Dynamic Object Detection Revolution: Deep Learning with Attention, Semantic Understanding, and Instance Segmentation for Real-World Precision

Karimunnisa.shaik[1], Dr. Dyuti Banerjee[2], Dr. R. Sabin Begum[3], Narne Srikanth[4], Jonnadula Narasimharao[5], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[6], Dr.E.Thenmozhi[7]

Assistant Professor, Department of Information Technology, Marri Laxman Reddy Institute of Technology and Management, Dundigal, Hyderabad, India-500043[1]

Assistant Professor, Department of Artificial Intelligence & Data Science (AI&DS), Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur District, Andhra Pradesh, India-522302[2]

Assistant Professor, Department of Computer Applications, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India[3]

Assistant Professor, Department of CSE (AI & ML), RVR & JC College of Engineering Andhra Pradesh, India[4]

Associate Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India – 501401[5]

Faculty of Informatics and Computing, UniSZA University, Malaysia[6]

Associate Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India[7]

***Abstract*—Semantic and instance segmentation are critical goals that span a wide range of applications, from autonomous driving to object recognition in different fields. The existing approaches have limitations, especially when it comes to the difficult task of identifying and detecting minute things in intricate real-world situations. This work presents a novel method that uses a hybrid deep learning architecture with the Python programming language to smoothly combine semantic and instance segmentation. The suggested approach takes care of the pressing necessity in challenging real-world settings for accurate localization and fine-grained object detection. By combining the strengths of a Convolutional Neural Network (CNN) with a Bidirectional Long Short-Term Memory Network (BiLSTM), the hybrid model effectively achieves semantic segmentation by using sequential input and spatial information. A parallel attention method is smoothly included into the segmentation process to further improve the model's capabilities and enable the recognition of important object attributes. This study highlights the difficulties caused by changing environmental elements, highlighting the need for precise object location and understanding in addition to the complexities of fine-grained object detection. The suggested approach has an outstanding accuracy rate of 99.66%, outperforming existing approaches by 25.22%. This significant increase highlights the benefits that the hybrid design has over individual techniques and shows how effective it is at resolving issues that arise in dynamic real-world circumstances. The research highlights the importance of attention processes in deep learning and demonstrates how they might improve the specificity and accuracy of object detection and localization in intricate real-world scenarios. The improved performance of the suggested methodology is with well-known techniques like RCNN, CNN, and DNN, reaffirming its status as a reliable means of developing object localization and recognition in difficult situations.**

Keywords*—Semantic segmentation; instance segmentation; convolutional neural network; bidirectional long short-term memory; attention mechanism

I. INTRODUCTION

Fine-grained object identification in visual computing is an important topic for addressing real-world issues by concentrating on minute distinctions and subtleties among comparable things, prioritizing precision in technologies such as manufacturing, self-driving vehicles, healthcare imaging, and wildlife preservation [1]. The phrase "challenging real-world environments" in this sense refers to a wide range of elements, such as obstructions, various lighting situations, various views, and abundantly generated scenery. Gaining an improved understanding of the minute differences that set items belonging to the identical grouping apart is the main objective of fine-grained object recognition. For example, fine-grained recognition of objects goes beyond standard object recognition to distinguish between certain dog kinds or cat species. standard object recognition could distinguish among a dog and a cat [2]. Innovative strategies are required to uncover small features and patterns that are invisible to the human eye, which frequently combine learning algorithms, deep neural networks, and characteristic engineering to achieve efficient analysis. Real-world scenarios with continually changing and unexpected characteristics provide considerable challenges to a seamless object recognition infrastructure. obstacles, visual obstacles generated by extra objects or features, and variations in light, such as shadows and views, can all complicate the identification process [3]. Recognition algorithms must be adaptable to variations in size, movement, and viewpoint caused by real-world views and complicated backgrounds. Sophisticated algorithms are required to discriminate between useful items and insignificant visual characteristics in busy and patterned

surroundings, providing smooth object detection in difficult conditions [4]. A comprehensive strategy comprising innovative algorithm creation, different training data sets, and in-depth knowledge of individual applications is required to increase the robustness and dependability of smooth object recognition algorithms for real-world applications.

Robots and AI systems use instance segmentation and semantic segmentation as core computer vision operations, allowing them to analyze visual input in a variety of circumstances and translate pixel-level data into useful information [5]. Semantic segmentation is an important approach for scene evaluation and object recognition that assigns pixel-level classification to certain sections of a picture. This allows computer systems to recognize object boundaries and grasp how the world is organized. Semantic segmentation improves its efficiency by distinguishing between object categories and individual occurrences, making it the cornerstone of scene interpretation [6]. In picture segmentation, each structure has a precise object mask, which is critical for robotics and AI systems to reliably discriminate and recognize instances of the same item category. This level of granularity is critical in complicated applications such as self-driving cars and healthcare imaging, where detecting unique instances in crowded settings is important. It is difficult to distinguish delicate elements from clutter limits segmentation algorithms in real-world scenarios. Conventional approaches fail to detect small-scale items accurately, while real-world difficulties like as obstructions, accessibility limits, and changing illumination conditions hamper object identification and localization even more. These issues demand unique tactics that go beyond existing approaches to provide dependable and seamless object assessment in difficult real-world scenarios. [7].

This paper introduces a hybrid deep learning architecture that combines semantic and instance segmentation capabilities to improve real-world segmentation algorithms. The framework employs attention processes to outperform existing strategies in tasks demanding precise object placement and identification. It integrates semantics and instance-level segmentation into a single framework, fixing specific flaws while maximizing benefits. The method employs numerous decoding paths and shared encoders to extract important information from input images, allowing the system to comprehend the image's overall meaning and design. This fundamental understanding facilitates scene design and relationships [8]. The framework employs a shared encoder to decrease processing needs, increase efficiency, and ensure a thorough grasp of input data. Strategic decoding branching switches the encoder, resulting in more accurate and consistent segmentation results while optimizing resource utilization and data transmission. The hybrid architecture uses attention processes to improve system efficiency, allowing it to concentrate on critical data locations, especially in complicated and crowded situations where fine-grained features are likely to be covered [9]. The system employs attention methods for segmentation, ensuring that computer resources are used efficiently and that meaningful instances are distinguished from background noise. The combination of attention processes and the shared encoder increases flexibility by handling occlusions, angles of view, and illumination conditions. This integrated technique encourages a more detailed comprehension of settings, resulting in considerable improvements in accurate item location and identification in demanding real-world scenarios [10]. The hybrid deep learning framework enhances semantics and instance segmentation through the use of twin decoding approaches. It separates pictures into relevant segments for semantic segmentation by generating pixel-wise categorization projections, allowing for comprehensive image interpretation by identifying significant portions of the image [11]. The second decoder is particularly built for instance segmentation, resulting in exact object masking to discriminate between distinct instances of objects. This method improves higher-level semantic comprehension and enables fluid localization, resulting in a more complete and accurate contextual understanding.

