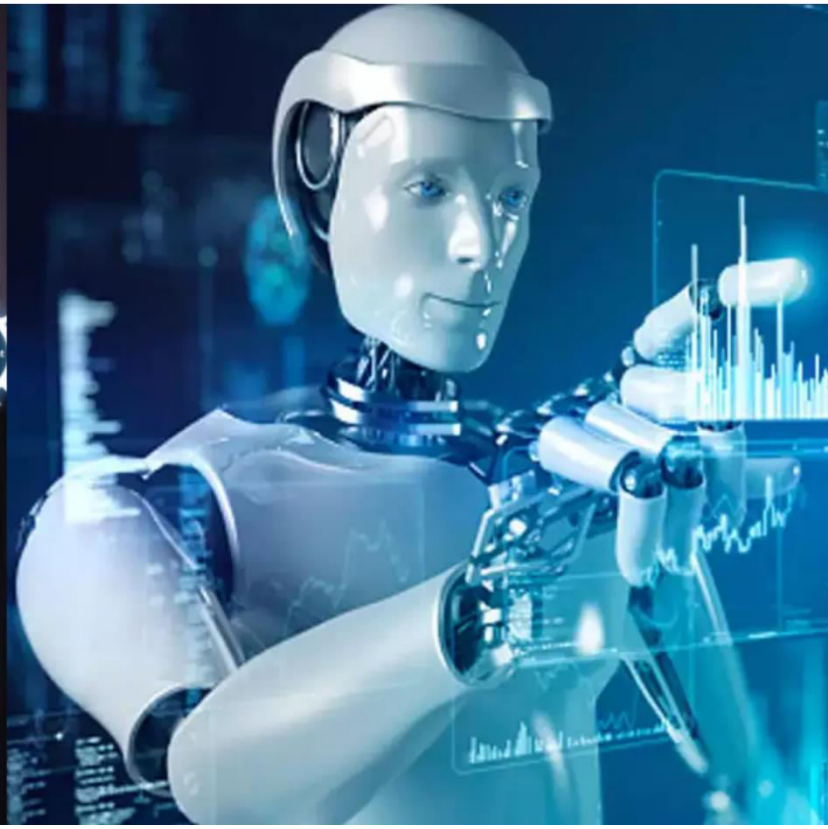
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

Exploring Deep Learning Innovations in Object Detection: a Detailed Survey

G.Banupriya[1], B.Sakthivel[2], J.Sivasankari[3], K.Kavitha[4]

Faculty, ULTRA College of Engineering and Technology, Madurai, Tamil Nadu, India[1,3,4]

Faculty, Pandiyan Saraswathy Yadav Engineering College, Madurai, Tamil Nadu, India[2]

ABSTRACT: In computer vision tasks, the most crucial issue is object detection. AlexNet's proposal, other studies on neural networks and various modifications to algorithm structures have emerged, as techniques based on Convolutional Neural Networks (CNNs) have gained popularity in the field of computer vision. There are many obstacles to overcome in order to produce quick and precise detection effects, which require leaving the current CNN framework. Researchers are investigating the use of transformers in computer vision tasks because to their remarkable performance in natural language processing (NLP). By redefining detection as a set prediction problem, DEtection TRansformer (DETR) introduces transformers to object detection tasks. As a result, post-processing and proposal generation are no longer necessary. At first, DETR struggled with sluggish training convergence and inefficient detection of tiny objects, even though it performed competitively. To address these problems, however, a number of enhancements are suggested, which will significantly boost DETR and allow it to perform at the cutting edge. It indicated new possibilities for research and demonstrated that transformer-based models outperform neural network techniques. This article presents traditional neural networks, talks about the benefits and drawbacks of convolutional neural networks for object detection algorithms, and presents the most recent cutting-edge Transformer techniques in computer vision. Lastly, the future progress and limitations of Transformers and convolutional neural networks in object detection are examined.

KEYWORDS: Transformer, Object Detection, DETR, Computer Vision, Deep Neural Networks

I. INTRODUCTION

When it comes to many computer visions (CV) applications, object detection is the most crucial area for research. Correctly classifying the object's category and placement within the provided image using a rectangular bounding box is the objective and aim.

The two stages of object detection can be separated based on the development process: the conventional object detection method and the current CNN methods. The conventional object detection models, like the color thresholding, [4], HOG [1], and others, are inadequate; they need a lot of computation, work slowly, and may yield several accurate recognition results rather than the ones we desire, so they no longer satisfy the demands of the present.

Numerous creative novel detection techniques have been developed by researchers, yet these techniques have also introduced new issues. For instance, the transformer-based detection approach necessitates a large amount of training data, the two-stage object identification method has a slow detection speed, and the one-stage object detection method has a low detection accuracy. It is crucial for researchers to understand how these constraints arise as well as how suggestions for enhancements arise. It is hoped that by examining, contrasting, and debating these detection techniques in the paper, researchers would gain a deeper comprehension of them and generate fresh concepts for new method proposals.State-of-the-art results on the PASCAL VOC object detection competition have been achieved using these discriminate local feature descriptors and shallow learnable architectures, and real-time embedded systems with little hardware load have been achieved. However, in 2010–2012, only the construction of ensemble systems and the use of slight modifications of effective techniques yielded modest advances the following factors are responsible for this fact:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

1) Using a sliding window approach to generate candidate bounding boxes is redundant, ineffective, and imprecise.

