Journal of Information System and Technology Research

journal homepage: https://journal.aira.or.id/index.php/jistr/

Pre-Review Convolutional Neural Network for Detecting Object in Image Comprehensive Survey and Analysis

[1]Gonten, [2]Fidelis Nfwan, [3]Abdulsalam Ya'u Gital

[1,2,3]Abubakar Tafawa Balewa University Bauchi (ATBU), Nigeria

ARTICLE INFO

Article history:

Received 25 February 2024
Accepted 31 May 2024
Available online 31 May 2024

Keywords:

Convolutional Network,
Object detection,
Background subtraction,
Image processing.
Computer vison

ABSTRACT

The *Convolutional neural network (CNN) has significantly exposed a great performances and growing desire in the field of image processing within the research community, through relevant innovations in object detection by magnificent capacity in transfer learning and feature learning. With the advancement of CNN in object detection, huge amount of data is process with great speed. In respect to CNN, object detection has greatly advanced and become popular in the research community, security experts, traffic experts, and remote sensing community etc. In this review, comprehensive study of various CNN architecture for object detection in images based on conventional approached, novelty, and achievement were analysed in details. Therefore, it is an important review on how to achieve high performance in object detection via CNN. We first introduced the basic idea on CNN models and their improvement in detecting object. Secondly, we review CNN and its variant such as, ResNet, VGG, GoogleNet and other CNN architectures. Thirdly, we mention some performance metrics used for object detection. Lastly, we analyse some main contribution of CNN algorithm with their remarkable achievement and further analyse the challenge and its future direction*

Corresponding Author:

Gonten, Fidelis Nfwan
Abubakar Tafawa Balewa University Bauchi, Nigeria
Email: gontenlur@gmail.com

1. INTRODUCTION

The Deep Learning (DL) framework has been extensively employed in a wide range of fields in recent years, and Convolutional Neural Networks (CNNs) are one of the most commonly used in tackling the real-time challenges of computer vision jobs since they allow for the most accurate acquisitions. The structure of biotic visual systems was used to inspire the design of CNNs[1]. In contemporary practice. Convolutional neural networks (CNNs) have found extensive use across various domains within computer vision, including tasks like image detection, classification, recognition, and tracking. Among these, image detection stands out as a particularly promising and actively researched area within computer vision, being among the most complex problems in computer vision. Image detection tasks are more intricate and challenging than object recognition tasks. Image detection is used to identify different objects in a section and assign labels to the bounding boxes of those objects. Object detection encompasses two distinct tasks: identifying the presence of objects and categorizing them. Object localization refers to pinpointing the position of an object within an image, typically achieved by delineating bounding boxes around the objects. Object localization was done before the CNN algorithm became famous by making each pixel in the image that contained the object look like Object detection, for example, was done using techniques like edge detection, drawing contours, and HOGs. This method is computationally demanding, sluggish, and inaccurate[2], [3].

Image detection and segmentation have improved dramatically thanks to the development of deep neural networks, particularly convolutional neural networks (CNN). In recent years, Schilling et al. have made a dramatic improvement by presenting

a CNN-based approach for the detection of personal vehicles constructed for the purpose of the execution of multi-sensor data attached to the Pseudo-Siamese body. The proposed system adopts output branches to enhance the efficiency of results in recognising a vehicles. presented a new fungus dataset with a CNN-based method for fungus detection and classification of various kinds of fungus, aimed at achieving conventional practices for the classification and detection of fungus. The proposed method applied different images of difficult fungal bacteria by sample extraction from spoiled fruits, records, and laboratory-hatched groups of fungus proposed a two-channel convolutional neural network (CNN) and Transfer Learning (TL) for hyperspectral image ship detection[4]. It was based on the spectral features extracted via one-dimensional spectrum vectors from the spectral medium. It was extracted through two-dimensional spatial images from spatial mediums. The TL model was designed to curb the shot-coming of two-dimensional spatial images and increase the rate of spatial feature extraction. proposed a CNN framework to enhance and simplify the detection of damage caused by natural disasters based on aerial scenes of previous disasters. The proposed method has the capacity to identify important areas with roofs, vehicles, debris, vegetables, and flooded scenes. When compared to non-CNN algorithms, object detection utilizing CNN approaches has been found to be faster and more accurate. The learning process is often computationally costly, but the real detection is quick and ideal for real-time detection[5]. The CNN method for object detection has advanced over time. In this review, we explored the different variations of the CNN algorithm employed in detecting objects, and also the remarkable achievement made recently.

The paper is organised into sections: section one introduced the basic concept of object detection with CNN, section two highlights the CNN and its variant, such as, LeNet, VGG, ResNet, GoogleNet, AlexNet, etc. Section three review various CNN architecture and their achievements. Section four explain the various metric used to evaluate performance[6].

2. RESEARCH METHODOLOGY

2.1 Convolutional Neural Networks And Their Variants

Hubel and Wiesel produced a study report on the visual cortices of primates and birds, which was the first step in developing CNN. Then, in the 1980s, Kunihiko Fukushima proposed a convolutional technique called neocognitron in the area of CNN, which was stimulated by Hubel and Wiesel's work. Nevertheless, Yann Le Cunn, who constructed a 7-level convolutional network called the LeNet-5 that applied back propagation and adaptive weight for different parameters, played a big part in getting CNN to the level it is at now. All of today's major architectures are different variations of LeNet-5[7].

CNN has been the most demonstrative deep learning model. The typical CNN design is known as the VGG-16, Each layer of a CNN is associated with a feature map, including the input layer, internal layers, filtering, and pooling. The input layer's feature map represents a three-dimensional matrix of pixel intensities corresponding to different color channels, such as RGB[8]. The feature map of internal layers forms a composite image representation where each pixel portrays a unique feature extracted from the input data. Every neuron in the preceding layer is connected to a tiny number of neurons in the layer above it. Filtering and pooling are two types of changes that can be executed on feature maps. A filter matrix transformation is convoluted during the filtering operation. To get final responses, A nonlinear function like sigmoid or ReLU is utilized along with the values within a receptive field of neurons. Pooling operations such as max pooling, average pooling, L2-pooling, and small contrast normalization aggregate the responses of a receptive field into a single value, enhancing the robustness of feature representations. Since its inception, neural networks' operation has been to emulate the functioning of the human brain to the greatest extent feasible. Convolutional neural networks get alleviated through the working principle of living beings' visual sensory organs and detect different types of objects, such as digits, images, or a specific action in any object, through a string of methods, such as convolutional operation, ReLu layer, pooling layer, flattening, and Softmax cross entropy, in a specific sequence[9].

Various research has shown that conventional CNN variants, such as LeNET, AlexNet, GoogleNet, VGG, ResNet, ResNext, and others, can Addressing complex tasks in recognition and localization, such as semantic and instance-based object segmentation, scene parsing, and scene positioning, presents significant challenges. Many prominent frameworks for object segmentation, including SSD (Single Shot Multibox Detector), region-based CNN (R-CNN), Faster R-CNN, Mask R-CNN, and Fully Convolutional Neural Network (FCN), are built upon architectures like ResNet, VGG, and Inception among others. Likewise, Libra R-CNN and others improved performance by modifying the frameworks indicated before[9], [10].

We therefore explain the following CNN variant in detail: LeNet was one of the first CNNs developed, and it was mostly utilized for digit identification. This work remains highly influential in the field of digit recognition. Introduced in 1998, this design served as a pioneering model that inspired the development of CNNs and subsequently Deep CNNs for digit recognition, particularly with the MNIST database[5], [11]. The network's fundamental architecture includes applying convolution to the input, followed by two rounds of average pooling with a stride of 2, and concluding with two fully connected layers. The dimension is 120 x 1 x 1 of the final input sent to the FCN. There are roughly 60,000 criteria that are taken into account.

The AlexNet Introduced in 2012, this architecture was among the initial CNN projects to achieve victory in the ILSVRC (ImageNet Large-Scale Visual Recognition Challenge) of that year. While sharing similarities with LeNet, this network's design is more intricate. Its operations involve convolutions of sizes 11x11, 5x5, and 3x3, alongside 3x3 maximum pooling, culminating in two fully connected layers, each containing 4096 units. Layer normalization and ReLU activation were adopted over tanh activation to optimize performance, resulting in a notable sixfold increase in speed. Dropout layers were utilized to avoid overfitting, although this increased the training duration. Because the architecture was trained for 6 days on two GTX 580 GPUs, the model is separated into two halves.