The framework incorporates attention processes into decoding operations to improve its efficacy in demanding settings. These tactics guarantee that just the most significant parts of a picture are analysed, allowing the algorithm to focus on areas critical for object localization and recognition. This enhances its capacity to handle complicated things, respond fast to visual cues, and function effectively in tough conditions [12]. The hybrid architecture's performance was evaluated using a variety of real-world datasets, such as interior settings, animal photos, and metropolitan landscapes. The system outperformed advanced approaches in semantics and instance segmentation, proving its capacity to handle accurate item localization tasks in demanding environments such as interior settings and metropolitan landscapes [13]. The hybrid architecture, as a flexible solution, exhibits its capacity to excel across a number of domains, providing potential paths for improvements in fine-grained visual processing and supporting applications ranging from autonomous cars to medical imaging.

The key contributions of the article is,

- The study introduces a novel pre-processing step using Gaussian functions, enhancing feature extraction and smoothing in input data, contributing to improved overall model performance.
- The research leverages a fusion of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory networks (BiLSTM) for semantic segmentation, enabling the model to capture both spatial and temporal dependencies in the data, resulting in more accurate and context-aware segmentation.
- The incorporation of attention mechanisms in instance segmentation significantly refines object delineation. The attention mechanism ensures that the model focuses on relevant regions, enhancing the precision and efficiency of instance segmentation in complex scenes.

- The study offers a nuanced view of the model's capabilities by introducing a thorough performance evaluation technique that takes into account variables including segmentation accuracy, computing efficiency, and resilience.

This article's remainder is organized as follows: In Section II, a summary of related research is provided. Section III presents the problem statement. The suggested approach's methodology and architecture are explained in Section IV of the article. The findings and subsequent discussion are covered in Section V. The conclusion is covered in Section VI.

II. Related Works

Automatic recognition of objects in 3D spaces is essential to working zone security, including ensuring adherence to building codes and averting accidents and fatalities at workplaces [14]. However, a number of difficulties, including correct three-dimensional object comprehension because of size changes and a lack of indicators in the three-dimensional environment, outstanding identification, exceptional segmentation of instances, and insufficient technical object databases with masking present significant challenges. These difficulties affect conventional manual techniques. The main discovery is to calculate pseudo-light recognition and reaching point clouds for three-dimensional object recognition using two-dimensional recognition of objects, segmentation of instances, and cameras perception. On the contrary hand, an upgraded cascading masks R-CNN is used to identify boundaries and masking for every two-dimensional object, while an enhanced characteristic pyramids system is presented for obtaining additional smooth object characteristics. An additional object classes with the boundaries and masking is introduced, and the AIM database for massive machinery identification is expanded. On the contrary hand, using deep learning, autonomous camera parameters estimation, a vision-based approach, and a spatial filtering, it is possible to retrieve pseudo-LiDAR point’s clouds of objects generated by boundary boxes and masking from a monochromatic image. Numerous tests and evaluations reveal that the recently developed approach can recognize three-dimensional items and autonomously assess the security working areas. On the AIM information set, the suggested object recognition system produced the most advanced outcomes, and for the enhanced information set. The fresh framework will act as a starting point for studies regarding three-dimensional object recognition for additional three-dimensional positions.

The study extends Trans10K-v1, the initial significant database for translucent item differentiation, by providing an additional, smooth database known as Trans10K-v2 [15]. The novel dataset has a number of enticing advantages over Trans10K-v1, which only contains two constrained classes. (1) It is better suited for practical use since it comprises eleven smooth classes of translucent items that are frequently seen in the average household surroundings. (2) Compared to its predecessor, Trans10K-v2 presents additional difficulties for currently sophisticated segmentation techniques. Additionally, the Trans2Seg pipelines, unique transformer-based segmented pipelines, are suggested. Initially Trans2Seg's transformers encoders, which offers an overall responsive field as opposed to CNN's localized one, outperforms standard CNN systems in terms of performance. Subsequently, the study builds a collection of accessible designs as the question parameters of Trans2Seg's transformers decoder, where every instance acquires the statistical information of a particular group in the entire data set by treating semantic segmentation as an issue of dictionaries. Researchers compare Trans2Seg with a variety of than twenty current semantically segmented techniques, revealing that it greatly exceeds all CNN-based techniques and potentially solving the problem of translucent object segmentation.

For the creation of a successful computerized diagnostic framework, the identification and fragmentation of the new coronavirus infection of 2019 abnormalities using CT images are extremely important [16]. One of the finest options for creating such an instance is deep learning. However, a number of issues, involving as information variation, a wide range in the dimensions and form of the inflammation, lesions imbalances and an absence of annotations, restrict the effectiveness of DL techniques. In order to overcome these difficulties, a unique multitasking regression networks for categorizing COVID-19 lesions is put forward in the present research. The model's designation is MT-nCov-Net. The lesions identification is formulated as a multitasking structure extrapolation issue, allowing for the sharing of low-, medium-, and excellent-quality information across multiple assignments. In order to effectively acquire tiny and substantial lesion characteristics while minimizing the semantic disparity between various scale visualizations, a multiscale characteristic learning modules is introduced. This component captures the multiscale knowledge of semantics. Also included is a smooth lesions identification module that employs an adaptable dual-attention technique to find infected regions. The inflammatory regeneration modules then segment the infected regions using the obtained position mapping and the merged multiscale depictions. By reducing the COVID-19 area's form, MT-nCov-Net can properly fragment the inflammation through absorbing all of its features. MT-nCov-Net is empirically assessed on two open multisource information sets, and the results support its general efficacy above the state-of-the-art methodologies and show how well it works to solve the issues with COVID-19 diagnostics.

Due to the intricacy of plant images, segmenting plants is a difficult automated vision problem [17]. The study has to do increasingly more challenging activities in order to tackle various real-world issues. Instead of looking at the entire plant, the study must make distinctions between plant sections. The lack of information with thorough annotations is the main obstacle to multi-part segmentation. Actively annotating databases at the object component levels takes a lot of effort and money. The study suggests using pseudo-annotation with inadequately supervised training. In the article, researchers examine the minimally supervised training techniques currently in use and offer a productive pipeline for agrarian applications. It is made to deal with close object overlaps. For the plant component example and the entire plant scenario, the pipeline outperforms the starting point solutions by twenty-three percent and forty percent, respectively. To improve

simulation effectiveness, researchers also use instance-level enhancement. The goal of the method is to create a weakened segmentation masking that can be used to trim items from the source images and paste them onto fresh backdrops while the participant is being trained. On object component segmentation operations, the strategy gives us a fifty-five percent mAP gain over the initial state, and on entire plant segmentation operations, a seventy-two percent gain.