2) The combination of discriminatively-trained shallow models with manually-engineered low-level descriptors is insufficient to close the semantic gap.

A greater benefit is achieved with the introduction of Regions with CNN characteristics (R-CNN), which is made possible by the development of Deep Neural Networks (DNNs). The most representative CNNs, or DNNs, behave very differently from conventional methods. At present, the CNN algorithm has an absolute position in both the general target detection method and the special object target detection method. Almost all object detection algorithms use CNN [2] as their backbone.

Numerous enhanced models have been proposed since R-CNN was first proposed. These include YOLO, which uses a fixed-grid regression to accomplish object detection [11], Faster R-CNN, which uses an additional sub network to generate region proposals [16], and Fast R-CNN, which simultaneously optimizes classification and bounding box regression tasks [17].

II. TRADITIONAL DETECTION METHODS

Due to their lengthy processing times, traditional object identification methods are not real-time. Additionally, the accuracy falls short of what is needed to execute actual applications. For non-neural networks to detect objects, features must first be extracted using any of the techniques such as the Viola-Jones object detection framework based on Haar features, Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG) features.

Traditional Vision-based Methods	High Efficiency, Low Computational Burden	Easy To Be Affected By The Complex Background; Depending On Accumulated Experience To Extract Image Features
Deep Learning-based Methods	High Accuracy And Feature Extraction Ability	Depending On Large Amounts Of Data.

FIGURE 1 Comparison of Traditional and Deep Learning Detection Methods

Color thresholding [4] is applied to detect discoloration or flashovers, relying on the distinct color variations between defective and non-defective regions.RGB color features [4], local binary patterns (LBP) [5], and histograms of oriented gradients (HOG) [6], which were applied to insulator defect detection. Scale-Invariant Feature Transform (SIFT) [4] extracts distinctive invariant features from images and it is a useful tool for matching between different views of an object .

An SVM classifier [33] is then used for classifications. SVMs may effectively handle high-dimensional feature spaces and are resistant to over fitting because of their margin-maximization approach, particularly when used in conjunction with feature extractors such as HoG, SIFT, or PCA. They can handle complex, non-linear data distributions because of their versatility in using different kernel functions. SVMs can be used as powerful lightweight classifiers in systems with limited resources because they require less computing power to train than deep learning models. Furthermore, because of their interpretability and stability, they are frequently employed as trustworthy baseline models in object identification studies.

However, complex outdoor conditions, including rain, snow, fog, and other weather factors, significantly affect detection accuracy. These traditional features are highly sensitive to variations in shape, lighting, and complex backgrounds. With advancements in computational power, deep convolutional neural networks (CNNs) have become increasingly popular for object detection tasks, offering a more robust solution.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

III. OBJECT DETECTION BASED METHODS USING DEEP LEARNING:

Object detection algorithms divided into two types according to the feature extraction methods used, one is a CNN-based object detection method, and the other is a transformer-based object detection method. Each type contains a different method of object detection, and in order to allow the reader to have a better understanding of these methods and more thinking about object detection, The paper list the advantages, disadvantages and areas for improvement of each method.

CONVOLUTIONAL NEURAL NETWORKS

Numerous academics have concentrated on CNN-based object identification models since deep learning's enormous success. Many object identification models have been developed with straightforward architecture and effective results, making them suitable for a wide range of applications. Object classification and object location are the two main issues that detection models must address. After determining whether there are any objects of interest and classifying them, the detection system outputs the objects' most likely category scores. Classifying images is the foundation of the object location problem. We also want to know the object's location within the picture. Utilize rectangular boxes to enclose the items that have been detected. For instance, the rectangular box can be created by calculating the coordinates of the upper left and lower right corners or by calculating the length and width of the rectangular box and the top left coordinate. A feature map in CNN is defined as a new image created by applying a convolution kernel to an existing image. The region of the original image that is impacted by the pixels in the high-level feature map is equal to the size of the receptive field on the feature map. One-stage object detection and two-stage object detection are the two general categories into which object detection algorithms fall.

Two stage algorithms (candidate-based algorithms)

Two-stage algorithms need two steps to complete tasks involving object detection. The two-stage methods first identify potential regions in the image, known as region proposals, and then categorize and locate the region proposals using CNN to obtain the results. Although the detection speed of two-stage algorithms is slow, the accuracy of the results is quite high. SPPNet, R-CNN, and Fast R-CNN, More fast two-stage object detection methods like R-CNN are common.