VGG, short for Visual Geometry Group, made its debut at the ILSVRC 2014 competition, where it gained significant recognition and popularity despite finishing second. VGG Net's architecture closely resembled that of AlexNet, featuring a substantial number of features. This network comprises approximately 138 million parameters. VGG-16 consists of a total of 16 convolutional layers organized into three blocks: each block includes two layers of 3x3 convolutions followed by 2x2 max pooling, and this sequence is repeated twice. The architecture concludes with two fully connected layers, each containing 4,096 hidden units.[12], [13], [14], [15]

By incorporating non-linearity within the blocks, this network achieves enhanced discriminative capabilities while consuming fewer computational resources. This efficiency is achieved by processing larger receptive fields through a series of 3x3 convolutions within each block, which reduces the number of parameters to be computed. Additionally, employing blocks allows for ReLU activation to be applied twice after each convolution within the block, further enhancing the network's performance and representational power.

The GoogLeNet followed the same block design as VGG-16, and each block is referred to as an Inception Module. In total, GoogLeNet employs nine inception modules. This design won the ILSVRC 2014 competition. This network has 22 layers; however, it compensates for this by having a smaller number of parameters (about 5 million) than the runner-up, VGG, which has 138 million[1], [16]. Each layer contains different filter sizes (1X1, 3X3, and 5X5 convolutions), and backprop determines which filter should be updated. A 3x3 max pooling layer was also included in each module. Each convolution layer was preceded by a bottleneck layer of 1x1 convolution to minimize the depth of the feature map, resulting in a considerable reduction in the number of parameters to be computed. The 1x1 convolutions were followed by ReLU activation to improve outcomes. Instead of using fully connected layers, this architecture uses "global average pooling," which effectively takes the average of each individual feature map[7], [17], [18], [19].

Microsoft ResNet, or Deep Residual Network, won the ImageNet competition for 2015, beating human accuracy errors for the first time. The network is quite deep, with 152 layers in the one presented at the challenge. Not only did the team have to tackle the Vanishing Gradient Problem as the network depth increased, but it was also discovered that as network depth climbed, accuracy became saturated and subsequently rapidly declined. ResNet countered this by introducing a skip connection feature[16], [20].

2.2 Advances Made In Detecting Objects In Images

2.2.1 Residual Network (ResNet)

proposed an improved Mask Region-Convolutional Neural Network (RCNN) based algorithm called the ResNet Group Cascade (RGC) Mask R-CNN use in object detection. Different ResNet backbone was compared and ResNeXt-101-64*4d was found to be better. On the testing stage: Mask R-CNN, the performance suffered from minor batch execution scale, causing an inaccurate computed mean and variance, therefore, group normalization was introduced to the backbone feature pyramid network neck and bounding box top of the network. The proposed model was tested on the COCO and PASCAL VOC 2007 datasets and it performed better than the baseline Mask R-CNN[20].

The Grid Convolutional Neural Network (G-CNN) architecture was introduced as a straightforward, precise, and efficient approach for object detection. This GCNN architecture was built upon ResNet as the foundation and integrated a cutting-edge object detection framework. The network introduced position-sensitive convolutional layers known as grid convolutional layers (GCL), which accurately localize objects within the feature map using grid-shaped activations. The experiment was conducted on PASCAL VOC 2007 data set which depicted that the proposed method strongly outperforms the strong baseline faster region-based convolutional neural network with a wide margin[4], [21].

Proposed an enhanced trained method via deep CNN for the automatic detection of colour-space detail within an unprocessed data pixel to encourage the deployment of HDTV and SDTV against Rec-BT-709 and Rec-BT -601 formats. The proposed method used three network kinds in training and testing the effect of the network model for precision of detection and classification. The study was conducted on Apache MXNet model and trained on three NVIDIA GTX GPU. The proposed system achieved an excellent result performed on ImageNet dataset and YouTube images for excellent colour version administered to digital images users.

proposed a Scene Mask R-CNN built on deep CNN for an accurate detection of ship and to also decrease the negative signals in unmark region. The proposed method is developed based on end-to-end technique with four network branches each with dissimilar design. The Scene Mask R-CNN joints the feature map and the inference region mask through edge probability weight, bad feature areas are exempted. The new ship detection method called Scene mask R-CNN evaluated on ship dataset sustain exactness in detection and also enhanced efficiency in decreasing the negative signals in unmark region.

Table 1. Summary of Residual Network (ResNet)

Proposed CNN architecture	Feature extraction method	Compared algorithm/s	Results
RGC Mask R-CNN[22]	ResNeXt-101-64*4d	Mask R-CNN	The model under consideration was evaluated using the COCO and PASCAL VOC 2007 datasets, demonstrating superior performance

			compared to the baseline Mask R-CNN for object detection in images.
G-CNN (GCL)[21]	ResNet-101 and Resnet-50	Faster-RCNN and R-FCN	The GCNN architecture showed significantly better performance than the baseline F-RCNN and R-FCN in object detection tasks using the PASCAL VOC 2007/12 and MS COCO datasets.
Residual Network[23]	Residual Network		System achieved an excellent result performed on ImageNet dataset and YouTube images for excellent colour version administered to digital images users.
Scene Mask R-CNN[24]	RestNet-101	Faster R-CNN	Scene mask R-CNN evaluated on ship dataset sustain exactness in detection and also enhanced efficiency in decreasing the negative signals in unmark region.
Mask R-CNN [25]	ResNet-101 and FPN	Deep neural auto-encoder, semantic segmentation algorithm, deep learning convolutional neural network based on VGG-16	The Mask R-CNN attained a total accuracy of 96.6% and 91.0% on oil spill detection and segmentation which performed better than state of the art method.
ResNet-50[26]	ResNet-50	Model-1, Model-3, Model-4, Resnet-18, and Resnet-34	The logo detection backbone of model-2 surpassed Resnet18, which had an accuracy of 0.065 mAP, and Resnet-34, which had an accuracy of 0.057 mAP, based on the experimental data while Model-2 on the other hand, had a lesser accuracy than ResNet-50, which was 0.556 mAP.

2.2.2 VGG

Proposes a VGG16 model advancement on Faster R-CNN by introducing a cross-layer fusion multi-object detection and recognition. From their experimental and data analysis, the result shows that the improved Faster RCNN architecture combines low-level and high-level image semantic attributes. This has improved image object classification and recognition. A mixed dataset was used, that is manually labelled for good detection outcome. The improved R-CNN model for object image classification and recognition has advanced the mean accuracy from the labelled cityscapes and KITTI datasets[27].

Proposed a deep convolutional neural network to handle the issue concerning poor foreground extraction in dynamic background images, an image foreground target detection technique. The foreground can accurately detect both in the complex background and foreground occlusion. VGG16 based network was used for the feature extraction map. They Applied the deconvolution technique with the pyramid pooling technique to upgrade the problem of VGG16 on classifying the entire image. The proposed model uses TensorFlow in training the network. The proposed model is compared to the traditional target model achieved good results in terms of solid robustness in a complex scene[28].

Proposed a model for object detection and tracking technique built on deep convolutional neural networks for wide-swath high-resolution optical remote sensing videos. The proposed model divides the video frames into sub-sample to avoid the challenges that occur due to data size. To detect images at the sliding window effectively, they used an overlapping sliding window sampling method. The proposed network was designed based on the region of interest (ROIs) to track images from the previous frames in the video and used directly on the current frame. The result shows the valid and general use of their dataset for the proposed algorithm[29].

Proposed A region-enhanced CNN used for detecting remote sensing images. The concept of the CNN model is based on the saliency constraints and the multilayer fusion strategy. In the proposed model, the concept of a multilayer fusion strategy is established to enable the representation of different object regions in different resolutions. Their concept adopts saliency information pixel-level loss function, being the beginning of saliency information capture through binary semantic and contextual information was used amidst multiresolution. Experiments on a public database (NWPU VHR) revealed the advantages of the RECNN over competing methods[23].

Presented a new technique called HSRI based on convolutional neural network used for detecting object through a proper measure feature. The ROI scale for detecting object is employed in designing CNN model in HSRI and also used in compiling statistical range in HSRIs. The features gotten from object via CNN is better globally. Furthermore, appropriate ROI gauged with CNN model used for detecting object is experimented with WHU-RSONE data set. A comparative analysis of the proposed model is conducted with Faster-RCNN model and the HSRIs model indicates high performance as compared to faster RCNN and enhanced better detection technique in HSRIs.

Presented CNN-Based hierarchical model with emphases on remote sensing image specifically for building detections. Gausian pyramid method was applied on the proposed model for training the hierarchical framework in view of extracting separate features at various ranges and expands spatial. The RCNN network is applied for extracting building areas which developed an efficient detection algorithm on building images. The research applied on different dataset (Dataset I, II, and III) improves the detection on building by mAP of 3.6%, 3.9% and 3.8% as compared to the state-of-the-art method and also robust in detecting building[30].

