2022, doi: 10.1109/TFUZZ.2022.3153745.

[11] Q. Wang, L. Zhang, Y. Li, and K. Kpalma, "Overview of deep-learning based methods for salient object detection in videos," *Pattern Recognit.*, vol. 104, p. 107340, Aug. 2020, doi: 10.1016/j.patcog.2020.107340.

[12] G. Huang, "Attention Guided Multi-Scale Regression for Scene Text Detection," in *2021 2nd International Conference on Computing and Data Science (CDS)*, Jan. 2021, pp. 498–502. doi: 10.1109/CDS52072.2021.00092.

[13] R. Gupta, D. Saini, and S. Mishra, "Posture detection using Deep Learning for Time Series Data," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Aug. 2020, pp. 740–744. doi: 10.1109/ICSSIT48917.2020.9214223.

[14] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Pedestrian Detection: Domain Generalization, CNNs, Transformers and Beyond," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.03176

[15] J. Kim and L. Wei, "Performance Analysis of Machine Learning-based Face Detection Algorithms in Face Image Transmission over AWGN and Fading Channels," in *2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Aug. 2021, pp. 1–5. doi: 10.1109/BMSB53066.2021.9547181.

[16] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild," *IEEE Trans. Multimed.*, vol. 22, no. 2, pp. 380–393, Feb. 2020, doi: 10.1109/TMM.2019.2929005.

[17] I. Cvisic, I. Markovic, and I. Petrovic, "Recalibrating the KITTI Dataset Camera Setup for Improved Odometry Accuracy," in *2021 European Conference on Mobile Robots (ECMR)*, Aug. 2021, pp. 1–6. doi: 10.1109/ECMR50962.2021.9568821.

[18] J. Kim, J. Bae, G. Park, D. Zhang, and Y. M. Kim, "N-ImageNet: Towards Robust, Fine-Grained Object Recognition with Event Cameras," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.01041

[19] S. Srivastava, A. V. Divekar, C. Anilkumar, I. Naik, V. Kulkarni, and V. Pattabiraman, "Comparative analysis of deep learning image detection algorithms," *J. Big Data*, vol. 8, no. 1, p. 66, Dec. 2021, doi: 10.1186/s40537-021-00434-w.

[20] Y. Kim, J. M. Kim, Z. Akata, and J. Lee, "Large Loss Matters in Weakly Supervised Multi-Label Classification," Jun. 2022, doi: 10.1103/PhysRevE.105.054214.

[21] I. Krylov, S. Nosov, and V. Sovrasov, "Open Images V5 Text Annotation and Yet Another Mask Text Spotter," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.12326

[22] O. C. Koyun, R. K. Keser, İ. B. Akkaya, and B. U. Töreyin, "Focus-and-Detect: A Small Object Detection Framework for Aerial Images," Mar. 2022, doi: 10.1016/j.image.2022.116675.

[23] L. Qinghe, C. Weihao, and Q. Hao, "Research on small target detection based on improved Faster R-CNN depth

network method," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Sep. 2020, pp. 833–838. doi: 10.1109/IFEEA51475.2020.00175.

[24] M. Mahendru and S. K. Dubey, "Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2021, pp. 734–740. doi: 10.1109/Confluence51648.2021.9377064.

[25] Q. Shuai and X. Wu, "Object detection system based on SSD algorithm," in *2020 International Conference on Culture-oriented Science & Technology (ICCST)*, Oct. 2020, pp. 141–144. doi: 10.1109/ICCST50977.2020.00033.

[26] X. Long, Z. Zheng, Y. Chi, and R. Liu, "A Mixed Two-stage Object Detector for Image Processing of Power System Applications," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*, Oct. 2020, pp. 1352–1355. doi: 10.1109/ICCT50939.2020.9295843.

[27] A. N. Amudhan, S. R. Vrajesh, A. P. Sudheer, and A. Lijiya, "RFSOD: a lightweight single-stage detector for real-time embedded applications to detect small-size objects," *J. Real-Time Image Process.*, vol. 19, no. 1, pp. 133–146, Feb. 2022, doi: 10.1007/s11554-021-01170-3.

[28] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[29] H. Li, X. Yue, Z. Wang, W. Wang, H. Tomiyama, and L. Meng, "A survey of Convolutional Neural Networks — From software to hardware and the applications in measurement," *Meas. Sensors*, vol. 18, p. 100080, Dec. 2021, doi: 10.1016/j.measen.2021.100080.