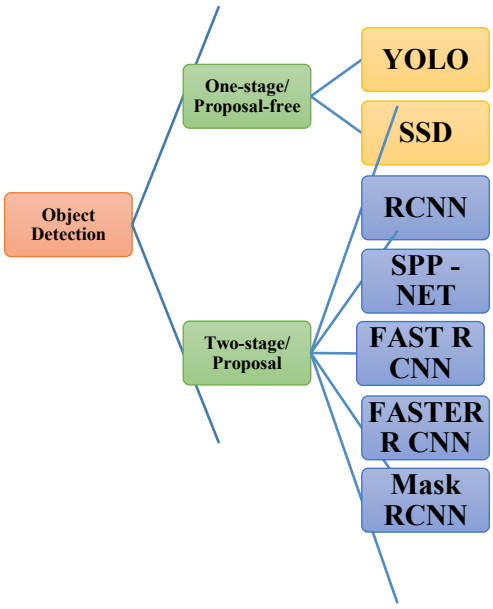


FIGURE 2: Object detection framework overview

These methods are highly accurate but computationally intensive. Regression/classification-based methods(One stage)[11], like YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector)[7], directly predict object classes and bounding boxes in a single pass, making them faster and suitable for real-time applications, though sometimes less accurate for detecting small or overlapping objects. The integration of advanced architectures, such as ResNet, EfficientNet, and Vision Transformers (e.g., Swin Transformer), has enhanced feature extraction capabilities,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

enabling multi-scale detection and improved robustness in challenging conditions. Techniques like data augmentation, transfer learning, and synthetic dataset generation address issues like class imbalance and small training datasets. Recent trends also include hybrid models, combining object detection with segmentation for pixel-level accuracy, and lightweight architectures optimized for edge devices and real-time deployment. Despite significant progress, challenges remain in detecting small, occluded, or densely packed objects and adapting models for unconstrained environments with varying lighting, occlusions, and viewpoints. Future research aims to address these issues through unsupervised learning, 3D object detection, and leveraging more efficient architectures. Deep learning continues to push the boundaries of object detection, enabling applications in autonomous driving, video surveillance, medical imaging, and robotics.Object detection models are widely used to locate and classify defects in insulators within an image.

R-CNN:

R-CNN (Region-Based Convolutional Neural Network) is a deep learning architecture operates in three key stages: generating region proposals using Selective Search, extracting features from each region using a CNN (such as AlexNet)[8], and classifying objects with Support Vector Machines (SVMs)[9] while refining bounding box coordinates through a regression model is shown in figure below.

R-CNN detection divided into four:

1. Receive input images.
2. Extract about 2000 region proposals.
3. Input region proposals and compute CNN features.
4. Use support vector machines (SVM) to determine classification.

Even though R-CNN produced positive results at the time, it had a lot of drawbacks.

- Training takes a long time because it is done in phases and the feature map for each candidate region is calculated independently.
- A significant amount of space will be used as a result of the feature map of each potential location being preserved for use in later procedures.
- Test duration is lengthy because each region proposal's feature map is computed independently and not shared. Consequently, there is a significant amount of computation required for every image in the test, which takes a long time.

SPP-NET (Spatial Pyramid Pooling Network)

The input images are fixed size images, and the feature extraction is inefficient, to name just two of R-CNN's shortcomings. SPP-NET feeds the full image to the convolutional layer without than requiring a specified size for the input images.\

A black feature map will be produced once an image has traversed a convolutional layer. In particular, the input feature map is split into blocks of various scales, such as 4*4, 2*2, and 1*1 in the three distinct scales (the block scale is not solid). Each block is then subjected to max pooling, resulting in a vector with a constant length of 16 + 4 + 1. As a result, after input images of various scales travel through the SPP layer, they will receive vectors of the same length.

Despite SPP-net's several enhancements over R-CNN, it still uses R-CNN's multi-stage procedure. For instance, extract 2000 candidate regions using the selective search approach, extract the feature map using a CNN network, and then enter the retrieved features into an SVM for classification, etc. But it also highlighted some of R-CNN's drawbacks, as outlined at the outset, and suggested fresh approaches to address them. In order to avoid repeating feature extraction for each region suggestion, SPP-Net innovated by immediately feeding an entire image into the convolutional neural network. Once the SPP-Net layer is used, the model is capable of processing images with varying sizes.

Drawbacks of SPP- NET

SPP-Net's drawbacks include the inability to fine-tune the network's parameters prior to the SPP-layer, which results in extremely low efficiency; when fine-tuning, each training sample (RoI: region of interest) originates from a

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

different image, creating extremely low back propagation efficiency in the SPP layer, which causes it to re-produce a new feature map for each image, which is inefficient; Fast-RCNN has improved it.