Table 2. Summary of VGG

Proposed CNN architecture	Feature extraction method	Compared algorithm/s	Results
Improved Faster R-CNN	VGG-16	Faster R-CNN	The improved Faster R-CNN performs better on cityscape and KITTI datasets than the compared algorithm
R-CNN and SPP-Net	VGG-16	Traditional Target algorithm	The proposed model as compared to the traditional target model, achieved good result in terms of solid robustness in complex scene.
Faster R-CNN	VGG-16	Faster R-CNN	The proposed model shows the accuracy and efficiency in solving object detection challenges
RECNN	VGG-16	Collection of Part Detectors (COPDs), RICNN, RAMs, SSD, YOLO, and Faster R-CNN.	The RECNN model Experiments on a public database (NWPU VHR) revealed the advantages of the RECNN over competing methods.
HSRIs	VGG-16	Faster R-CNN	The HSRIs model indicates high performance as compared to faster RCNN and enhanced better detection technique in HSRIs.
CNN-based Hierarchical	VGG-16	DPM (deformable part model), Fast RCNN, faster RCNN, RICNN and YOLO	The research applied on different dataset (Dataset I, II, and III) improves the detection on building by mAP of 3.6%, 3.9% and 3.8% as compared to the state-of-the-art method and also robust in detecting building.

2.2.3 GoogleNet

Proposed a new technique named object scale adaptive convolutional neural network (OSA CNN), comprise of object-based image analysis (OBIA) and CNN for high spatial resolution remote (HSR) sensing image classification. The proposed model accepts images with the areas of the object gotten from images dissection, also, they retrieved the squeeze and excitation network from the SE model through GoogleNet structure which further knows the weighted fusion features, improves vital features, and rejects unwanted areas. The study shows that the development further improved image classification accuracy as compared to OBIAS[17].

Table 3. Summary of GoogleNet Inception model

Proposed CNN architecture	Feature extraction method	Compared algorithm/s	result
OSA-CNN	GoogleNet	OBIA and CNN	The study shows that the development further improved image

		(GN-CNN, GNSE-CNN and GNSEM_CNN)	classification accuracy as compared to OBIAS

2.2.4 Hybridized CNN

Proposed the combination of deep-CNN and scale invariant feature transform (SIFT) for detection and classification of object through the enhancement of saliency technique and target extraction. The proposed system subsequently applied an entropy technique, executed on Deep-CNN with SIFT for strong feature selection. The strong feature selection was matrix bond, passed to the classifier for detection and classification. The performance of the proposed system was appraised on Caltech101, Barkley 3D and Pascal3d dataset with exactness of 93.82%, 88.62%, and 99.1% compared to state of the art approach for object detection and classification[31].

Proposed convolutional neural network technology with YOLOv3 detection algorithm for targeted objects are implored to achieve the identification and detection in the targeted images, analysis, and processing images via computer vision algorithm. The proposed model and the MS COCO data set are applied in training the network. The dataset (COCO) is based on the study of images of supermarket products. The proposed network is built on ResNet as a backbone network to handle degradation issues and also act as a component for feature extraction, the regression algorithm, and non-maximum suppression were used for the prediction box. The overall performance was good[32].

Proposed a hybrid method that implored a Faster R-CNN to accomplish a robust system for detecting objects with a new clustering algorithm by clustering connected fraction detections and subdue negative detections. The proposed system was designed to address the problem concerning low Object Region Percentages (ORPs) for detecting an elongated object. First, the proposed system was a train with Faster R-CNN and a fraction of the region proposal fit for ORPs. Secondly, DCNN was applied for adequate orientation and classification on ORPs. The proposed model was evaluated using the MS COCO dataset and achieved high-level development for the detection and localization of elongated[24].

Proposed a new CNN architecture based on SSD network and ResNet structure called SSD-ResNet which substitute the actual network model in other to increase the figure of layers detecting different types of dangerous goods within various background scenes. The SSD-ResNet introduced SSD as the primary or main network architecture and change the internal VGG-16 with ResNet-101 architecture. Th proposed model result depict an improvement of 17.40% accuracy through deep network as compared to the replaced VGG-16 architecture. The proposed network achieves optimally in detecting dangerous good than the main model in terms of accuracy[19].

Proposed an enhanced very-deep CNN for precision and substantial object detection built on VGG with ResNet architecture. The improved very-deep CNN deeply classifies data, with the VGG model created through very-deep CNN. Due to the training and localization issues associated with the VGG network, ResNet was introduced in VGG in other to improve the technique in handling training error and enhanced the ability of detecting small objects. To obtain high performance in classification of images and detecting object, the proposed model extracts high features. The result of the proposed model attained 85.8mAP average accuracy in object detection[33].

Table 4. Summary of Hybridized CNN architecture

Proposed CNN architecture	Feature extraction method	Compared algorithm/s	results
DCNN and SIFT	VGG and AlexNet	Ejbali 2018	the proposed system was appraised on Caltech101, Barkley 3D and Pascal3d dataset with exactness of 93.82%, 88.62%, and 99.1% compared to state of the art approach for object detection and classification.
CNN and YOLOv3	ResNet and Darknet	Traditional algorithm	The proposed CNN and YOLOv3 network performance were good on COCO dataset
Faster R-CNN and Clustering algorithm	VGG-16 and ResNet 101	Faster R-CNN and R-CNN/YOLOv3, Mask R-CNN	The proposed model achieved high-level development for detection and localization of elongated objects in image
GoogleNet and ResNet-50	GoogleNet and ResNet-50	AlexNet	Their results show that the Google inception and ResNet-50 performed better in mAP as compared to AlexNet on object detection.
Res-YOLO-R	ResNet	Tiny-YOLO and improved Tiny-YOLO	The proposed model reduced the false rate and miss detection rate, improve the detection accuracy and have a good real-

			time and generalization ability based on the comparison on study on commodity data sets.
SSD-ResNet	ResNet-101	VGG-16	Th proposed model result depict an improvement of 17.40% accuracy through deep network as compared to the replaced VGG-16 architecture. The proposed network achieves optimally than the main model in terms of accuracy.
VGG with ResNet	VGG		The result of the proposed model attained 85.8mAP average accuracy in detecting object.

2.2.5 Other CNN Architecture

Proposed two-channel convolutional neural network (CNN) and Transfer Learning (TL) for hyperspectral images ship detection. It was based on the spectral features, extracted via one-dimensional spectrum vectors from the spectral medium. It was extracted through two-dimensional spatial images from spatial mediums. The TL model was designed to curb the shot-coming of two-dimension spatial images and increase the course of spatial feature extraction[34]. The investigation result shows that the proposed algorithm enhanced the image detection accuracy of the ship while reducing the rate of false alarms.

Presented an overall concept for insulator defects detection built on new CNNs cascading framework aims at conducting localization and insulator faults detection. The CNNs cascading model was built on region proposal network (RPN) aims at changing the fault inspection twice the class of the detection problem. The proposed model employed the concept of augmenting data to handle the problem of shortage of fault images in the inspection scene based on four processes. They also used an end-to-end network training on a hug dataset. The proposed model achieved robustness and accuracy with precision of 0.91 and recall of 0.96 on insulator dataset when subjected to different fault detection settings as compared to CNNs based method[35].

Presented a lightweight deep convolutional neural network architecture called (LD-CNN) for detecting aerial vehicle images with speed and exactness. The proposed model improves the reduction of execution cost, drastically enhanced the exactness in detecting aerial vehicles, and further created a multi conditional constrained generative adversarial network also called MC-GAN for effective image generation. For performance and evaluation, Munich public database was used in training the LD-CNN model, which obtained 86.9% mAP, 0.875 F1-score, and 1.64s time for detection, implemented on Nvidia Titan XP compared to GAN and Pix2PixGAN. LD-CNN attained a state-of-the-art result.

Proposed a new feature fusion-based network architecture to detect object in UAV aerial images, built on three major layers. The proposed model is quite different from other common method; the structural learning architecture was fixed to a network to provide better robust spatial information[36]. Unsupervised deep learning techniques are employed to extract deep features and spatial information concurrently with few labelled data. The proposed FFDN result shows a great achievement and accuracy on objects with small size, occlusion and out of sight which was performed on UAV123 dataset. The tests are executed on Intel Core i9-7900 3.3-GHz CPU, a NVIDIA GTX-1080Ti GPU[37].

Proposed a new model for detecting large remote sensing images called sample update-based CNN (SUCNN) in order to advance the accuracy. The SUCNN model works with two-techniques: SSD model, used to trained the dataset and image samples are created to enhanced the dataset training by joining the background with targets images. The two-techniques were evaluated through the dataset suitable for large area remote sensing. The SUCNN model result depicts the efficiency and supremacy for detecting large remote sensing image compared to R-CNN and Faster R-CNN performed on G-2 satellite dataset[38].