Recent developments in navigational autonomy have shown a greater preference for computation vision over conventional methods [18]. The majority of the places are built with individual movement in mind, which is how this works. They are therefore chocking full of visual indicators. In this way, the capacity to recognize objects visually is crucial for self-driving cars to prevent impediments while interacting with the outside environment. It is laborious and costly to gather information employing unmanned aerial vehicles that are capable of operating in the real environment. A database consisting of areas and conversations constitutes one of IT businesses' greatest resources as a result. Adopting an image-realistic three-dimensional simulation as the source of information is one way to address this issue. It is feasible to obtain a substantial quantity of information with this asset. Therefore, utilizing images from a frontend UAV camera moving through a three-dimensional simulator, the present study builds a collection of images for example segmentation. The Mask-RCNN, a cutting-edge deep learning approach, is used in the present research. The framework estimates per-pixel segmentation of instances from an image input. According to empirical findings, Mask RCNN performs better in the data we provide when improving the algorithm generated from the COCO dataset. Additionally, the intriguing findings in real-life information provide the suggested technique a solid generalization potential.

The inconsistent and fragmented nature of points cloud information in a non-Euclidean environment makes it difficult to fully employ smooth semantic properties [19]. A max pooling procedure is frequently employed to draw attention to particularly significant characteristics in the immediate area in attempting to illustrate the local characteristic for every centering point that is beneficial for improved contextual understanding. The max pooling method, however, ignores any additional geometrical local connections between every central location and its related neighborhood. In order to do this, the focused attention method shows promise in preserving node representations on graph-based information by paying consideration to every node in its immediate vicinity. Using stacking MLP units and a unique neural network called GA Point Net, the study offers a new method for analyzing point clouds. GA Point Net can acquire localized geometrical descriptions. To effectively utilize local characteristics, the study emphasizes various focus weights on every focal point's neighborhood. To completely retrieve localized geometrical patterns and improve the system's resilience, the study additionally mixes attentive characteristics with the regional identity characteristics produced by the focus grouping. The suggested GA Point Net structure performs at the cutting edge across the form categorization and segmentation operations when evaluated on a variety of datasets used as benchmarks.

To be more precise, the characteristics with multiple scales are combined top-down to blend specifics and broad semantics to improve tiny item detection [20]. A pyramidal tiered attention mechanism made up of channels and spatial focus is created throughout the fusion of multi-scale information in order to see the item. Additionally, enhancing self-paced learning is used to direct the model to study challenging data. Two real-world datasets, an AMMW dataset and a publicly accessible PMMW dataset, are used for validating the technique that is suggested. The results of experiments show that the suggested strategy is preferable due to its versatility in identifying various devices concealed inside clothing while remaining harmless; the AMMW scanner has become a common tool for checking the safety of people in public settings in recent years. Nevertheless, due to intrinsic image noise, unknown item kind, and ambiguous status, it is very difficult to identify all concealed objects autonomously and precisely. Recent improvements in concealed object identification have been made by various current algorithms, particularly those that utilize deep learning. These techniques are effective for finding a few specific types of huge things, but they are ineffective for finding dim or imperfect hard objects. The state-of-the-art approaches are provided in this paper as a hidden finding of objects model with SPFAFN to handle this problem. SPFAFN obtains improved results on the two datasets with Maximum Precision.

The importance of object detection for systems that drive autonomously is rising [21]. Nevertheless, the use of existing object detectors to autonomously drive is constrained by their low accuracy or low inference ability. By integrating dilated convolutions and a SAM into the YOLOv3 design, a quick and precise object detector known as SA-YOLOv3 is suggested in this study. In order to enhance the accuracy of detection, the loss functional determined by GIoU and focused loss is rebuilt. With immediate inference, the suggested SA-YOLOv3 enhances YOLOv3 by 2.58 mAP and 2.63 mAP on the KITTI and BDD100K standards, respectively. Its improved compromise in terms of quickness and - accuracy in comparison to other cutting-edge detectors suggests that it is appropriate for use in self-driving vehicle applications. It is believed that this strategy, which integrates YOLOv3 with an attention mechanism for the first time will serve as a model for upcoming studies on autonomous vehicles.

To differentiate objectives from inferior categorization is the goal of smooth visual categorization [22]. It is thought to be a very challenging assignment since smooth images naturally exhibit significant inter-class variations and tiny intra-class variability. The majority of current methods employ CNN-based systems as characteristic extractors, which results in the generated exclusive areas including the majority of the object's components and missing to identify the truly crucial components. The perception transformers, which employ a mechanism for focus to gather broad context-relevant data to create a distant reliance on the desired object and then extracts more potent characteristics, has subsequently proven its effectiveness on a variety of imagine activities. The ViT approach might function inadequately in the

categorization of smooth images because it continues to place greater emphasis on overall coarse-textured data than localized smooth data. The study enhances the ViT framework and develops a concentration accumulating transformers to more effectively catch small variations across images. To improve communication between every transformer layering, the study specifically suggests an essential consideration accumulator. Additionally, the study provides an original data entropy selection to direct the algorithm in accurately obtaining discriminatory portions of the image. Numerous tests demonstrate that the suggested model architecture is capable of operating at an innovative modern facilities level on a number of widely used databases.

It's crucial and difficult to recognize facial expressions employing a DCNN [23]. Although significant efforts have been performed to improve FER accuracy using DCNN, earlier experiments have not yet been adequately generalized for use in practical situations. Conventional FER research is mostly restricted to regulated lab-posed fronted facing images, which do not encounter the difficulties associated with motion disintegrate head causes, obstructions, face distortions, and illumination under unregulated circumstances. The study suggested a SqueezExpNet architecture for extremely efficient FER systems that can tolerate fluctuations in the environment. It can benefit from global as well as local face data. The network was split into two distinct phases: a geometric concentration phase, which uses a SqueezeNet-like structure to collect localized highlighting data, and a geographic texturing phase, which consists of numerous compressed and extended levels to take use of the highest-level broad characteristics. To highlight significant localized facial areas, the study specifically developed the weighted masking of three-dimensional facial characteristics and employed element-wise combination with a geographical characteristic in the initial step. The subsequent phase of the network is then fed with the facial geographical imagine and its enhancements. To help overcome the unpredictability, a network of recurrent neural networks was created to cooperate with the emphasized data from two phases instead of only employing the SoftMax function, much like a classification system. The three top expressions databases were used in investigations encompassing simple and complex FER objectives. The technique produced cutting-edge findings and surpassed the current DCNN algorithms. Instantaneous FER could uncover possible applications in monitoring, well-being, and feedback mechanisms according to the created construction, chosen investigation approach, and published outcomes.