Fast R-CNN:

Fast R-CNN, an enhanced version of R-CNN, has been successfully applied to insulator defect detection due to its efficiency and accuracy in object detection tasks. The architecture integrates a Region of Interest (RoI) pooling layer that enables the CNN to process the entire image once, extracting shared features for all region proposals, which significantly reduces computation time. For insulator defect detection, Fast R-CNN[17] utilizes a pretrained CNN[8] backbone (e.g., VGG16 or ResNet) to capture high-level features, followed by the RoI pooling to standardize feature maps for region-specific detection. These features are then passed to fully connected layers for classification and bounding box regression, identifying defects such as cracks, contamination, and missing parts in insulator surfaces. Studies have demonstrated that Fast R-CNN[8] provides higher detection accuracy and faster processing times compared to traditional methods, making it suitable for practical applications in power line inspection where real-time performance is critical. Its ability to generalize and adapt to complex backgrounds further enhances its effectiveness in challenging field environments.

Faster R-CNN:

Faster R-CNN [16] is a highly efficient and accurate object detection technique, primarily known for integrating Region Proposal Networks (RPNs) into the CNN architecture, enabling it to simultaneously propose regions of interest (ROIs) and classify objects within those regions. Compared to other CNN-based object detection techniques, such as R-CNN and Fast R-CNN, Faster R-CNN offers a significant performance improvement. While R-CNN and Fast R-CNN rely on an external process for generating region proposals, Faster R-CNN eliminates this bottleneck by introducing the RPN, which allows for end-to-end training, streamlining the detection pipeline. This integration not only speeds up the process but also improves the accuracy of region proposals, as the RPN is trained alongside the rest of the network. When compared with single-pass methods like YOLO and SSD, Faster R-CNN is often more accurate, particularly in handling complex object scenarios or smaller objects, as it processes the image in a two-stage manner (first proposing regions, then classifying and refining them). However, Faster R-CNN tends to be slower than YOLO and SSD because of its more computationally intensive two-stage process. While YOLO and SSD offer faster detection speeds, they may sometimes sacrifice accuracy, especially in challenging environments with multiple overlapping or small objects. Overall, Faster R-CNN strikes a balance between accuracy and speed, making it a preferred choice in scenarios where precision is paramount, but its computational demands can be a limitation for real-time applications. Faster R-CNN is typically more accurate and precise, particularly in challenging detection tasks, while YOLO excels in speed and efficiency, making it more suitable for real-time applications where detection speed is prioritized over perfect precision.Backbone architecture of CNN is shown in figure

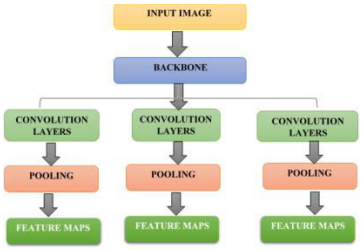


FIGURE 4: CNN Backbone Architecture for Feature Extraction

Mask R-CNN:

Mask R-CNN, an extension of Faster R-CNN, has been effectively applied to insulator defect detection due to its capability to perform object detection, classification, and instance segmentation simultaneously. The architecture consists of a backbone network (e.g., ResNet or ResNeXt) for feature extraction, a Region Proposal Network (RPN) to generate candidate regions, and two parallel branches for bounding box regression and class prediction. A third branch, unique to Mask R-CNN, predicts pixel-wise binary masks for each detected object, allowing precise localization of defect regions such as cracks, flashovers, and contaminations on insulator surfaces. The RoIAlign layer ensures accurate mapping of feature maps for region proposals, improving segmentation accuracy over the traditional RoIPool.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

Mask R-CNN [18] not only detects defects but also provides precise segmentation masks, helping in better visualization and measurement of defect areas. This makes it particularly suitable for applications in insulator maintenance and inspection using UAVs, where fine-grained analysis of surface conditions is essential.

One-Stage Detectors (Regression and Classification based):

YOLO (You Only Look Once):

YOLO (You Only Look Once) is a deep learning-based object detection method known for its speed and efficiency. Unlike traditional object detection methods that apply a region proposal network (RPN) and classify objects in separate stages, YOLO frames object detection as a single regression problem. It divides an image into a grid and predicts bounding boxes and class probabilities directly from each grid cell. This allows YOLO to detect multiple objects in a single forward pass, making it extremely fast and suitable for real-time applications.

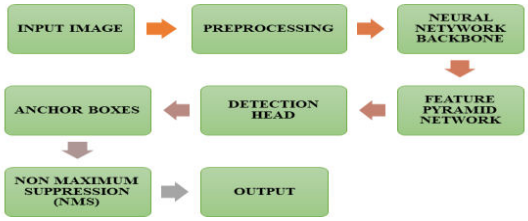


FIGURE 5: YOLO Processing Pipeline for Object Detection

The input image is processed through a Convolutional Neural Network (CNN) to extract features, after which fully connected layers predict both the class probabilities and the bounding box coordinates. The image is divided into a grid, with each cell responsible for predicting multiple bounding boxes. To refine the results, Non-Maximum Suppression (NMS) is applied to filter out overlapping boxes, keeping only the most probable ones for accurate object detection.the processing pipeline architecture is shown in figure

YOLO's architecture has evolved through several versions, such as YOLOv1, YOLOv2, YOLOv3, and the more recent YOLOv5 and YOLOv8, with each version offering improvements in accuracy, speed, and robustness.
YOLOv1: The initial version used a simple grid approach but struggled with small object detection.