Presented a new fungus dataset with CNN based method for fungus detection and classification of various kind of fungus, aim at achieving conventional practices for classification and detection of fungus. The proposed method applied different images of difficult fungal bacteria by samples extraction through spoiled fruits, records and laboratory hatched groups of fungus. The proposed method images comprise of five fungus bacteria. The developed fungus dataset enhanced accurate fungus classification and detection through the use of 40, 800 branded images. The proposed method attained 94.8% precision, which outperformed with 6.8% upgrade in detection and classification[39].

Table 5. Summary of others CNN model

Proposed CNN architecture	Feature extraction method	Compared algorithm/s	Result
CNN and TL	SPE-CNN	CNN	The CNN and TL improved the image detection accuracy of ship whereas reduces the rate of false alarm.
Cascading CNN	CNN	ILN-CNN, Faster R-CNN,	The proposed model achieved robustness and accuracy with precision of 0.91

		ILN+ACF, Cascade DNN	and recall of 0.96 on insulator dataset when subjected to different fault detection settings as compared to CNNs based method.
LD-CNN	CNN	GAN and Pix2PixGAN	The LD-CNN obtained 86.9% mAP, 0.875 F1-score and 1.64s time for detection, on Nvidia Titan XP compared to GAN and Pix2PixGAN
FFDN	CNN	Mask R-CNN, YOLOV3 and SingleNet	The proposed FFDN result shows a great achievement and accuracy which was performed on UAV123 dataset.
RCNN and YOLOv3	CNN	Traditional Matched Filtering	The results showed that DSp performed optimally via CNN-based region target detection and classification as compared to traditional matched filtering (MF) with MF dataset.
CNN Features	CNN	handcrafted BoVW features	The results depicted that the features of the CNN realise high performance as compared to handcrafted BoVW features.
Multi-class detection method	CNN	YOLOv2 FR-O, Faster R-CNN, RRPN, R-DFPN	The research evaluated on DOTA, VEDAI and VisDrone dataset achieved high detection in mAP as compared to different current practices method.
MIP-based CNNs	CNN		The MIP-based CNN model achieved a sensitivity of 95.4% compared to conventional approach for nodule detection.
SUCNN	CNN	R-CNN and Faster R-CNN	The SUCNN model result depicts the efficiency and supremacy for detecting large remote sensing image compared to R-CNN and Faster R-CNN performed on G-2 satellite dataset.
CNN	CNN	(K. Zhang, Zuo, Gu, & Zhang, 2017)	The proposed method attained 94.8% precision, which outperformed with 6.8% upgrade in detection and classification.
Crowd-SDNet	CNN	LSC-CNN, Facedness-WIDER Faster R-CNN, CSP HR-Tiny Face, PSDNN	The proposed method evaluated on many benchmarks ‘dataset outplayed the conventional methods with 10% AP and also decreases the error in counting with 31.2% for object detection and counting within a crowded area.
CNN	CNN	Traditional algorithm	The proposed method obtained excellent result both in high and low height, trained

			based on YOLO method with transfer learning TL.
CNN	CNN	KLab1, AZ3, CASIA2, RIT4, DST5, SVL3	Th system achieves powerful detection outputs and better executional capacity in a vehicle detection compared to conventional method. Their method achieved high result with F1 score of 97%.
Cascade CNN	CNN	DPM and Original Faster R-CNN	The proposed system was evaluated on railway dataset which shows accuracy and robustness in detecting different defect and extracting parts.
CNN based classifier	CNN		The proposed model performance was evaluated with CNN classifier an obtained the accuracy of 95% and also obtained 98% true positive for marine birth detection.
RRPN	CNN	Faster R-CNN, ERPN, R-CNN, SAN, CoupleNet	The proposed system can be applied to real-time computer vision applications. (mAP) and that of VGG16 is increased from 2.6% to 69.1% mAP.
CNN	R-CNN		The proposed model achieved the state-of-the-art in terms of detecting objects in oriented remote sensing images.
CLU-CNN	CNN		The CLU-CNN model was evaluated at the REFUGE CHALLENGE 2018, which achieved current best practices.
Faster R-CNN + SLD + PM + SM	CNN	Faster R-CNN, Faster R-CNN+SLD and Faster R-CNN+SLD+PM	The proposed model compared to state of the art method reduced the false and the executional time conducted on THz security dataset with good performance in precision and efficacy.

3. RESULTS AND DISCUSSION

3.1 Datasets And Performance Evaluation Metrics

3.1.1 Datasets

Different datasets were provided and utilized for object detection applications. The most widely utilized datasets are: MS COCO, PASCAL VOC, and ImageNet VID dataset. The serve as the state-of-the-art benchmark for object detection. Most datasets are group into training set and validation set. The objects existing in each image frame for ImageNet VID is less compared to a static object detection dataset like, MS COCO and PASCAL. applied PASCAL VOC and MS COCO datasets as benchmark. The datasets consist of various objects with image sizes uniformly adjusted to a ratio of 1300 x 800 pixels. Another object detection dataset called EPIC KITCHENS was introduced, which comprises 32 various kitchens in different cities, having 11, 500, 000 frames with 454, 158 ground truth across 290 classes. Furthermore, different datasets exist for generic or specific applications: NWPU VHR-10 and Airport data, Cityscape and KITTI, DIOR dataset, JONATHAN, KITTI, UAV123 data set, GF1-test-dev and GF2-test-dev WHU-RSONE[1], [21], [37], [40], [41], [42], [43].

There is presently no open domain dataset that provides compressed annotations for multiple complicated scenarios for object detection in image with classification labels and tight bounding boxes. further work is needed to develop comprehensive datasets in order to progress object detection.

Table 6. Image datasets

Data	Source	Data size
NWPU VHR-10 and Airport data	and Google Earth	NWPU VHR-10: Has 650 images in different sizes. Airport data > 10000 × 10000 size.
PASCAL VOC 2007/12	www.doi.org, GitHub repository, and	PASCAL VOC 207/12 comprises of 20 natural images that are labelled with 20 different types of objects.
MS COCO and PASCAL VOC07	GitHub repository and Kaggle	The image size is uniformly adjusted to a ratio of 1300 × 800 pixels.
PASCAL VOC 2007/12 and MS COOC	Kaggle	PASCAL VOC 2007/12 Consist of 800 images. MS COCO has 80 objects category with resolution of 1300 x 800 pixels
Sonar data	obtained by Marine Sonic Technology and EdgeTech.	The optimal patches of the image size are 37 × 37 with almost 81 348 data sample
		X
PASCAL VOC and CLAHE	Girshick, et al., 2016	PASCAL VOC 2007/12 Consist of 800 images.
Cityscape and KITTI	Kaggle	-
MS COCO	Kaggle	MS COCO has 80 objects category with resolution of 1300 x 800 pixels
Urban Atlas	GitHub repository	The Urban Atlas comprises images of 21 classes and satellite images of 224x224
HIS data	Geospatial Data Cloud, http://www.gscloud.cn/	HIS dataset with400 hyperspectral images and 3000 multispectral images
RSOD-Dataset and Jilin-1 Video-03	Geospatial Data Cloud, http://www.gscloud.cn/	The width and height of the video is 12000x5000 pixels. The randomly crop 200 images with size 300x300
Microsoft COCO	Kaggle	Foveated image sampling reduced in size from 416 × 416 to 128 × 128 at interval of 32 pixels
Microsoft COCO	Zhang, Yang, Zhang, and Zhu, 2016	Image resolution of 64 x 64 pixel
Microsoft COCO	GitHub Repository	The model was trained on 21 object class with input of size 416×416
Dbp2-11 with 627 X-ray images	Github	627 X-ray images: 256 × 256 fixed-sized
MSTAR (DSp and DNsp)	Geospatial Data Cloud, http://www.gscloud.cn/	Sparse SAR image with resolution of 0.3 × 0.3
DIOR dataset	Geospatial Data Cloud, http://www.gscloud.cn/	The image resolution is 800 × 800. The feature map sizes of five scales are 200 × 200, 100 × 100, 50 × 50, 25 × 25, and 13 × 13, in that regard.