[30] V. Mandal and Y. Adu-Gyamfi, "Object Detection and Tracking Algorithms for Vehicle Counting: A Comparative Analysis," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.16198

[31] A. Benali Amjoud and M. Amrouch, "Convolutional Neural Networks Backbones for Object Detection," 2020, pp. 282–289. doi: 10.1007/978-3-030-51935-3_30.

[32] X. Zhao, J. Chen, M. Liu, K. Ye, and L. Shen, "Multi-scale Attention-Based Feature Pyramid Networks for Object Detection," 2021, pp. 405–417. doi: 10.1007/978-3-030-87355-4_34.

[33] Y. Lu, L. Zhang, and W. Xie, "YOLO-compact: An Efficient YOLO Network for Single Category Real-time Object Detection," in *2020 Chinese Control And Decision Conference (CCDC)*, Aug. 2020, pp. 1931–1936. doi: 10.1109/CCDC49329.2020.9164580.

[34] C. Zhuang, "DetNAS: Design Object Detection Network via One-Shot Neural Architecture Search," in *2021 2nd Asia Symposium on Signal Processing (ASSP)*, Nov. 2021, pp. 28–37. doi: 10.1109/ASSP54407.2021.00013.

[35] A. Krueangsai and S. Supratid, "Effects of Shortcut-Level Amount in Lightweight ResNet of ResNet on Object Recognition with Distinct Number of Categories," in *2022 International Electrical Engineering Congress (iEECON)*, Mar. 2022, pp. 1–4. doi: 10.1109/iEECON53204.2022.9741665.

[36] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net Structures for Speaker Verification," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2021, pp. 301–307. doi: 10.1109/SLT48900.2021.9383531.

[37] A. Jain, A. Shafi, Q. Anthony, P. Kousha, H. Subramoni, and D. K. Panda, "Hy-Fi: Hybrid Five-Dimensional Parallel DNN Training on High-Performance GPU Clusters," 2022, pp. 109–130. doi: 10.1007/978-3-031-07312-0_6.

[38] Z. Wentao, G. Lan, and Z. Zhisong, "Garbage Classification and Recognition Based on SqueezeNet," in *2020 3rd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM)*, Dec. 2020, pp. 122–125. doi: 10.1109/WCMEIM52463.2020.00032.

[39] B. Koonce, "MobileNet v1," in *Convolutional Neural Networks with Swift for Tensorflow*, Berkeley, CA: Apress, 2021, pp. 87–97. doi: 10.1007/978-1-4842-6168-2_8.

[40] S. Ghosh, M. J. Mondal, S. Sen, S. Chatterjee, N. Kar Roy, and S. Patnaik, "A novel approach to detect and classify fruits using ShuffleNet V2," in *2020 IEEE Applied Signal Processing Conference (ASPCON)*, Oct. 2020, pp. 163–167. doi: 10.1109/ASPCON49795.2020.9276669.

[41] X. Wu, R. Liu, H. Yang, and Z. Chen, "An Xception Based Convolutional Neural Network for Scene Image Classification with Transfer Learning," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, Dec. 2020, pp. 262–267. doi: 10.1109/ITCA52113.2020.00063.

[42] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A Real-Time Object Detection System on Mobile Devices," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.06882

[43] X. Xiao and X. Tian, "Research on Reference Target Detection of Deep Learning Framework Faster-RCNN," in *2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA)*, Sep. 2021, pp. 41–44. doi: 10.1109/ICDSBA53075.2021.00017.

[44] T. D. D and K. V, "Deep Learning based Object Detection using Mask RCNN," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Jul. 2021, pp. 1684–1690. doi: 10.1109/ICCES51350.2021.9489152.

[45] M. F. Haque, H.-Y. Lim, and D.-S. Kang, "Object Detection Based on VGG with ResNet Network," in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, Jan. 2019, pp. 1–3. doi: 10.23919/ELINFOCOM.2019.8706476.

[46] R. Ashraf *et al.*, "Deep Convolution Neural Network for Big Data Medical Image Classification," *IEEE Access*, vol. 8, pp. 105659–105670, 2020, doi: 10.1109/ACCESS.2020.2998808.

[47] A. M. Hafiz and G. M. Bhat, "A Survey on Instance Segmentation: State of the art," Jun. 2020, doi: 10.1007/s13735-020-00195-x.