YOLOv2 (YOLO9000): Introduced anchor boxes and multi-scale processing, improving detection for various object sizes. YOLOv3: Enhanced multi-scale detection using feature pyramids and better loss functions.

YOLOv4: Implemented Spatial Pyramid Pooling (SPP) and Cross-Stage Partial (CSP) connections to boost accuracy.
YOLOv5: Transitioned to an efficient architecture with anchor-free detection and transfer learning techniques.
YOLOv6: Focused on lightweight architecture for faster computation using EfficientNet-Lite.

YOLOv7: Added improvements like nine anchor boxes and focal loss for better small-object detection.

SSD (Single Shot MultiBox Detector):

The Single Shot MultiBox Detector (SSD) is a fast and efficient object detection model that performs both object classification and bounding box regression in a single pass through the network. It uses a pre-trained Convolutional Neural Network (CNN) as the backbone to extract multi-scale feature maps from the input image, allowing it to detect objects at different sizes. SSD applies convolutional layers on each feature map to predict class scores and bounding box coordinates for predefined anchor boxes at each spatial location. These anchor boxes serve as reference points, with SSD predicting adjustments to better match the objects in the image. After generating predictions, SSD uses Non-Maximum Suppression (NMS) to eliminate redundant and overlapping bounding boxes, leaving only the most confident detections. The model is end-to-end trainable, combining a localization loss for bounding box prediction and a confidence loss for class prediction. While SSD is efficient and suitable for real-time applications, it may struggle with very small or densely packed objects. However, its use of multi-scale feature maps and the ability to detect objects of various sizes makes it versatile and competitive in terms of both speed and accuracy, particularly in scenarios requiring fast detection.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

In summary, SSD strikes a balance between accuracy and speed, making it suitable for diverse object sizes, while YOLO excels in speed but may sacrifice precision, and CNN-based methods are more accurate but slower.

RetinaNet:

The RetinaNet architecture consists of three main components: a backbone, a Feature Pyramid Network (FPN), and subnetworks for classification and bounding box regression. The backbone, typically a ResNet (e.g., ResNet-50 or ResNet-101), extracts feature maps from input images, while the FPN builds a multi-scale feature representation to capture both small and large object details. Each level of the FPN produces feature maps at different resolutions. The classification subnetwork predicts the probability of each anchor box containing a defect, and the regression subnetwork refines the coordinates of these boxes. A key innovation in RetinaNet is the focal loss function, which assigns higher weights to hard-to-classify examples, addressing the class imbalance issue inherent in object detection tasks.It strikes a balance between speed and accuracy, making it ideal for applications such as defect detection in insulators. .The ResNet backbone extracts multi-scale features from input images, passing them through multiple convolutional layers to create feature maps at various resolutions. These feature maps are then fed into the FPN, which generates a feature hierarchy with consistent semantic information across scales, enhancing the model's ability to detect objects of varying sizes. At each pyramid level, anchor boxes are evaluated using the classification subnetwork to predict object classes and the regression subnetwork to adjust the coordinates of bounding boxes. The classification subnet uses fully connected layers with ReLU activations and focuses on predicting class probabilities, while the regression subnet fine-tunes anchor boxes to fit detected objects accurately. This multi-scale and hierarchical design makes RetinaNet a powerful model for dense object detection.

IV. TRANSFORMERS-BASED

Transformers [23] are revolutionary in object detection because they allow models to capture long-range dependencies and global context in an image, which are difficult for regular CNNs to do. Transformers are a fully end-to-end trainable architecture that uses self-attention techniques to directly predict object classes and bounding boxes from image characteristics. They are first used in models such as DETR (DEtection TRansformer), which replace hand-designed components (such as region proposal networks). As a result, anchor boxes and heuristic post-processing stages can be removed, overlapping or occluded objects can be handled better, and object localization becomes more accurate. Additionally, transformers allow for collaborative reasoning across all objects in an image, which makes them very useful for dense object recognition jobs and complicated scenarios.

Several works related to Transformer in the field of object detection will introduce in
below: DETR [22] and D-DETR [34] are examples of previous work that combined Transformer and CNN for classification and instance segmentation. ViT [35], which developed a Transformer-based computer vision model, employed pure Transformer directly for classification. & The concept of ViT proposes Swin Transformer [36], a new universal backbone that employs Transformer to manage computer vision tasks.

Detection Transformer (DETR)

In the area of target detection, DETR [6] can be considered Transformer's groundbreaking achievement. After several adjustments, the DETR based on ResNet50 matches the performance of Faster-RCNN, and the DETR result is excellent.

Actually, it is simple to comprehend the DETR design as a whole. It is made up of three primary parts:

- Feed forward network
- Encoder-decoder transformer
- CNN backbone network

Using the input images, the CNN backbone network creates feature maps. The CNN backbone network's output is then transformed into a one-dimensional feature map and sent as input to the transformer encoder. N fixed-length embeddings (vectors), where N is the number of objects in the image that the model assumes, are the encoder's output. These embeddings are decoded into bounding box coordinates by the Transformer decoder using the encoder-decoder attention technique.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

Lastly, the linear layer employs the SoftMax function to predict the category label using normalized center coordinates, the height and breadth of the bounding box predicted by the feedforward neural network.