CARLA, KITTI and Panoramic dataset	GitHub Repository	CARLA has 200 images (2048 × 300 resolution), KITTI has 7482 images (1242 × 375 resolution) and 5,000 image (4608 × 3456 resolution) for panoramic dataset
HSR dataset	GitHub repository	The patches of the image are resized to 224×224×3 resolution
Munich public dataset	GitHub repository	
NWPU VHR-10 data set	Google Earth	spatial resolution ranging from 0.5 to 2 m and spatial resolution of 0.08 m.
UAV123 data set	(Mueller, Smith, & Ghanem, 2016)	15 filters of 7×7 pixels at the 1-th layer, and both 30 filters of 5×5 pixels at the 3-th and 5-th layers.
GF1-test-dev and GF2-test-dev	Google Earth	GF-1 and GF-2 images with resolution of 2.0-m (18 000×18 192 pixels) and 0.8m (27 620×29 200 pixels)
WHU-RSONE	http://pan.baidu.com	The image resolution is 0.8 m with 2500×2500 pixels
Database I, II, and III	Github.com and Digital Globe's WorldView-2 satellite.	Image with the resolution of 5400x6000 m^2,8000x6000m2, and 4700x5000 respectively.
NWPU VHR-10 and aircraft dataset.	Github.com and Digital Globe's	
DOTA, VEDAI and VisDrone	GitHub repository	DOTA has 2806 images with 800 to 4000 pixel, VEDAI has 1241 images with 1024x1024 and 512x512 resolution
MNIST, CIFAR-100 and CIFAR-10	GitHub repository	The CIFAR-10 dataset takes 32x32 colour images and 6000 images /class, with 60000 images. CIFAR-100 takes 100 classes of images with 600 images each.

3.2 Performance Evaluation Metrics

It is critical to standardize how to evaluate the performance of techniques using integrated datasets. In computer vision, various performance metric was deployed to measure the percentages of detected object. Different metrics are applied in CNN algorithm for object detection in image to evaluate their performances and analysis the best detection technique. Detection in complex scene is the same as detection in a non-challenging scene, the algorithms use the same performance evaluation metrics

To understand the metrics proper, we need to understand the various parameters of the metrics[44], [45]:

1. **True Positive (TP):** is the total number of detected image pixels matching the detected pixels within the ground truth. It can also refer to hit.
2. **False Positive (FP):** true positive is the total number of detected image pixels matching the un-detected pixels of the ground truth. Also refers to as false alarm.
3. **True Negative (TN):** TN compute the total number of detected image pixel matching the un-detected pixel within the ground truth. Is also called correct rejection.
4. **False Negative (FN):** FN is the total number of un-detected image pixel matching the pixels detected within the ground truth. It is also referred to as miss.

We hereby, discuss different most utilized evaluation metrics used in object detection:

Precision: when the ground truth fits into the percentage of a predicted region, then it is known as precision. The following is the precision formula:

$\frac{Predicted\ area\ in\ ground\ truth}{Total\ area\ of\ predicted\ region} = \frac{TP}{TP+FP}$. where TP means true positives and FP denotes.

Recall: The percentage of the ground truth region that is present in the anticipated region is calculated as recall. Recall is computed using the following formula: $\frac{\text{Ground truth area in predicted region}}{\text{Total area of ground truth region}} = \frac{TP}{TP+FN}$, where TP stands for true positives and FN stands for false negatives.

F-measure: is calculated through the mean of precision and recall. It can be expressed mathematically as: $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$,.

Specificity: Evaluates all the background pixel based on percentage which is adequately un-detected. It can be mathematically represented as: Specificity = $\frac{TN}{(TN+FP)}$.

Accuracy: the accuracy metric evaluates all the object pixels in percentage for images that are perfectly detected and excluded. It can be mathematically represented as: Accuracy = $\frac{TP+TN}{(TP+TN+FN+FP)}$.

IOU (Intersection Over Union): presented an object detection metric, commonly known as the Jaccard index, is one of the widely used evaluation metrics for determining the effectiveness of object detection technique. the developed training sets for object detection by drawing bounding boxes around labelled objects. The bounding box and the ground truth overlapped. It's a metric for determining how much the projected region overlaps with the real ground. Region of truth. The following is how an IOU is defined: $\frac{\text{Area of Overlap region}}{\text{Area of Union region}}$.

Average Precision: calculates the average precision value across different recall levels. The greater the AP computed, the higher the performance, respectively. The formular below is used for computing average precision: $AP = \sum_n (R_n - R_{n-1}) P_n$, Where the nth threshold for precision and recall are, Rn and Pn

Mean absolute error: the mean absolute error evaluates the average pixel-level total change between predicted value and ground truth. Computed as follow: $\sum_{i=1}^{D} |x_i - y_i|$,

The mean Average Precision (mAP) is widely used as performance evaluation metrics in detecting category-specific object. The mAP is calculated over different classes. It can be express, mathematically as: $mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$, where AP_i is the average precision for a required category and *N* denotes the overall number of category

Table 7. Performance metric

Precision	Recall	F1-score	IOU	Specificity
√	√	X	X	X
√	X	X	X	X
√	X	X	X	X
√	√	X	X	X
X	X	√	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	√	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X

√	√	X	X	X
√	√	X	X	X
√	√	√	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	√	X
√	√	X	X	X
√	√	X	X	X
√	√	X	X	X
√	√	X	√	X
√	√	X	X	X
√	√	X	X	X
√	√	X	√	X

3.3 Object Detection

Object detection is regarded as one of the fundamental and basic fields in computer vision with two famous approaches. Each has its own set of benefits and drawbacks based on CNN. Detection and recognition of objects is made up of two different concepts, such as localization and classification techniques. Object detection is defined as the process of determining where objects are situated in an image (localization) and to which class objects belong (classification). As a result, old techniques for object recognition include three stages in their pipeline: informative region selection, extraction of features, and classifications.

Current advancements in algorithm modification were accomplished, such as feature extraction and detection subnets, which are applied in two-stage detectors and one-stage detectors. Certain ideas, like the region proposals, which are proposed to modify the RPN process, are focused on the two-stage. Others, like loss function development, which is recommended to decrease sample imbalance, are based on one-stage. These enhancements improve detection quality and efficiency, allowing object identification based on CNN to be subjected to a wider range of computer vision applications, such as autonomous driving, security systems, etc. Due to the high intra-class and low inter-class variance, image segmentation or object detection is a difficult challenge. The result of different items belonging to a single class, such as varied stances of individuals or having different backgrounds in an image, is high intra-class variance. Low inter-class variance occurs when similar-looking objects from distinct classes, such as class chair samples, are easily incorrectly classified as class bench samples, and vice versa. Sliding windows were used in one of the first techniques for object detection algorithms, with classification applied to each window to discover objects. Consequently, RPN was used to narrow the search before applying categorization, replacing the sliding window paradigm. Object identification systems, among other domains, have benefited from the current explosion in deep learning.

Presented an object detection paper which subsequently splits object detection methods into three categories: object detection, salient object and category-specific. Object detection is the process of detecting objects irrespective of their class. The detection technique often generates a huge number of viable region proposals, from which the top images are chosen based on a set of criteria. Salient Object Detection, these methods highlight and recognize objects in an image or video using the human attention technique. Category-specific Object Detection seeks to recognize many images in a single category. Unlike OD and SOD, they must forecast the object's category class and location in the image. presented Object detection methods based on CNN are divided into two types: two-stage object detectors and one-stage object detectors. R-CNN, Fast R-CNN, and Faster R-CNN are two-stage object detection architectures that separate the problem of object localization from the task of object categorization. They use region-suggestive algorithms to identify possible places where objects exist. Mask R-CNN introduced later segmentation output and superior detection pooling approaches. On the other hand, one-stage object detection algorithms produce candidate regions before classifying them as objects or no-objects.

One-stage detectors like YOLO and SSD, for example, use feature pyramid networks (FPNs) as a backbone to detect objects at many scales in a single pass rather than predicting regions and then classifying them. Object detection's output, like those of other computer vision domains, is heavily reliant on spatial data. As a result, when occlusions obscure the objects, objects have varying sizes, or background information is combined with the objects, the object detection system's performance suffers. In real-time applications, it's common for the object detection network to receive input images that aren't very detailed or were taken in low-light conditions. Over time, several pre-processing approaches have been used to increase image quality and improve object detection performance in difficult settings. Traditional methods rely on image enhancement and manual feature selection methods to improve image quality. Due to their robustness and generalization capabilities, Deep Neural Networks (DNNs) eventually supplanted these techniques.