[48] L. Jiang, J. Chen, H. Todo, Z. Tang, S. Liu, and Y. Li, "Application of a Fast RCNN Based on Upper and Lower Layers in Face Recognition," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–12, Sep. 2021, doi: 10.1155/2021/9945934.

[49] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.09017

[50] S. Li, L. Wang, J. Li, and Y. Yao, "Image Classification Algorithm Based on Improved AlexNet," *J. Phys. Conf. Ser.*, vol. 1813, no. 1, p. 012051, Feb. 2021, doi: 10.1088/1742-6596/1813/1/012051.

[51] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in timm," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.00476

[52] Z. Yu, Y. Dong, J. Cheng, M. Sun, and F. Su, "Research on Face Recognition Classification Based on Improved GoogleNet," *Secur. Commun. Networks*, vol. 2022, pp. 1–6, Jan. 2022, doi: 10.1155/2022/7192306.

[53] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," May 2016, [Online]. Available: http://arxiv.org/abs/1605.06409

[54] J. Tang, T. Lu, and J. Wei, "Multi-Scale Approach for Document Detection Based on the Cascade mask RCNN," in *2021 4th International Conference on Artificial Intelligence and Pattern Recognition*, Sep. 2021, pp. 511–517. doi: 10.1145/3488933.3489039.

[55] N. Xiang, C. Pan, and X. Li, "An Object Detection Algorithm Combining FPN Structure With DETR," in *2021 4th International Conference on Control and Computer Vision*, Aug. 2021, pp. 57–63. doi: 10.1145/3484274.3484284.

[56] Z. Qiu, T. Yao, C.-W. Ngo, and T. Mei, "Optimization Planning for 3D ConvNets," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.04021

[57] X. Huang *et al.*, "PP-YOLOv2: A Practical Object Detector," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.10419

[58] M. Alkhaleefah, N. B. Tatini, H.-T. Lee, T.-H. Tan, S.-C. Ma, and Y.-L. Chang, "YOLOv3-mobile for Real-time Pedestrian Detection on Embedded GPU," in *2021 the 5th International Conference on Graphics and Signal Processing*, Jun. 2021, pp. 27–31. doi: 10.1145/3474906.3474915.

[59] R. Araki, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "MT-DSSD: multi-task deconvolutional single shot detector for object detection, segmentation, and grasping detection," *Adv. Robot.*, vol. 36, no. 8, pp. 373–387, Apr. 2022, doi: 10.1080/01691864.2022.2043183.

[60] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, "SE-SSD: Self-Ensembling Single-Stage Object Detector From Point Cloud," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.09804

[61] J. Royo-Miquel, S. Tolu, F. E. T. Schöller, and R. Galeazzi, "RetinaNet Object Detector based on Analog-to-Spiking Neural Network Conversion," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.05624

[62] Q. Zhao *et al.*, "M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 9259–9266, Jul. 2019, doi: 10.1609/aaai.v33i01.33019259.

[63] R. Wang *et al.*, "DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems," Aug. 2020, doi: 10.1145/3442381.3450078.

[64] M. Zhu, G. Hu, S. Li, S. Liu, and S. Wang, "An Effective Ship Detection Method Based on RefineDet in SAR Images," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, May 2021, pp. 377–380. doi: 10.1109/CISCE52179.2021.9445958.

[65] S. Zhang, E. Nezhadarya, H. Fashandi, J. Liu, D. Graham, and M. Shah, "Stochastic Whitening Batch Normalization," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.04413

[66] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 658–666. doi: 10.1109/CVPR.2019.00075.

[67] H. Wang, C. Hao, and B. Li, "A water area illegal intrusion detection algorithm based on yolov3 algorithm modification with higher detection accuracy," in *2021 International Conference on Computer, Blockchain and Financial Development (CBFD)*, Apr. 2021, pp. 54–59. doi: 10.1109/CBFD52659.2021.00018.

[68] A. Bagaskara and M. Suryanegara, "Evaluation of VGG-16 and VGG-19 Deep Learning Architecture for Classifying Dementia People," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, Sep. 2021, pp. 1–4. doi: 10.1109/IC2IE53219.2021.9649132.