The thorough literature review explores a wide range of computer vision and deep learning subjects. By balancing camera perception, instance segmentation, and two-dimensional object recognition, the first study presents a novel hybrid deep learning architecture that adeptly tackles real-world complexities in three-dimensional object recognition. The results are promising and have significant implications for the identification of machinery and the development of autonomous driving technologies. Moreover, it presents Trans2Seg, a novel transformer-based segmentation pipeline that significantly outperforms the state-of-the-art CNN-based techniques. To enhance and broaden the research landscape in this dynamic and ever-evolving domain, the study complements these breakthroughs by creating Trans10K-v2, a novel dataset featuring a variety of translucent object classes.

III. Problem Statement

Deep convolutional neural networks (DCNNs) must operate accurately in uncontrolled, real-world contexts, making face emotion recognition an important yet difficult challenge. Advances in facial expression recognition (FER) accuracy have mostly concentrated on controlled lab environments, ignoring problems like head obstacles, motion disintegration, face distortions, and changing lighting. Effective generalization in such real-world scenarios is not possible with the current DCNN-based models [23]. In order to overcome these shortcomings, the SqueezExpNet architecture is presented in this paper. It is intended for very effective FER systems that can adapt to changes in the environment and make use of both local and global face data. The network is divided into two phases: a geographic texturing phase and a geometric concentration phase. For enhanced performance, a network of recurrent neural networks and weighted masking of three-dimensional face characteristics are included. The suggested technology outperforms existing DCNN algorithms and has potential uses in feedback systems, monitoring, and wellbeing applications.

IV. Proposed Hybrid Learning Architecture with Attention Mechanisms

The approach entails putting forth a unique architecture that combines instance and semantic segmentation to improve precise item localization and recognition in difficult real-world circumstances. The integration recognizes that although semantic segmentation delivers high-level contextual comprehension, instance segmentation offers detailed instance-level information. The CNN-BiLSTM for semantic segmentation is a novel addition that makes use of the spatial context and sequence data gathered by BiLSTMs and CNNs. In particular in complicated situations, the addition of attention processes, such as segmentation, assists in focusing on key visual areas for exact instance classification. In addition to demonstrating the architecture's superiority over stand-alone approaches and emphasizing the importance of attention processes in boosting instance segmentation's accuracy and reliability, the research makes sure to provide a thorough review of the architecture through quantitative and qualitative assessments. It is depicted in Fig. 1.

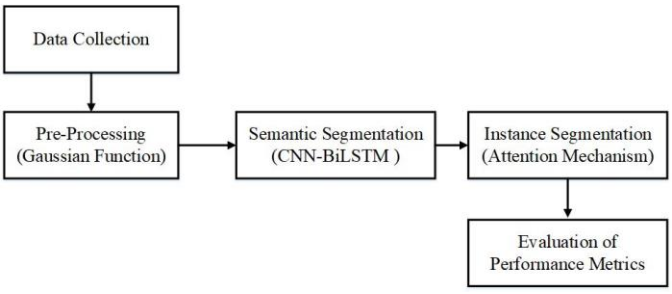

Fig. 1. Proposed methodology.

A. Data Collection

The VEDAI dataset, which supports the growth of fine-grained vehicle recognition algorithms in remote sensing images, is a dataset for fine-grained vehicle identification. The total number of images and occurrences in VEDAI is 1210, and each image has a 1024 by 1024 pixel resolution. The order of occurrences in this collection can be considered sparse, as can be observed by the quantity of instances and images [25].

B. Pre-Processing using Gaussian Function

When preprocessing an image using a Gaussian function, a Gaussian filter is applied to it to assist smooth it out and remove noise. The kernel or mask of the filtering procedure is the bell-shaped Gaussian function, a mathematical function. The filter operates using a Gaussian kernel, and each pixel is assigned a weight based on how near to its neighbors it is. The weights are computed using the values of the Gaussian function at each location along the kernel. By changing the Gaussian filter's parameters, such as the kernel size and standard deviation, the amount of image smoothing may be changed. While bigger kernel sizes and greater standard deviations provide more comprehensive smoothing, smaller kernel sizes and lower standard deviations maintain finer features. The final preprocessed image has a smoother overall look, less noise, and fewer high-frequency features. Gaussian filtering, a preprocessing technique frequently used in image processing, is particularly useful for tasks like demising, feature extraction, and increasing the quality of images. Because the Gaussian filter can blur and smooth images well, it is a crucial tool in computer vision and image processing applications. This filter, which gets its name from the Gaussian function, is frequently used to improve images and reduce noise. Through the use of a Gaussian kernel to convolve the input picture, the filter reduces high-frequency noise while maintaining important image characteristics. Its natural smoothing ability helps to lessen pixel-level differences, producing an output that is more aesthetically pleasant and visually cohesive. The blurring effect of the Gaussian filter also helps to emphasize important structures while minimizing unimportant details, which makes it useful in applications like edge detection and feature extraction. Because of its adaptability, the filter may be used in a wide range of domains, such as object identification, robotics, and medical imaging. As such, it is an essential tool for preprocessing and picture refinement in a variety of settings. The Gaussian function equation is shown below in Eq. (1).

$$H(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{x^2}{2\pi\sigma^2}} \quad (1)$$

where the standard deviation of the distribution is denoted as σ. The distribution is assumed to have a mean of 0.

C. Hybrid CNN-BiLSTM for Semantic Segmentation

Considering it effectively handles image interpretation issues by combining the capabilities of CNN and BiLSTM, the Hybrid CNN-BiLSTM architecture is essential to semantic segmentation. Through the combination of the sequential information-taking capacity of the BiLSTM and the spatial feature-capturing capability of the CNN, this innovative approach allows the model to interpret both contextual and spatial subtleties in the image data. These two components work together in the Hybrid CNN-BiLSTM architecture to significantly increase semantic segmentation accuracy and precision, particularly where precise localization and fine-grained object recognition are crucial. The resilience and reliability of semantic segmentation tasks are enhanced by this design, which makes it possible for the model to function effectively in dynamic and complicated real-world environments. This makes it vital for a wide range of applications, including autonomous driving, computer vision, and medical imaging.

A convolutional neural network (CNN) is used in this step of feature extraction and is a crucial step in deep learning for pattern and recognition of image applications. CNN is developed in order to autonomously determine organizational and discriminating qualities from the original input images. In a series of convolutional layers, tiny filters are organized over the input image to collect regional trends and attributes. These characteristics capture the edges, the surfaces, and shapes which make up the visual environment. Then, the feature maps are down sampled by pooling layers, which lessens the computational cost while maintaining important data. Fully connected layers accept the results of the learnt attributes and put them into effect future classification or similar activities.