Table 8. Object detected in the images

Objects detected by the propose CNN architecture
Airplane, ship, storage tank, baseball diamond, tennis court, basketball court, harbour, bridge, vehicle, Mairport and ground track field.
Person, cat, dog, cow, motor-bike, car, bottle, sheep, chair, table, bicycle.
Jug, bird, Zebra, rugby ball, plate, Person, cat, dog, cow, motor-bike, car, bottle, sheep, chair, table, bicycle, etc.
Aeroplane, bicycle, birds, boats, bottle, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorbikes, people, potted plants, sheep, sofas, trains, TV/Monitors
Dam, Shipwrecks, Coral reef etc.
Ship
Vehicle, pedestrian, motor-bike, cyclist
Supermarket Product (canned drinks and ice cream, etc.)
Water, Forest and Agriculture
Ship spectrum, Land spectrum and Sea spectrum
Car, trucks, person, cyclist
aircraft, oil tank, playground and overpass
Person, bicycle, car, motorbike, aeroplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, cow.
Baseball bat, toothbrush, wrench, pen, screwdriver
Motorbike, person, bird, bottle, train, sofa, car, aeroplane, sheep, dog, potted plant, cat, chair, bicycle, etc.
handguns, glass bottle, Firearm Component, Ceramic Knife, Laptop, Camera, Firearm and Knife
Airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, gold field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle and wind mill.
Vehicle
Water, Forest, Building, Land, Road, grassland, bare land
Aerial Vehicles
Airplane, Ship, storage tank, baseball court, tennis court, basketball court, ground track field, harbor, bridge, vehicle, mean AP
Bike, Boat, Car, Person, Car, Boat, cyclist, and bus.
Airport, Airplane
Airplane, tank storage and ship
Airplane, ship, storage tank, baseball, diamond tennis court, baseball court, ground track field, harborm, bridge, vehicle
Building
Airplane, ship, storage tank, baseball, diamond tennis court, baseball court, ground track field, harborm, bridge, vehicle
bridge, small vehicle, storage tank, large vehicle, plane and ship

Bed, airplane, bicycle, bird, cat, bus, chair, deer, couch, motorcycle, dog, frog, streetcar, horse, table, train, ship, wardrobe, truck.
Airplane, basketball diamond, basketball court, bridge, crossroad, ground track field, harbour, parking lot, ship, storage tank, T-junction, tennis court and vehicle
Airport, civil aircraft, fighter, helicopter, transport plane, bridge and oil tank
Nodule image
Airplane, ship, storage tank, residential areas, industrial areas, grasslands, and woodlands

Table 9. Object detection method in images

Background Modelling method	Ground truth
The model is generated best on statistical feature extracted from the image background	Pixel wise labelling
Current image frame is computed with matrix transform	Bounding Box
Background and foreground subtraction with feature selection	Bounding Box
Current image frame is computed with inverse transform	Bounding Box
Background and foreground subtraction with feature selection	Bounding Box
Background and foreground subtraction with feature selection	Pixel wise labelling
Background and foreground subtraction with feature selection	Bounding Box
The cluster algorithm is approved to get the important anchor boxes.	Anchor boxes.
The cluster algorithm is approved to get the important anchor boxes.	Contour on images
Background and foreground subtraction with feature selection	Rectangle boxes
Background and foreground subtraction with feature selection	Bounding Box – Regression
Background and foreground subtraction with feature selection	Bounding Box – Regression
Current image frame is computed with matrix transform	Bounding boxes
Current image frame is computed with matrix transform	Bounding boxes (Rectangular boxes)
multi-stage system. Separates out background from the foreground object in sequence in images.	Bounding boxes (Rectangular boxes)
Background and foreground subtraction with feature selection	Pixel wise labelling with bounding box
Background and foreground subtraction with feature selection	Pixel wise labelling with bounding box

multi-stage system. Separates out background from the foreground object in sequence in images.	Bounding boxes (Rectangular boxes)
Current image frame is computed with matrix transform	3D Bounding boxes
multi-stage system. Separates out background from the foreground object in sequence in images.	Pixel wise labelling
Current image frame is computed with matrix transform	Prediction area for Ground truth box
multi-stage system. Separates out background from the foreground object in sequence in images.	Bounding boxes (Rectangular boxes)
multi-stage system. Separates out background from the foreground object in sequence in images	Bounding boxes (Rectangular boxes)
multi-stage system. Separates out background from the foreground object in sequence in images.	Pixel wise labelling with bounding box
multi-stage system. Separates out background from the foreground object in sequence in images.	Pixel wise labelling with bounding box
The model is generated best on statistical feature extracted from the image background	Bounding boxes (Rectangular boxes)
The model is generated best on statistical feature extracted from the image background	Bounding boxes (Rectangular boxes)
The model is generated best on statistical feature extracted from the image background	Bounding boxes (Rectangular boxes)
multi-stage system. Separates out background from the foreground object in sequence in images.	Bounding boxes (Rectangular boxes)
multi-stage system. Separates out background from the foreground object in sequence in images.	Bounding boxes (Rectangular boxes)

3.4 Meta-Data Analysis

After a comprehensive search technique which enable us to identify virtually must relevantly studies on object detection in images via CNN, we performed quantitative appraisal and synthesize outcomes across studies to obtain information on significant and relevant achievement. We therefore, discuss in detail those major CNN achievements extracted from the literature review.

Various research has shown that conventional CNN variants, such as LeNET, AlexNet, GoogleNet, VGG, ResNet, ResNext, and others, can solve difficult recognition and localization challenges such as semantic and instance-based object segmentation, scene parsing, and scene positioning. Most well-known object segmentation frameworks, including Sigle Short Multibox Detector (SSD), region-based CNN (R-CNN), Faster R-CNN, Mask R-CNN, and Fully Convolutional Neural Network (FCN), are based on ResNet, VGG, and Inception, among others. Likewise, Libra R-CNN and others improved performance by modifying the frameworks indicated before.

Achieved result applied on different dataset (Dataset 1, 2, and 3) which improves the detection on building by mAP of 3.6%, 3.9% and 3.8% as compared to the state-of-the-art method and also robust in detecting building. proposed model was tested via the least trained and validation loss value, which attained a total accuracy of 96.6% and 91.0% on oil spill detection and

segmentation which performed better than state of the art method achieved 4.7 seconds with Resnet-50 faster than Model-2 and the result was also superior to Resnet-18 and ResNet-34 Values for one stage logo detection. proposed model achieved robustness and accuracy with precision of 0.91 and recall of 0.96 on insulator dataset when subjected to different fault detection settings as compared to CNNs based method. For performance and evaluation, Munich public database was used in training the LD-CNN model, which obtained 86.9% mAP, 0.875 F1-score, and 1.64s time for detection. presented MIP-based CNN model which achieved a sensitivity of 95.4% compared to conventional approach for nodule detection. developed fungus dataset which enhanced accurate fungus classification and detection through the use of 40, 800 branded images. The proposed method attained 94.8% precision, which outperformed with 6.8% upgrade in detection and classification. proposed method evaluated on many benchmarks' dataset outplayed the conventional methods with 10% AP and also decreases the error in counting with 31.2% for object detection and counting within a crowded area. achieves powerful detection outputs and better executional capacity in a vehicle detection compared to conventional method. Their method achieved high result with F1 score of 97%. proposed model performance was evaluated with CNN classifier an obtained the accuracy of 95% and also obtained 98% true positive for marine birth detection. The proposed system demonstrates that the ZF-Net parameters was compress by 81.2% and save 66% computing power. proposed system achieved 4x speed in frame rate, ranging from 3.59 FPS – 15.244 FPS under 416 x 416 and 128 x 128 pixel. The reduction of the image experience less in recall outcome of 92.0%, and a decrease of 50.1% against the baseline in full image size proposed model attained 38.7% accuracy and 0.6 seconds per image on 3D-detection, higher compared to state-of-the-art algorithms on various datasets. investigation result shows that the proposed model effectively increases by 12.7% mAP with 0.307s compared to Faster R-CNN in detecting object. proposed model evaluated on the hand-rising dataset produced an impressive result with detection accuracy of 90% in mAP, sufficient for real-world application when compared to state-of-the-art object detection methods.