DETR's benefits are that it requires less prior knowledge and suggests an alternative approach for object detection. On the COCO dataset, the accuracy and efficiency are comparable to the highly optimized Faster R-CNN, although not being the best. Larger objects will have a stronger detecting effect. The components needed for DETR may be obtained in any deep learning framework, and since it doesn't require any custom layers, rebuilding the model is simple. The disadvantages are also obvious:

DETR has problems such as slow convergence speed, poor detection accuracy, and low
Operating efficiency

Vision Transformer (ViT)

The original Vision Transformer's (ViT) primary function was image classification. ViT's main concept is to use the Transformer Encoder, which was first created for natural language processing, to classify picture data directly. ViT flattens and embeds an image into a series of vectors by dividing it into fixed-size patches (such as 16x16 pixels) as an alternative to convolutional layers. The Transformer Encoder processes these vectors and use self-attention mechanisms to capture global correlations between patches. The final representation of the patch sequence is utilized for classification, and a special [CLS] token is prepended to it. Especially in large and complicated datasets, this method enables the model to incorporate global context and long-range dependencies, which are frequently advantageous for classification tasks.

ViTs to function consistently across different object scales, handle occlusions, and recognize objects in complex scenarios. They are well-suited for end-to-end learning, as demonstrated by models like DETR and Swin Transformer, because they can analyze whole images as sequences of patches, removing the requirement for manually constructed elements like anchor boxes or region proposal networks. However, because ViTs lack the powerful inductive biases (such as locality and translation invariance) that CNNs have, they may perform poorly on smaller datasets and need big datasets and significant computational resources to train well. Notwithstanding these drawbacks, they are an effective tool for contemporary object detection jobs due to their scalability and adaptability.

Swin Transformer

For image classification transformers are used. However, there are two significant issues with these methods:

1) To convey its meaning, an image must have a minimum of a few hundred pixels. Transformer is not good at processing a large number of pixels in a row.

2) All of the other techniques were designed to address object classification problems; nevertheless, in theory, Transformer is more effective at solving detection problems; however, it still requires improvement in its capacity to handle the complex instance segmentation prediction scenario.

In order to address the aforementioned issues, Swin Transformer was used which effectively enabled the model to produce various algorithmic outcomes in the areas of classification, detection, and instance segmentation. Swin Transformer is now the preferred BackBone when utilizing Transformer for visual creation. Instead of using small-resized images as input images like the prior ViT and iGPT, Swin Transformer directly enters the initial images to avoid information loss from shrinking. One advantage of Swin Transformer is that it makes use of CNN's most widely used hierarchical network topology. The receptive field of the node expands with the complexity of the hierarchical network structure. The more complicated the hierarichical network structure, the larger the node's receptive field will be. The Swin Transformer is likewise endowed with this property, allowing it to identify and separate items like FPN and U-Net structures.

V. COMPARTIVE STUDY OF DEEP LEARNING MODELS

.Ning.P [27] proposed the improved Faster R-CNN model's mean Average Precision (mAP) from 72.7% to 85.9%.The method used in Fast R-CNN by Girshick (2015)[15] introduces key improvements over the earlier R-CNN architecture for object detection, focusing on computational efficiency and streamlined training.It allows

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

backpropagation through the RoI pooling layer, enabling end-to-end learning for region proposal processing and feature extraction. Ren et al[16] replaces the computationally expensive Selective Search used in Fast R-CNN,generates region proposals by sliding a small neural network across the feature map achieves mAP of 73.2 %..Li, W.[3] [17] introduced a improved fast R CNN by multi-scale feature fusion technique to enhance the detection of small defects and utilized a more efficient feature extraction network to improve accuracy of 93.5%.Li, Q [18] used Mask R-CNN to perform instance segmentation, which allowed for the precise pixel-level localization and classification of defect regions within the aerial images achieved precision of 96.8%. Zhou,W[19] proposed improved YOLOv3 model for detecting defects in insulators using images captured by Unmanned Aerial Vehicles (UAVs).Li,S[20]chose YOLOv4, a state-of-the-art object detection model, for this task due to its accuracy, speed, and robustness in real-time object detection.It is known for improved performance in detection tasks over previous versions, and it is suitable for detecting small and complex objects. Carion, N., Massa[23] unlike traditional methods, DETR does not use anchors or region proposals. Instead, it formulates object detection as a direct set prediction problem. Lin,T[1][24] developed FPN on Faster R-CNN (with ResNet-101 backbone) achieved a mAP of 36.2% on the COCO test-dev set, which was a substantial improvement over the baseline Faster R-CNN without FPN (which had a mAP of 34.0%).Zhang Y [29] introduced a Depth wise Separable Convolution fused with a Squeeze-and-Excitation (DSC-SE) module to replace standard convolution in the YOLOv7 backbone, aiming to reduce network parameters and improve feature extraction and produce accuracy of 95.2.Qiang ,H[30]propose enhancements to the YOLOv8s model to improve detection accuracy in complex backgrounds. The improvements include integrating the Triplet Attention (TA) mechanism to enhance feature extraction, employing an improved Bidirectional Feature Pyramid Network (BiFPN) for better feature fusion with precision of 92.1.Jiang.T [31] proposes CDDCR–YOLOv8 algorithm enhances insulator defect detection by segmenting images into grid cells, each predicting target presence and produce location. Li et al.[33] introduced a novel defect recognition method for power insulators utilizing convolutional neural networks (CNNs). The proposed approach constructs an efficient binary classification model that accurately identifies insulator defects within complex backgrounds