CNNs have shown outstanding results in a range of applications involving computer vision, demonstrating modern performance in problems including object detection, image segmentation, and analysis of images in healthcare. This is because of their ability to voluntarily pick up and remember important aspects from visuals. By merging information from different methods, the CNN technique maximizes the benefits of each modality while compensating for its limitations and providing a more complete knowledge. Recurrent neural networks of the Bidirectional Long Short-Term Memory (BiLSTM) variety handle sequential information while simultaneously considering into account both past as well as future circumstances. It has both forward and backward LSTM components that create hidden states and collect data from components that come before and after them in the sequence. When combined with convolutional neural networks (CNNs), BiLSTMs can improve pixel-wise image categorization in applications like semantic segmentation. Bidirectional connections are utilized into consideration by BiLSTMs to enhance the framework's comprehension of intricate spatial and contextual interactions, producing more precise and context-sensitive semantic segmentation that benefits applications related to computer vision. The diagrammatic representation of the proposed hybrid CNN-BiLSTM method is depicted in Fig. 2 and the equations of CNN-BiLSTM are depicted in Eq. (2) to Eq. (7) below:

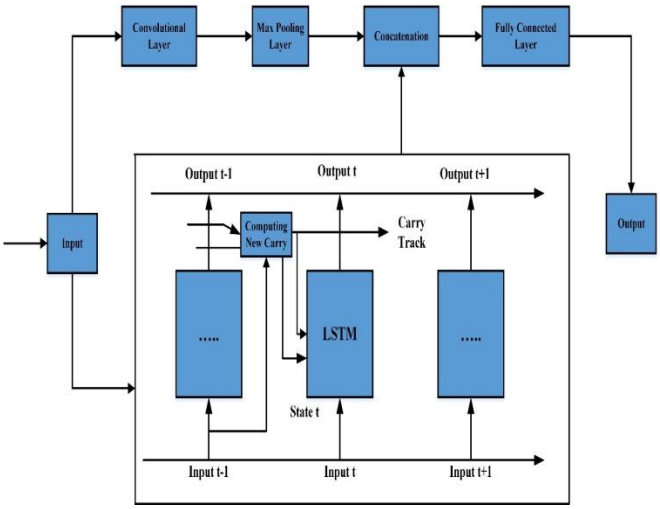

Fig. 2. CNN-BiLSTM framework.

$$g_v = \sigma(N_g y_d + K_g g_{v-1} + d_g) \quad (2)$$

$$h_v = tan_g(N_h y_v + K_h g_{v-1} + d_h) \quad (3)$$

$$j_v = \sigma(N_j y_v + K_q g_{v-1} + d_j) \quad (4)$$

$$p_v = \sigma(N_i y_v + K_p g_{v-1} + d_p) \quad (5)$$

The current qv state can be calculated by (6).

$$q_v = g_v \times q_{v-1} + j_v \times h_v \quad (6)$$

$$x_v = g_v = p_v \times tan_g(q_v) \quad (7)$$

Here, Kg, Kh, Kq, Kp signifies the weight matrices of the previous short-term state gv-1. Ng, Nh, Nj, Ni signifies the weight matrices of the current input state yd, dh, dg, dj, and dp are labelled as the bias terms, qv-1 characterizes the preceding long-term state.

D. Attention Mechanism for Instance Segmentation

The Attention Mechanism is a fundamental component of instance segmentation and serves as a basis for enhancing and optimizing object recognition in intricate visual datasets. The primary function of this method is to enhance the accuracy and precision of instance segmentation by allowing the model to recognize and concentrate on certain object features and regions. Even in congested or densely populated environments, the model can efficiently separate and isolate individual object instances because to its dynamic capacity to balance the value of different components in the visual input. This adaptable approach proves its worth in a multitude of domains, ranging from robotics and object recognition to the intricate realm of medical imaging, where it enhances instance segmentation accuracy and consistency. This paves the door for innovative developments in these domains by laying the foundation for a richer comprehension of complicated visual data and the things it includes.

The attention model improves the CNN by maintaining its context-relevant characteristics. In the prior-based model, each block's attributes are integrated into the ones from the layer below it. This method equally weights each attribute acquired from the previous CNN blocks. To learn precise feature values, important characteristics from the previous blocks must be given a high weight relative to other features. In order to facilitate the acquisition and selection of noteworthy qualities from previous blocks, a mechanism tracking attention was consequently introduced to the CNN architecture. This model generates an attention mask that equalizes the relative importance of spatial characteristics on that feature map. Leveraging an attention mechanism throughout blocks, the CNN architecture generates a weighted function for simulating activations from the prior blocks. The connections from the previous blocks that were skipped were then weighed throughout the depth axis for every single pixel in that layer's spatial range [26].

Two operational channels, H(x) and S'(x) are used to guide the layer of convolution to produce 'x' output in both the first and subsequent blocks. H(x) shows the set of methods that were implemented to take the input value "x" and just feed it forward to the block after it. The group of techniques collectively referred to as S'(x) has weighting with attention and skipping the 'x' through convolutional and maximum-pooling layers. The balanced summation is used to obtain the outputs G(x) from the CNN block and is shown in Eq. (8).

$$G(X) = H(X) + S'(X) \quad (8)$$

Eq. (9) is used to determine the functional route S'(x).

$$S'(X) = S'(X) * \varphi \quad (9)$$

where S(x)'s spatial dimensions and the attention weight matrix's parameters are equivalent. The appropriate cross-section of S(x) is multiplied point-wise (broadcast throughout the depths) by the attention matrix's weights "." CNN can incorporate the inputs from the current moment and its results from the previous instant to automatically allocate the weighting for each element of the network as a whole by including an attention method. Important image details may be focused on to improve the classification's precision and adaptability. Based on this, CNN's attention concept is added to form the Attention-CNN model. Pay attention carefully: To identify and classify the framework and airborne particulates in SEM images, CNN is used. The attention-CNN architecture consists of the input, a convolution, attention to detail, full connection, and output layers. The layer that receives input is made up of four nodes, which are the labelled images of four separate kinds of particles, as the input data is a pictorial representation of four distinct particle types. Each of the convolution layers of the four phases that together make up a single layer of convolution is followed by a layer of attention in order to successfully accomplish weight transportation. The dimensions of the convolution kernels are 8x3x3, 16x5, 32x3, and 32x3 for each layer, accordingly. A pooling layer connects the first and last convolution layers. The final result of the layer used for convolution serves as the layer's input, and the total number of nodes in the full connections layer is set to 64. The maximum number of the output layer nodes is 4, and the output layer classifies the smallest particles into four groups.