3.5 Challenges and future research

3.5.1 Challenges

The basic concept of object detection using various CNN technique is to accept image frames from either a static or moving environment through camera in other to displays a binary mask representation of object for different frame, sequentially. However, this procedure is not a simple job due to number of complexity and challenges associated with. further presented in details, some of the challenges enumerated bellow:

1. Light challenges: fluctuation of light in the environment on the object target as a result of movement of light source, light from bright surfaces, climate in outdoor scenarios, light interrupt from other object, and changing time of the day, etc. the straight effect of these changes causes FN detection for technique through background appearance modelling.
2. Unstable object appearance: many objects can change in 3-dimension space in real scene, they only have the 3-dimensional projection motion on a 2-dimension images in sequence, therefor rotation in the path of the third base will likely affect the appearance of the object.
3. Sudden variations in movement: another issue in object detection and tracking is the sudden variation in the speed and direction of the object's movement, as well as sudden camera movement. These object movement or camera movement if not corrected, the background modelling algorithms will fail to appropriately detect objects. more so, a quick movement create a ghost detected area.
4. Occlusion: the objects in the image scene may be block or obscure by another object. Partial occlusion occurs when some parts of the object are block behind other objects, while complete occlusion occur when objects are entirely block by others. Occlusion has a significant effect on object detection via the background modelling algorithm.
5. Background complexity: the complexity of background is based on highly textured image, particularly in outdoor scene where there is a lot of texture variation. Furthermore, some backgrounds are subject to frequent changes due to, traffic light, fountain, clouds, shaking trees, and water waves, etc.
6. Shadow: shadow present in image makes it more difficult to detect an object. The shadows are caused when a light ray cannot pass through and opaque object. Furthermore, the shadow form by the object will greatly affect the accuracy of the object detection.
7. Camera issues: a variety of issues such as video capture devices, acquisition technique, compression strategies, and camera stability, can all have an impact on the quality of a images. Additionally, block artifacts and blur degrade image quality.
8. Non-rigid deformed object: varied segment of dynamic object may have different motion based on speed and orientation. Most methods detect various segment as different object in motion when dealing with object in motion. It poses a significant problem, particularly with non-rigid object in the presence of moving cameras.

3.5.2 Future research

The management of different state-of-the-art techniques through CNN algorithm model have remarkably change and elevated the research community, especially in computer vision such as, object detection, pedestrian detection, autonomous vehicle, traffic control, etc. in future, more powerful CNN architectural models will be developed in the research community:

1. presented one of the promising research field in CNN called Ensemble learning. The proposed technique will combine many and different methods to guide the technique in improving the resilience and generalization on many image categories through extracting distinct levels of semantic representation.

2. In image segmentation tasks, CNNs' ability to learn generatively has been leveraged, showing promising performance. Harnessing CNNs' generative learning capabilities during feature extraction can enhance the model's representational capacity. Novel approaches are needed to enhance CNNs' learning efficacy by enriching the feature maps with informative content.
3. In vision system, one of the key methods in acquiring information from image is attention. The method of attention works The attention mechanism is designed to capture important information from an image while also maintaining contextual relationships with other visual components. In future research, there is potential to explore approaches that preserve not only the spatial importance of objects but also their distinguishing characteristics during subsequent stages of learning.
4. CNNs involve numerous hyperparameters, including activation functions, kernel sizes, number of neurons per layer, layer configurations, and more. Tuning these hyperparameters is a time-consuming and intuitive process that cannot be explicitly specified. Genetic algorithms offer a solution by automatically optimizing hyperparameters through a combination of random searches and guided searches based on previous results.
5. The concept of a parallel pipeline can be employed to increase the scale of CNN training and overcome hardware limitations. This pipelining approach could be leveraged in the future to accelerate the training of large models and improve performance scalability without the need for hyperparameter tuning.
6. Cloud based platforms is projected to fully utilized the creation of highly compute-intensive CNN domains in the future. Cloud computing has the capability of not only permits the processing of large amounts of data, In addition, it offers substantial computational power at an affordable price. Moreover, the cloud environment simplifies the setup of libraries for researchers, including those new to the field.
7. Because CNNs are primarily used for computer vision, implementing Applying state-of-the-art CNN architectures to sequential data involves transforming 1-dimensional data into 2-dimensional data. The adoption of 1D-CNNs for sequential data is gaining popularity due to their strong feature extraction capabilities and efficient computation with a reduced number of parameters.
8. Recent time, researchers at CERN specializing in high-energy physics have started employing the learning capabilities of CNNs to analyze particle collisions. The utilization of machine learning, particularly CNNs, in high-energy physics is expected to grow in the future.

4. CONCLUSION

The advancement of CNN has made an astonishing development, specifically in the field of computer vision and also rejuvenated the inquisitiveness of researchers and academicians in artificial intelligence (AI). The enhancements in CNN are considered in diverse ways, such as optimization, regularization, activation, novelty in architecture loss function, and learning algorithm. To enhance the efficiency and functionality of the CNN algorithm, several investigative works where performed.

This paper review convolutional neural network for detecting objects in images, especially based on building concepts of the processing techniques and the proposed framework for recent CNN algorithm and its variant. We first introduced the general concept of object detection, looking at the traditional method and the contemporary practice. Furthermore, we review and categorised the CNN algorithm into various Variant and other classes such as, ResNet, VGG, GoogleNet, hybridized CNN and CNN novelties with the achievement. We observed in recent research work that the substitution of conventional layer structure with blocks was the remarkable and main achievement in CNN performance. Conventionally, the paradigm shift of research in building CNN structure is the development of novelty and efficient block design. The function of the block network is that of the auxiliary learner which used the spatial or feature map detail to enhance performance. The blocks contribute in enhancing CNN efficacy through feature learning and transfer learning. We further introduced different performance metrics used for object detection and also explain the background of object detection. Lastly, we analyse some main contribution of CNN algorithm with their remarkable achievement which attained the state-of-the-art and also analyse the challenge and its future direction.

REFERENCES

[1] Y. Liu, Z. Zhang, X. Liu, L. Wang, and X. Xia, “Ore image classification based on small deep learning model: Evaluation and optimization of model depth, model structure and data size,” *Miner. Eng.*, vol. 172, 2021, doi: 10.1016/j.mineng.2021.107020.

[2] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, “An ensemble stacked convolutional neural network model for environmental event sound recognition,” *Appl. Sci.*, vol. 8, no. 7, 2018, doi: 10.3390/app8071152.

[3] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, 2018, doi: 10.1016/j.neunet.2018.07.011.

[4] H. M. Al-Jawahry, A. H. A. Hussein, G. Sunil, M. A. Alkhafaji, and P. K. Pareek, “Faster Region-Convolutional Neural Network with ResNet50 for Video Stream Object Detection,” in *3rd IEEE International Conference on Mobile Networks and Wireless Communications, ICMNWC 2023*, 2023. doi: 10.1109/ICMNWC60182.2023.10435864.

[5] S. Sahu, R. Kumar, M. S. Pathan, J. Shafi, Y. Kumar, and M. F. Ijaz, “Movie Popularity and Target Audience Prediction