[69] D. R. D. S, B. S. Negara, S. Sanjaya, and E. Satria, "COVID-19 Classification for Chest X-Ray Images using Deep Learning and Resnet-101," in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, Jul. 2021, pp. 1–4. doi: 10.1109/ICOTEN52080.2021.9493431.

[70] J. Sales, J. Marcato Junior, H. Siqueira, M. De Souza, E. Matsubara, and W. N. Goncalves, "Retinanet Deep Learning-Based Approach to Detect Termite Mounds in Eucalyptus Forests," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, pp. 586–589. doi: 10.1109/IGARSS47720.2021.9555177.

[71] T. Zhang and L. Li, "An Improved Object Detection Algorithm Based on M2Det," in *2020 IEEE International Conference on Artificial Intelligence and Computer*

Applications (ICAICA), Jun. 2020, pp. 582–585. doi: 10.1109/ICAICA50127.2020.9181938.

[72] S.-H. Lee and H.-C. Chen, "U-SSD: Improved SSD Based on U-Net Architecture for End-to-End Table Detection in Document Images," *Appl. Sci.*, vol. 11, no. 23, p. 11446, Dec. 2021, doi: 10.3390/app112311446.

[73] C. Lin, Y. Zheng, X. Xiao, and J. Lin, "CXR-RefineDet: Single-Shot Refinement Neural Network for Chest X-Ray Radiograph Based on Multiple Lesions Detection," *J. Healthc. Eng.*, vol. 2022, pp. 1–11, Jan. 2022, doi: 10.1155/2022/4182191.

[74] J. Park, S. Yoo, J. Park, and H. J. Kim, "Deformable Graph Convolutional Networks," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.14438

[75] L. Aziz, M. S. Bin Haji Salam, U. U. Sheikh, and S. Ayub, "Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review," *IEEE Access*, vol. 8, pp. 170461–170495, 2020, doi: 10.1109/ACCESS.2020.3021508.

[76] X. Gu and Y. Fu, "Curvature-Driven Deformable Convolutional Networks for End-To-End Object Detection," *Mob. Inf. Syst.*, vol. 2022, pp. 1–11, Feb. 2022, doi: 10.1155/2022/7556022.

[77] Y. Wang *et al.*, "Remote sensing image super-resolution and object detection: Benchmark and state of the art," *Expert Syst. Appl.*, vol. 197, p. 116793, Jul. 2022, doi: 10.1016/j.eswa.2022.116793.

[78] L. Shine and C. V Jiji, "Comparative Analysis of Two Stage and Single Stage Detectors for Anomaly Detection," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICCCNT51525.2021.9580079.

[79] Y. Umeki, I. Funahashi, T. Yoshida, and M. Iwahashi, "Salient Object Detection With Importance Degree," *IEEE Access*, vol. 8, pp. 147059–147069, 2020, doi: 10.1109/ACCESS.2020.3014886.

[80] G. Singh and A. K. Goel, "Face Detection and Recognition System using Digital Image Processing," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Mar. 2020, pp. 348–352. doi: 10.1109/ICIMIA48430.2020.9074838.

[81] L. Khelifi and M. Mignotte, "Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020, doi: 10.1109/ACCESS.2020.3008036.

[82] J. Nataprawira, Y. Gu, I. Goncharenko, and S. Kamijo, "Pedestrian Detection on Multispectral Images in Different Lighting Conditions," in *2021 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2021, pp. 1–5. doi: 10.1109/ICCE50685.2021.9427627.

[83] S. Q. Nisa, A. R. Ismail, M. A. B. M. Ali, and M. S. Khan, "Medical Image Analysis using Deep Learning: A Review," in *2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Dec. 2020, pp. 1–3. doi: 10.1109/ICETAS51660.2020.9484287.

[84] P. Viola and M. Jones, "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*, vol. 1, pp. I-511-I–518. doi: 10.1109/CVPR.2001.990517.

[85] H. Zhou and X. Song, "Lane Detection Algorithm Based on Haar Feature Based Coupled Cascade Classifier," in *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Apr. 2021, pp. 286–291. doi: 10.1109/IPEC51340.2021.9421278.

[86] A. Kumar, K. M. Baalamurugan, and B. Balamurugan, "Real-Time Facial Components Detection Using Haar Classifiers," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, May 2022, pp. 01–08. doi: 10.1109/ICAAIC53929.2022.9793034.