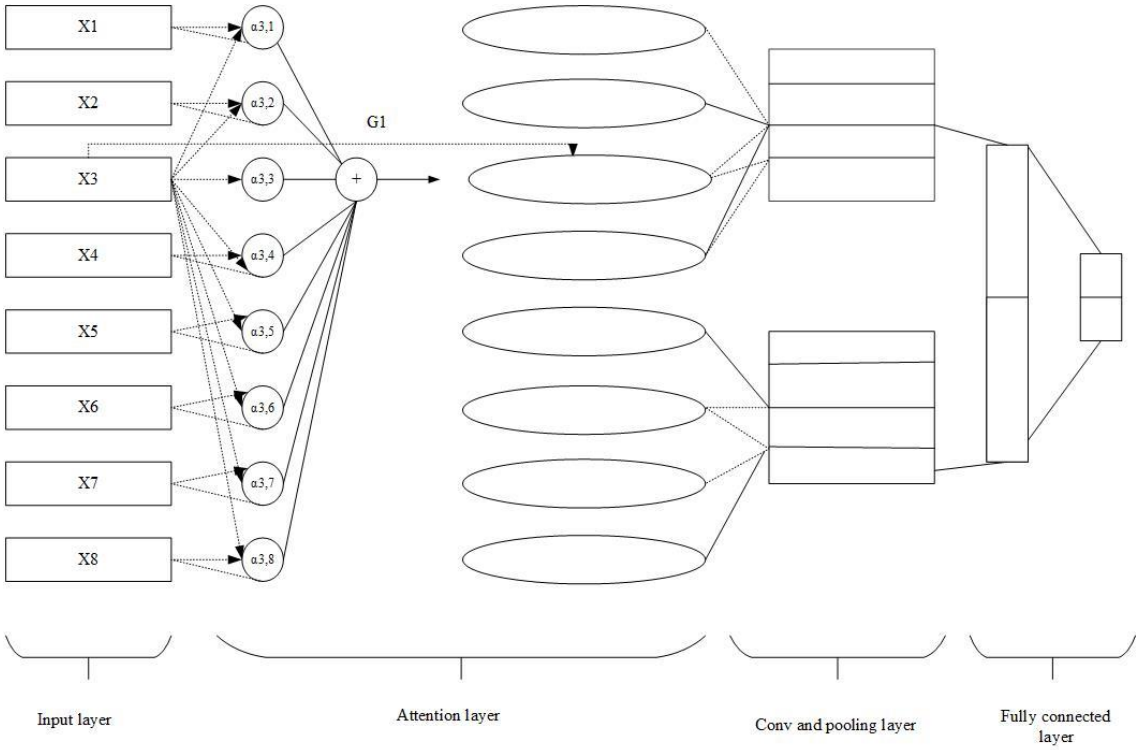

Fig. 3. Attention mechanism integration in CNN.

The CNN Attention Mechanism for Integration is shown in Fig. 3. A nonlinear relationship might be inserted amongst the different layers of an ensemble of neurons by changing the function that causes activation. The network's output no longer looks linear, which increases the network's expression and allows it to fit a wider range of patterns. The Attention-CNN model uses two activating coefficients, Relu (rectified linear unit) as well as, for its concealed and outputs components, respectively. Relu can deal with elevation dispersion throughout the element transfer process. When the value of the Relu function is greater than 0, its derivative is 1. It is simple to identify the gradient and may greatly accelerate the gradients' downward speed of convergence. Eq. (10) shows the Relu function's formulation.

$$ReLu = max(0, X) \tag{10}$$

By transferring the outputs of several neurons to the coordinates (0, 1), Softmax is able to classify data in a variety of ways. Assuming there is one, j denote the last component of an input array; the softmax value assigned to that component is determined by

$$G_j = \frac{s^j}{\sum_{l=1}^{k} s^j} \tag{11}$$

where, k stands for all of the input items. Since both the first-order and second-order instant means of the gradient are completely used by the Adam optimization approach, the forward momentum component is taken into account during the updating procedure. Adam's computation is shown in Eq. (12) to Eq. (16):

$$u_t = \alpha_1 \delta_{t-1} + (1 - \alpha_1) k_t \tag{12}$$

$$n_t = \alpha_2 \delta_{t-1} + (1 - \alpha_2) k_t^2 \tag{13}$$

$$\hat{u}_t = \frac{u_t}{1 - \beta_1^t} \tag{14}$$

$$\hat{n}_t = \frac{n_t}{1 - \delta_1^t} \tag{15}$$

$$k_{t+1} = k_t - \frac{\partial}{\sqrt{\hat{n}_t + \varepsilon}} \hat{u}_t \tag{16}$$

where u_t is the first-order moment estimates that nt is the second-order momentum term, α_1, α_2 are actually dynamic values, $k_t$is the gradient of the cost operates after t iterations, u_t is the first moment's correction value, n_t is the second moment's correction value, k_t is the model's variables, and is a small amount that can circumvent the zero denominators. In neural network [24] training, the loss of functions is used to quantify the difference between the predicted result and the actual value. In addition, the effectiveness of the computational framework is evaluated using this component as a benchmark. The cross-entropy cost functioning, which may be thought of as the loss of function for Attention-CNN in Eq. (17),

$$H = -\frac{1}{n} \sum_l l_j (\rho(a) - b) \tag{17}$$

where, y is the resultant value, b is the actual value, n is the total of the samples l is the sample, and n is the sample. The following formula is used to determine the gradient.

$$\frac{\partial b}{\partial a} = \frac{1}{n} \sum_l l_j (\omega(z) - b) \tag{18}$$

where the error between the output and the actual value is $\omega(z) - b$ [27].

V. Results and Discussions

To improve accurate item localization and recognition in challenging real-world scenarios, the approach proposes a novel architecture that combines instance and semantic segmentation. The integration acknowledges that while instance segmentation provides detailed instance-level information, semantic segmentation provides high-level contextual comprehension. A novel addition that leverages the spatial context and sequence data collected by CNNs and BiLSTMs is the CNN-BiLSTM for semantic segmentation. The addition of attention processes, like segmentation, helps to focus on important visual areas for precise instance classification, especially in complex scenarios. The research ensures that the architecture is thoroughly reviewed through quantitative and qualitative assessments, highlighting its superiority over stand-alone approaches and highlighting the role of attention processes in improving the accuracy and reliability of instance segmentation.

A. Performance Metrics

Training and Testing Accuracy: An indicator of a machine learning model's success during the training stage is training accuracy. It displays the percentage of instances (or samples) in the training dataset that were properly predicted in relation to all of the occurrences in that dataset. In other words, the degree to which the model's predictions match the actual labels for the data it was trained on is indicated by the degree of training accuracy.