Using the Content-Based Recommender System," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3168161.
[6] A. M. Ali, B. Benjdira, A. Koubaa, W. El-Shafai, Z. Khan, and W. Boulila, "Vision Transformers in Image Restoration: A Survey," *Sensors*, vol. 23, no. 5, 2023, doi: 10.3390/s23052385.
[7] U. S. Bhargavi *et al.*, "Brain Tumor Detection Using Convolutional Neural Network," in *Communications in Computer and Information Science*, 2023. doi: 10.1007/978-3-031-35641-4_40.
[8] T. Sinha and B. Verma, "A Non-iterative Radial Basis Function Based Quick Convolutional Neural Network," in *Proceedings of the International Joint Conference on Neural Networks*, 2020. doi: 10.1109/IJCNN48605.2020.9206798.
[9] J. H. Yoon and B. Jang, "Evolution of Deep Learning-Based Sequential Recommender Systems: From Current Trends to New Perspectives," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3281981.
[10] M. Dhurkunde, N. Kadam, M. Trivedi, S. Maru, and P. Shirke, "Multi-class Brain Tumor Detection using Convolutional Neural Network.," *Grad. Res. Eng. Technol.*, 2023, doi: 10.47893/gret.2023.1158.
[11] M. A. Mahjoubi, S. Hamida, O. El Gannour, B. Cherradi, A. El Abbassi, and A. Raihani, "Improved Multiclass Brain Tumor Detection using Convolutional Neural Networks and Magnetic Resonance Imaging," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, 2023, doi: 10.14569/IJACSA.2023.0140346.
[12] D. Kusumawati, A. A. Ilham, A. Achmad, and I. Nurtanio, "Vgg-16 and Vgg-19 Architecture Models in Lie Detection Using Image Processing," in *Proceeding - 6th International Conference on Information Technology, Information Systems and Electrical Engineering: Applying Data Sciences and Artificial Intelligence Technologies for Environmental Sustainability, ICITISEE 2022*, 2022. doi: 10.1109/ICITISEE57756.2022.10057748.
[13] Y. Tao, "Image Style Transfer Based on VGG Neural Network Model," in *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications, AEECA 2022*, 2022. doi: 10.1109/AEECA55500.2022.9918891.
[14] C. Shu and X. Hu, "Improved Image Style Transfer Based on VGG-16 Convolutional Neural Network Model," in *Journal of Physics: Conference Series*, 2023. doi: 10.1088/1742-6596/2424/1/012021.
[15] Z. Zhao and S. Zhang, "Style Transfer Based on VGG Network," *Int. J. Adv. Network, Monit. Control.*, vol. 7, no. 1, 2022, doi: 10.2478/ijanmc-2022-0005.
[16] N. I. Mahbub, F. Naznin, M. I. Hasan, S. M. R. Shifat, M. A. Hossain, and M. Z. Islam, "Detect Bangladeshi Mango Leaf Diseases Using Lightweight Convolutional Neural Network," in *3rd International Conference on Electrical, Computer and Communication Engineering, ECCE 2023*, 2023. doi: 10.1109/ECCE57851.2023.10101648.
[17] S. Zhang, Y. H. Gong, and J. J. Wang, "The Development of Deep Convolution Neural Network and Its Applications on Computer Vision," *Jisuanji Xuebao/Chinese J. Comput.*, vol. 42, no. 3, 2019, doi: 10.11897/SP.J.1016.2019.00453.
[18] I. Ahmed, M. Ahmad, A. Chehri, M. M. Hassan, and G. Jeon, "IoT Enabled Deep Learning Based Framework for Multiple Object Detection in Remote Sensing Images," *Remote Sens.*, vol. 14, no. 16, 2022, doi: 10.3390/rs14164107.
[19] S. Arya and R. Singh, "A Comparative Study of CNN and AlexNet for Detection of Disease in Potato and Mango leaf," in *IEEE International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2019*, 2019. doi: 10.1109/ICICT46931.2019.8977648.
[20] N. Yamsani, M. B. Jabar, M. M. Adnan, A. H. A. Hussein, and S. Chakraborty, "Facial Emotional Recognition Using Faster Regional Convolutional Neural Network with VGG16 Feature Extraction Model," in *3rd IEEE International Conference on Mobile Networks and Wireless Communications, ICMNWC 2023*, 2023. doi: 10.1109/ICMNWC60182.2023.10435819.
[21] Z. Gu, Z. Luo, M. Huang, Y. Cai, and S. Li, "A Graph-Convolutional-Network based Prototype Mixing Model for Few-shot Segmentation," in *Proceedings - 11th International Conference on Information Technology in Medicine and Education, ITME 2021*, 2021. doi: 10.1109/ITME53901.2021.00028.
[22] B. Azam *et al.*, "Aircraft detection in satellite imagery using deep learning-based object detectors," *Microprocess. Microsyst.*, vol. 94, 2022, doi: 10.1016/j.micpro.2022.104630.
[23] H. Lyu, "Research on Corrosion Recognition Method of Steel Based on Convolutional Neural Network," in *2023 IEEE 6th International Conference on Information Systems and Computer Aided Education, ICISCAE 2023*, 2023. doi: 10.1109/ICISCAE59047.2023.10393077.
[24] D. O. Melinte and L. Vladareanu, "Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified adam optimizer," *Sensors (Switzerland)*, vol. 20, no. 8, 2020, doi: 10.3390/s20082393.
[25] A. J. Xiang, A. B. Huddin, M. F. Ibrahim, and F. H. Hashim, "An Oil Palm Loose Fruits Image Detection System using Faster R -CNN and Jetson TX2," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, 2021. doi: 10.1109/ICEEI52609.2021.9611111.
[26] L. Yang *et al.*, "An Improving Faster-RCNN With Multi-Attention ResNet for Small Target Detection in Intelligent Autonomous Transport With 6G," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, 2023, doi: 10.1109/TITS.2022.3193909.
[27] A. Bagaskara and M. Suryanegara, "Evaluation of VGG-16 and VGG-19 Deep Learning Architecture for Classifying Dementia People," in *Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021*, 2021. doi: 10.1109/IC2IE53219.2021.9649132.
[28] R. Dash, M. Bag, D. Pattnayak, A. Mohanty, and I. Dash, "Automated signature inspection and forgery detection utilizing VGG-16: a deep convolutional neural network," in *2023 2nd International Conference on Ambient Intelligence in Health Care, ICAIHC 2023*, 2023. doi: 10.1109/ICAIHC59020.2023.10431430.
[29] W. Zhang, "Computer Identification of Mango Leaf Disease Based on Adversarial Denoising Autoencoder Model," in *Proceedings - 2022 International Conference on Machine Learning and Intelligent Systems Engineering, MLISE 2022*,

2022. doi: 10.1109/MLISE57402.2022.00101.
[30] Venkatesh, Y. Nagaraju, T. S. Sahana, S. Swetha, and S. U. Hegde, "Transfer Learning based Convolutional Neural Network Model for Classification of Mango Leaves Infected by Anthracnose," in *2020 IEEE International Conference for Innovation in Technology, INOCON 2020,* 2020. doi: 10.1109/INOCON50539.2020.9298269.
[31] Z. Zhang, Q. Jiang, Y. Zhan, X. Hou, Y. Zheng, and Y. Cui, "VAE-GAN Data Enhancement Networks-based Model for Rolling Bearing Few-shot Fault Classification," *Yuanzineng Kexue Jishu/Atomic Energy Sci. Technol.*, vol. 57, 2023, doi: 10.7538/yzk.2023.youxian.0198.
[32] A. N. Yumang, C. J. N. Samilin, and J. C. P. Sinlao, "Detection of Anthracnose on Mango Tree Leaf Using Convolutional Neural Network," in *2023 15th International Conference on Computer and Automation Engineering, ICCAE 2023*, 2023. doi: 10.1109/ICCAE56788.2023.10111489.
[33] N. Srivastava, S. Ruhil, and G. Kaushal, "Music Genre Classification using Convolutional Recurrent Neural Networks," in *2022 IEEE 6th Conference on Information and Communication Technology, CICT 2022*, 2022. doi: 10.1109/CICT56698.2022.9997961.
[34] M. Kiruthiga Devi, U. Surya, K. Unnamalai, and R. K. Tharani, "Treatment for Insomnia using Music Genre prediction using Convolutional Recurrent Neural Network," in *2022 1st International Conference on Computational Science and Technology, ICCST 2022 - Proceedings*, 2022. doi: 10.1109/ICCST55948.2022.10040279.
[35] Y. Li, "Research on Application of Convolutional Neural Network in Intrusion Detection," in *Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, 2020. doi: 10.1109/IFEEA51475.2020.00153.
[36] P. Mittal, A. Sharma, R. Singh, and V. Dhull, "Dilated convolution based RCNN using feature fusion for Low-Altitude aerial objects," *Expert Syst. Appl.*, vol. 199, 2022, doi: 10.1016/j.eswa.2022.117106.
[37] H. Long, Y. Chung, Z. Liu, and S. Bu, "Object Detection in Aerial Images Using Feature Fusion Deep Networks," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2903422.
[38] Y. Hu, X. Li, N. Zhou, L. Yang, L. Peng, and S. Xiao, "A Sample Update-Based Convolutional Neural Network Framework for Object Detection in Large-Area Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, 2019, doi: 10.1109/LGRS.2018.2889247.
[39] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An ensemble of convolutional neural networks for audio classification," *Appl. Sci.*, vol. 11, no. 13, 2021, doi: 10.3390/app11135796.
[40] R. Aulia, Suendri, and M. Alda, "Build Android-Based Go Service Electronic Applications Using React Native Framework and Firebase Realtime Database," *J. Inf. Syst. Technol. Res.*, vol. 2, no. 3, pp. 102–114, 2023, doi: 10.55537/jistr.v2i3.615.
[41] Kaggle.com., "No Title." [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification
[42] T. Liu, L. Yang, and D. Lunga, "Change detection using deep learning approach with object-based image analysis," *Remote Sens. Environ.*, vol. 256, 2021, doi: 10.1016/j.rse.2021.112308.
[43] A. Ikhwan, R. A. A. Raof, P. Ehkan, Y. Yacob, and M. Syaifuddin, "Data Security Implementation using Data Encryption Standard Method for Student Values at the Faculty of Medicine, University of North Sumatra," *J. Phys. Conf. Ser.*, vol. 1755, no. 1, 2021, doi: 10.1088/1742-6596/1755/1/012022.
[44] A. A. Kabanov, "Application of Support Vector Machines to the Multiclass Classification Electromyography Signal Patterns," in *Proceedings of the 2021 15th International Scientific-Technical Conference on Actual Problems of Electronic Instrument Engineering, APEIE 2021*, 2021. doi: 10.1109/APEIE52976.2021.9647434.
[45] P. C. Upadhyay, L. Karanam, J. A. Lory, and G. N. Desouza, "Classifying Cover Crop Residue from RGB Images: A Simple SVM versus a SVM Ensemble," in *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*, 2021. doi: 10.1109/SSCI50451.2021.9660147.