[87] Y. Xia, S. Yan, and B. Zhang, "Combination of ACF detector and multi-task CNN for hand detection," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, Nov. 2016, pp. 601–606. doi: 10.1109/ICSP.2016.7877903.

[88] Y. Li, Y. Tian, J. Tian, and F. Zhou, "An Efficient Method for DPM Code Localization Based on Depthwise Separable Convolution," *IEEE Access*, vol. 7, pp. 42014–42023, 2019, doi: 10.1109/ACCESS.2019.2905638.

[89] R. Andrie Asmara, M. Ridwan, and G. Budiprasetyo, "Haar Cascade and Convolutional Neural Network Face Detection in Client-Side for Cloud Computing Face Recognition," in *2021 International Conference on Electrical and Information Technology (IEIT)*, Sep. 2021, pp. 1–5. doi: 10.1109/IEIT53149.2021.9587388.

[90] T. Honda and H. Takano, "Development of Feature Extractor for Visible-Light Iris Recognition Using Multi-task CNN," in *2021 20th International Symposium on Communications and Information Technologies (ISCIT)*, Oct. 2021, pp. 83–87. doi: 10.1109/ISCIT52804.2021.9590629.

[91] S. Baek, M. Song, J. Jang, G. Kim, and S.-B. Paik, "Face detection in untrained deep neural networks," *Nat. Commun.*, vol. 12, no. 1, p. 7328, Dec. 2021, doi: 10.1038/s41467-021-27606-9.

[92] L. Hai and H. Guo, "Face Detection with Improved Face R-CNN Training Method," in *2020 the 3rd International Conference on Control and Computer Vision*, Aug. 2020, pp. 22–25. doi: 10.1145/3425577.3425582.

[93] E. S. Babüroğlu, A. Durmuşoğlu, and T. Dereli, "Novel hybrid pair recommendations based on a large-scale comparative study of concept drift detection," *Expert Syst. Appl.*, vol. 163, p. 113786, Jan. 2021, doi: 10.1016/j.eswa.2020.113786.

[94] A. Z. Agghey, L. J. Mwinuka, S. M. Pandhare, M. A. Dida, and J. D. Ndibwile, "Detection of Username Enumeration Attack on SSH Protocol: Machine Learning

Approach," *Symmetry (Basel).*, vol. 13, no. 11, p. 2192, Nov. 2021, doi: 10.3390/sym13112192.

[95] Y. Zhong and W. Deng, "OPOM: Customized Invisible Cloak Towards Face Privacy Protection," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2022, doi: 10.1109/TPAMI.2022.3175602.

[96] Z. Yu, J. Yin, Q. Zhang, W. Yang, J.-H. Xue, and Q. Liao, "Hourglass Face Detector for Hard Face," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9533674.

[97] Z. Wen, "Large-scale Face Clustering Method Research Based on Deep Learning," in *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Dec. 2021, pp. 731–734. doi: 10.1109/MLBDBI54094.2021.00143.

[98] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Res.*, vol. 40, no. 10–12, pp. 1489–1506, Jun. 2000, doi: 10.1016/S0042-6989(99)00163-7.

[99] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient Object Detection in the Deep Learning Era: An In-Depth Survey," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.09146

[100] Z. Li and S. Lou, "An Extracting and Labeling Algorithm for Connected Components in Images," in *2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, May 2021, pp. 210–213. doi: 10.1109/CAIBDA53561.2021.00051.

[101] S. Valipour, M. Siam, M. Jagersand, and N. Ray, "Recurrent Fully Convolutional Networks for Video Segmentation," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.00487

[102] T. Gomez, T. Fréour, and H. Mouchère, "Metrics for saliency map evaluation of deep learning explanation methods," Jan. 2022, doi: https://doi.org/10.1101/2021.05.05.21256683.

[103] M. Feng, K. Liu, L. Zhang, H. Yu, Y. Wang, and A. Mian, "Learning from Pixel-Level Noisy Label : A New Perspective for Light Field Saliency Detection," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.13456

[104] A. Appice, A. Cannarile, A. Falini, D. Malerba, F. Mazzia, and C. Tamborrino, "Leveraging colour-based pseudo-labels to supervise saliency detection in hyperspectral image datasets," *J. Intell. Inf. Syst.*, vol. 57, no. 3, pp. 423–446, Dec. 2021, doi: 10.1007/s10844-021-00656-7.