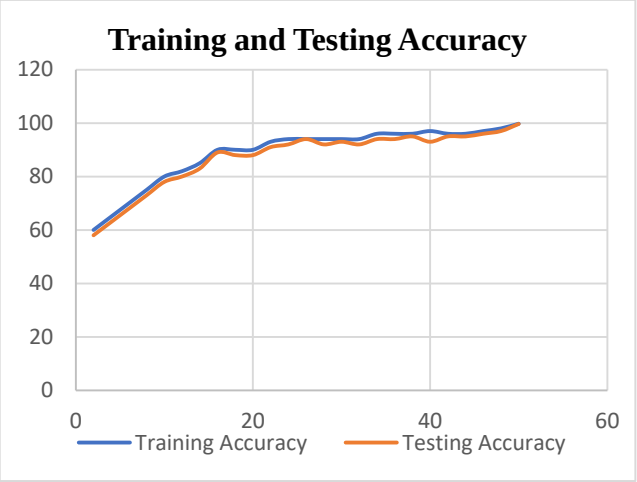

Fig. 4. Training and testing accuracy.

In this Fig. 4, accuracy refers to a performance parameter that assesses the percentage of fine-grained objects that the proposed hybrid deep learning architecture successfully recognized and localized in difficult real-world contexts. The model's excellent precision and efficacy in successfully recognizing and localizing items with complex features and difficult backgrounds are indicated by the accuracy of 99.66% that was attained.

Training and Testing Loss: A machine learning model's training loss, often referred to as the objective or cost function, is a quantifiable indicator of how well the model is doing. It shows the difference between the actual target values (ground truth) found in the training dataset and the projected values produced by the model. Making the model's predictions as near to the actual values as is practical is the main objective of training; this is to minimize this loss.

Fig. 5. Testing loss.

In Fig. 5, the gap between anticipated and ground truth values during training is measured as loss, which is shown graphically, and iteratively evolves over time for the hybrid deep learning architecture. As the model is trained, the loss curve shows how the architecture is coming together to minimize mistakes and enhance its capacity to precisely recognize and localize fine-grained objects in challenging real-world circumstances.

ROC Curve: In binary classification tasks, a graphical depiction known as the Receiver Operating Characteristic (ROC) curve is frequently used to assess how well a machine learning model is doing. As the discriminating threshold for distinguishing positive and negative occurrences is changed, it demonstrates the trade-off between the TPR, also known as sensitivity or recall, and the FPR.

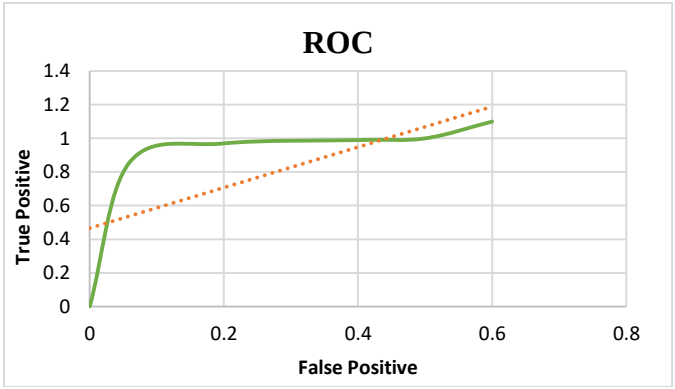

Fig. 6. ROC curve.

In Fig. 6, the best threshold for detection assessments is present; the ROC assesses the model's capacity to discriminate between the presence and absence of objects. ROC values over a certain threshold indicate a more precise object identification model. The true positive rate, also referred to as sensitivity or recall, is shown on the vertical axis, while the false positive rate is represented on the horizontal axis. The proportion of actual positive events that the classification model correctly identified is measured by the true positive

rate, while the percentage of no object data that is incorrectly classified as an object is shown by the false positive rate. These rates for various classification thresholds, which specify the point at which the model classifies a data point as an event or a non-event, are plotted to create the ROC curve.

Matthews Correlation Co-efficient (MCC): The MCC is one of the most popular measures for classification effectiveness. It is generally recognized as a reliable estimate that may be used even when class sizes vary significantly. Eq. (19) contains the formula for the Matthews correlation coefficient.

$$MCC = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P+F_P)-(T_P+F_N)(T_N+F_P)(T_N+F_N)}} \quad (19)$$

Negative Predictive Value (NPV): The subject-to-outcome ratio calculates the percentage of patients who have really poor test findings overall. The NPV measures the proportion of times each forecast was entirely wrong. Eq. (20) has the formula.

$$NPV = \frac{T_N}{T_N+F_N} \quad (20)$$

TABLE I. COMPARISON OF MCC AND NPV

Methods	NPV (%)	MCC (%)
RCNN	89.82	69.47
CNN	85.67	57.03
DNN	79.58	38.75
Hybrid DL-Attention Mechanism	**93.77**	**73.12**

Based on their NPV and MCC performance indicators, the various techniques are compared in Table I. The RCNN, CNN, DNN, and a hybrid deep learning (DL) model with an attention mechanism are the four techniques that are assessed. The comparison table's findings highlight the importance of the Hybrid DL-Attention Mechanism as a useful strategy for the assigned job, surpassing more established approaches like RCNN, CNN, and DNN in terms of NPV and MCC. The results highlight the rising significance of attention processes in deep learning by demonstrating their capacity to improve correlation metrics and prediction accuracy across a range of applications. It is shown in Fig. 7.

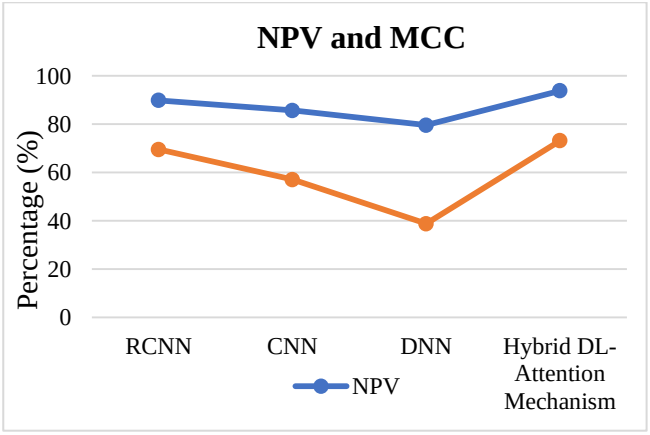

Fig. 7. Comparison of NPV and MCC.

False Positive Rate: The percentage of situations, where a favorable outcome was predicted but it did not materialize. It is illustrated in Eq. (21).

$$FPR = \frac{F_P}{T_N+F_P} \quad (21)$$

False Negative Rate: Eq. (22) displays the percentage of positive situations that were expected to be negative but ended up being positive.

$$FNR = \frac{F_N}{T_P+F_N} \quad (22)$$

TABLE II. COMPARISON OF FPR AND FNR

Methods	FPR (%)	FNR (%)
RCNN	10.17	20.35
CNN	14.32	28.64
DNN	20.41	40.82
Hybrid DL-Attention Mechanism	**3.51**	**4.44**

The contrast provided Table II highlights the FPR and FNR performance of the Hybrid DL-Attention Mechanism. The results highlight the potential of deep learning's attention mechanisms by showing how they may considerably improve the model's performance in tasks that call for striking a careful balance between reducing false alarms and missed detections. By demonstrating the advantages of attention processes in optimizing detection and classification models for practical applications, this research adds to the larger area of computer vision (see Fig. 8).