Author	year	Model	Precision	Recall	mAP ($AP/AP_{50}/AP_{75}$)
J. Tang[1]	2024	YOLOv8-ISDD	-	-	88.9
Zhendong He [1]	2024	YOLOv8s-SwinT	95.6	95.3	97.7
Jiang, T[1]	2024	CDDCR–YOLOv8	-	-	97.5
Yue Zhang[1]	2023	GAN improved Faster R CNN	-	-	AP_{50} =96.18 AP_{75}=91.08
Qiang, H[1]	2023	YOLOv8s	92.1	88.4	92.4
Zhang Y	2023	YOLOV7	95.2	-	-
Li, Q[1]	2021	Improved Mask R CNN	96.8	-	-
Li, S.[1]	2021	Yolov4	92.4	90.3	- 91.2
Carion, N., Massa[1]	2020	DEtection TRansformers (DETR)	-	-	42
Zhang, X[1]	2018	Improved Fast R-CNN	93.5	-	-
Ren et al[1]	2015	Faster R-CNN	-	-	73.2

TABLE 1 Comparison of Experimental results of insulators for different object detection models

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

VI. CONCLUSION

Region proposal-based deep learning models, such as Faster R-CNN, achieve high accuracy but come with significant computational complexity. In contrast, one-stage detectors like YOLO and SSD deliver faster inference, making them ideal for real-time applications, though they may face challenges in detecting smaller or overlapping defects. Recent innovations, including transformer-based models and hybrid CNN-Transformer architectures, have enhanced detection accuracy, particularly in complex environments. Lightweight YOLO models, including YOLOv4-tiny and YOLOv5n, provide an optimal balance between speed and efficiency, making them well-suited for deployment on resource-limited devices. On the other hand, transformer-augmented YOLO variants enhance feature extraction and spatial awareness but require greater computational power. Feature enhancement techniques such as Feature Pyramid Networks (FPN) and BiFPN further improve multi-scale object detection. Ultimately, selecting the right model depends on application needs. Lightweight YOLO models are preferable for real-time defect detection on devices with limited processing capabilities, while transformer-based and hybrid models offer greater precision for more complex detection tasks, albeit with increased computational costs. Future research should prioritize improving model robustness, refining lightweight architectures, and leveraging unsupervised learning techniques to enhance the accuracy and efficiency for object detection

REFERENCES

[1] Pramanik, A., Harshvardhan, Djeddi, C, Sarkar, S., & Maiti, J. (2020). Region proposal and object detection using HoG-based CNN feature map. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy(ICDABI)(pp.1–5).IEEE. https://doi.org/10.1109/ICDABI51230.2020.9325708

[2] Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., & Maningo, J. M. Z. (2018). Object detection using convolutional neural networks. In 2018 IEEE Region 10 Conference(TENCON)(pp.2023–2027).IEEE. https://doi.org/10.1109/TENCON.2018.8650517

[3] Gao, C., & Li, J. (2014). Research on the surface defect detection technology of porcelain insulators. Advances in Mechanical Engineering, 6, 252803. https://doi.org/10.1155/2014/252803

[4] Nguyen, T., Park, E. A., Han, J., Park, D. C., & Min, S. Y. (2014). Object detection using scale invariant feature transform. In J. S. Pan, P. Krömer, & V. Snášel (Eds.), Genetic and evolutionary computing (Vol. 238, pp. 65–72). Springer. https://doi.org/10.1007/978-3-319-01796-9_7

[5] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[6] Ng, H.-F. (2006). Automatic thresholding for defect detection. Pattern Recognition Letters, 27(14), 1644–1649. https://doi.org/10.1016/j.patrec.2006.02.013

[7] Li, Y., Huang, H., Xie, Q., Yao, L., & Chen, Q. (2018). Research on a surface defect detection algorithm based on MobileNet-SSD. Applied Sciences, 8(9), 1678. https://doi.org/10.3390/app8091678

[8] Patel, S. (2020). Object detection with convolutional neural networks. In Proceedings/Book Title, Chapter 52. Springer. https://doi.org/10.1007/978-981-15-7106-052

[9] Muralidharan, R., & Chandrasekar, C. (2011). Object recognition using SVM-KNN based on geometric moment invariant. International Journal of Computer Trends and Technology, 215. Available at: http://www.internationaljournalssrg.org