[105] Y. Luo, X. Liu, and X. Cao, "Improvement and Comparison of Traditional CNN and SVM Classification Based on Hog Descriptor in Pedestrian Detection," in *2021 International Conference on Artificial Intelligence and Blockchain Technology (AIBT)*, Dec. 2021, pp. 12–16. doi: 10.1109/AIBT53261.2021.00009.

[106] W. Li, W. Li, F. Yang, and P. Wang, "Multi-Scale Object Detection in Satellite Imagery Based On YOLT," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2019, pp. 162–165. doi: 10.1109/IGARSS.2019.8898170.

[107] L. Li, C. Zou, Y. Zheng, Q. Su, H. Fu, and C.-L. Tai, "Sketch-R2CNN : An RNN-Rasterization-CNN Architecture for Vector Sketch Recognition," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 9, pp. 3745–3754, Sep. 2021, doi: 10.1109/TVCG.2020.2987626.

[108] L. Bynum, T. Doster, T. H. Emerson, and H. Kvinze, "Rotational Equivariance for Object Classification Using xView," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, Sep. 2020, pp. 3684–3687. doi: 10.1109/IGARSS39084.2020.9324015.

[109] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019, doi: 10.1109/TIP.2018.2867198.

[110] S. S. L. Parvathi and H. Jonnadula, "A Comprehensive Survey on Medical Image Blob Detection and Classification Models," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, Oct. 2021, pp. 1–6. doi: 10.1109/ICAECA52838.2021.9675575.

[111] Z. Yang, Q. Li, L. Wenyin, and J. Lv, "Shared Multi-view Data Representation for Multi-domain Event Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019, doi: 10.1109/TPAMI.2019.2893953.

[112] L. Shi, L. Liu, Y. Wu, L. Jiang, and A. Ayorinde, "Event Detection and Multi-source Propagation for Online Social Network Management," *J. Netw. Syst. Manag.*, vol. 28, no. 1, pp. 1–20, Jan. 2020, doi: 10.1007/s10922-019-09493-0.

[113] K. Xiao, Z. Qian, and B. Qin, "A Survey of Data Representation for Multi-Modality Event Detection and Evolution," *Appl. Sci.*, vol. 12, no. 4, p. 2204, Feb. 2022, doi: 10.3390/app12042204.

[114] W. Zhao, Y. Hu, H. Wang, X. Wu, and J. Luo, "Boosting Entity-aware Image Captioning with Multi-modal Knowledge Graph," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.11970

[115] F. Hou, Y. Zhang, X. Fu, L. Jiao, and W. Zheng, "The Prediction of Multistep Traffic Flow Based on AST-GCN-LSTM," *J. Adv. Transp.*, vol. 2021, pp. 1–10, Dec. 2021, doi: 10.1155/2021/9513170.

[116] D. Endalie, G. Haile, and W. Taye, "Deep learning model for daily rainfall prediction: case study of Jimma, Ethiopia," *Water Supply*, vol. 22, no. 3, pp. 3448–3461, Mar. 2022, doi: 10.2166/ws.2021.391.

[117] Y. Zhou, T. Xu, H. Yang, and S. Li, "Improving Spatial Visualization and Mental Rotation Using FORSpatial through Shapes and Letters in Virtual Environment," *IEEE Trans. Learn. Technol.*, pp. 1–1, 2022, doi: 10.1109/TLT.2022.3170928.

[118] J. Sun, R. Futahashi, and T. Yamanaka, "Improving the Accuracy of Species Identification by Combining Deep Learning With Field Occurrence Records," *Front. Ecol.*

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 10 Issue: 1s
DOI: https://doi.org/10.17762/ijritcc.v10i1s.5795
***Article Received**: 10 August 2022 **Revised**: 14 October 2022 **Accepted**: 09 November 2022*

Evol., vol. 9, Dec. 2021, doi: 10.3389/fevo.2021.762173.

[119] P. Dendorfer *et al.*, "MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 845–881, Apr. 2021, doi: 10.1007/s11263-020-01393-0.

[120] "No Title." https://motchallenge.net/data/MOT16/

[121] Y. Cao *et al.*, "VisDrone-DET2021: The Vision Meets Drone Object detection Challenge Results," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 2847–2854. doi: 10.1109/ICCVW54120.2021.00319.

[122] "No Title." [Online]. Available: https://pyimagesearch.com/2022/05/02/mean-average-precision-map-using-the-coco-evaluator/.