[10] Panda, C. (2019). Object detection and tracking using Faster R-CNN. International Journal of Recent Technology and Engineering (IJRTE), 8(3), 4894. https://doi.org/10.35940/ijrte.C5580.098319

[11] Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3212–3232. https://doi.org/10.1109/TNNLS.2019.2912315

[12] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 580–587). https://doi.org/10.1109/CVPR.2014.81

[13] Antwi-Bekoe, E., Liu, G., Ainam, J.-P., Sun, G., & Xie, X. (2021). A deep learning approach for insulator instance segmentation and defect detection. Springer-Verlag London Ltd., part of Springer Nature. https://doi.org/10.1007/springer-nature-2021

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406026|

[14] Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS), 28, 91–99.

[16] Zhang, X., Zhou, Y., & Li, W. (2018). Detection of insulator defects using an improved Fast R-CNN model. IEEE Transactions on Instrumentation and Measurement, 67(9), 2177–2189 https://doi.org/10.1109/TIM.2018.2811486.

[17] Li, Q., Han, T., & Wang, J. (2021). High-precision insulator defect detection using UAV aerial images and Mask R-CNN. Electric Power Systems Research, 197, 107263. https://doi.org/10.1016/j.epsr.2021.107263.

[18] Zhou, W., Liu, Y., & Zhang, X. (2020). Insulator defect detection based on UAV images using improved YOLOv3 model. IEEE Access, 8, 202329–202338. https://doi.org/10.1109/ACCESS.2020.3034894.

[19] Li, S., Liu, F., & Zhang, Y. (2021). Detection of power line insulator defects using YOLOv4 and UAV images. Energy Reports, 7, 1382–1390. https://doi.org/10.1016/j.egyr.2021.05.028.

[20] Wang, L., Han, T., & Zhang, Q. (2019). Insulator fault detection based on YOLO with aerial image inspection. Electric Power Systems Research, 174, 105878. https://doi.org/10.1016/j.epsr.2019.105878.

[21] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.

[22] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European Conference on Computer Vision (ECCV) (pp. 213–229). https://doi.org/10.1007/978-3-030-58452-8_13.

[23] Lin, T.-Y., Dollar, P., Girshick, R., He, K., & Hariharan, B. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2117–2125).

[24] Tan, M., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10781–10790).

[25] He, Z., Yang, W., Liu, Y., Zheng, A., Liu, J., Lou, T., & Zhang, J. (2024). Insulator defect detection based on YOLOv8s-SwinT. Information, 15(4), 206. https://doi.org/10.3390/info15040206

[26] Ning, P., Jin, J., Xu, Y., Kong, C., Zhang, C., Tang, D., Huang, J., & Xu, Z. (2024). Enhanced detection of glass insulator defects using improved generative modeling and Faster R-CNN. Procedia CIRP, 129, 31–36. https://doi.org/10.1016/j.procir.2024.10.007

[27] Detection of insulator defects with improved ResNeSt and region proposal network. (2020).IEEEAccess, 8, 184841–184850. https://doi.org/10.1109/ACCESS.2020.3029857

[28] Zhang, Y., Li, J., Fu, W., Ma, J., & Wang, G. (2023). A lightweight YOLOv7 insulator defect detection algorithm based on DSC-SE. PLoS ONE, 18(12), e0289162. https://doi.org/10.1371/journal.pone.0289162

[29] Qiang, H., Tao, Z., Ye, B., Yang, R., & Xu, W. (2023). Transmission line fault detection and classification based on improved YOLOv8s. Electronics, 12(21), 4537. https://doi.org/10.3390/electronics12214537

[30] Jiang, T., Hou, X., & Wang, M. (2024). Insulator defect detection based on the CDDCR–YOLOv8 algorithm. International Journal of Computational Intelligence Systems, 17(1), 245. https://doi.org/10.1007/s44196-024-00654-x

[31] Tang, J., & Peng, H. (2024). Insulator defect detection algorithm based on YOLOv8-ISDD. Journal of Physics: Conference Series, 2849(1), 012087. https://doi.org/10.1088/1742-6596/2849/1/012087

[32] Li, N., Zeng, D., Zhao, Y., Wang, J., & Wang, B. (2024). A convolutional neural network-based defect recognition method for power insulator. Journal of Electrical Engineering & Technology, 12(4), 1201–1210. https://doi.org/10.1109/jeet.2024.04018

[33] Fathima, A. A., Vaidehi, V., Rastogi, N., Kumar, R. M., & Sivasubramaniam, S. (2013). Performance analysis of multiclass object detection using SVM classifier. In 2013 International Conference on Recent Trends in Information Technology (ICRTIT) (pp. 157–162). IEEE. https://doi.org/10.1109/ICRTIT.2013.6844198

[34] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR) (Oral presentation, pp. 1–16).

[35] Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929

[36] Liu, Z., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9992–10002). https://doi.org/10.1109/ICCV48922.2021.00986

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details