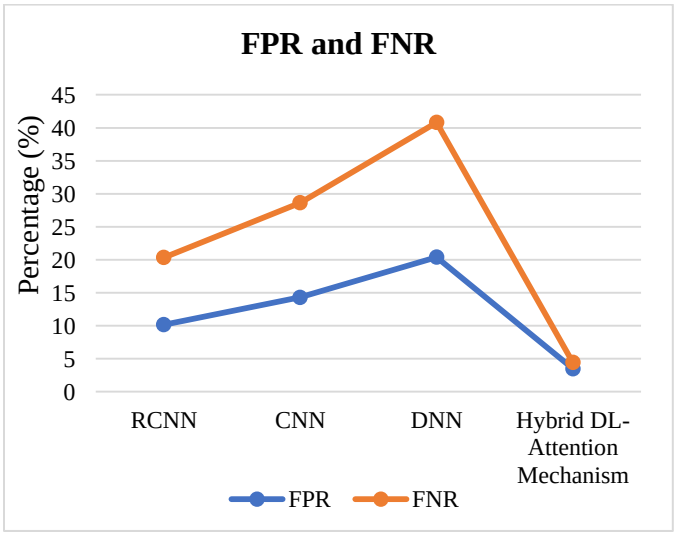

Fig. 8. Comparison of FPR and FNR.

Accuracy: Accuracy is used to evaluate the system model's performance as a whole. Every interaction can be properly foreseen is its central tenet. Eq. (23) provides the precision.

$$Accuracy = \frac{T_{Pos} + T_{Neg}}{T_{Pos} + T_{Neg} + F_{Pos} + F_{Neg}} \quad (23)$$

Precision: Precision also describes how closely two or more computations resemble one another in addition to being right. The link between precision and accuracy shows how often a judgment may be formed. Precision may be calculated using (24).

$$P = \frac{T_{Pos}}{T_{Pos} + F_{Pos}} \quad (24)$$

Recall: The percentage of all pertinent results that were effectively sorted by the procedures is known as recall. The suitable positive for such numbers is determined by dividing the genuine positive by the erroneous negative values. It is referenced in Eq. (25).

$$R = \frac{T_{Pos}}{T_{Pos} + F_{Neg}} \quad (25)$$

F1 Score: Accuracy and recall are combined in the F1-Score calculation. Eq. (26) computes the F1-Score using precision and recall.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (26)$$

A comparison of many object identification techniques, including RCNN, CNN, DNN, and a hybrid DL architecture with attention mechanisms, is shown in Table III. Key performance indicators including accuracy, precision, recall, and F1-Score are used to assess these approaches. With an astounding accuracy of 99.66%, the hybrid DL architecture with Attention Mechanisms emerges as the most accurate technique. The accuracy, recall, and F1-Score values for this approach are also superior, coming in at 98.12%, 97.65%, and 96.54%, respectively. The outcomes highlight how combining attention mechanisms into a hybrid DL architecture has a substantial influence and leads in remarkable object detection performance. Overall, the evaluation's findings highlight the promise of cutting-edge strategies that make use of attention processes to improve object detection resilience and accuracy in difficult situations (see Fig. 9).

TABLE III. COMPARISON OF PERFORMANCE METRICS

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RCNN	86.43	79.64	79.64	79.64
CNN	80.90	71.35	71.35	71.35
DNN	72.78	71.35	71.35	59.17
Hybrid DL-AM	**99.66**	**98.12**	**97.65**	**96.54**

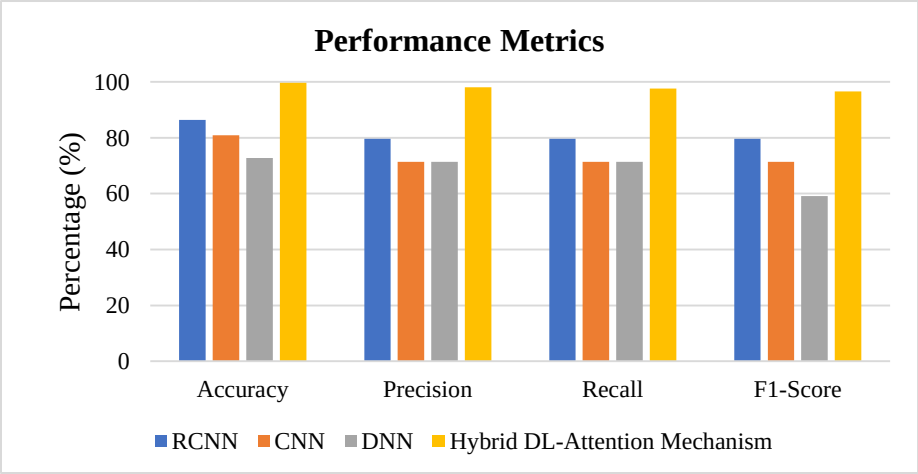

Fig. 9. Comparison of performance metrics.

VI. DISCUSSION

The paper addresses recent advances in computer vision, focusing on the application of deep learning algorithms for workplace safety, automated diagnostic frameworks for COVID-19 abnormalities [16], and plant segmentation. It also looks at navigational autonomy, proposing the use of 3D simulations in self-driving automobiles and Mask-RCNN for picture segmentation. The study also examines object detection and introduces SA-YOLOv3 for self-driving applications [21]. It also emphasizes the need of seamless visual classification through enhanced perception transformers. The study on facial expression recognition proposes the SqueezExpNet architecture [23], which efficiently solves real-world issues such as motion, obstacles, and lighting changes. The study emphasizes the ongoing advancement of computer vision techniques, as well as their potential influence on safety, healthcare, agriculture, autonomous systems, and other fields.

VII. CONCLUSION AND FUTURE WORK

This research examines the profound challenges associated with precise object localization and identification in complex real-world scenarios. The subject of computer vision has advanced significantly with the introduction of the new hybrid deep learning architecture. By integrating semantic and instance segmentation approaches with attention mechanisms to maximize their respective strengths, the strategy has demonstrated exceptional performance. Comprehensive tests and studies have clearly shown significant improvements over

conventional methods in terms of precision, memory, and overall correctness. Among the notable achievements are the integration of CNN and BiLSTM for semantic segmentation and the astute use of attention processes in instance segmentation. The model can now find fine-grained objects with the maximum accuracy, even in dynamic and complicated real-world scenarios, thanks to these advancements. This discovery has extensive implications for many disciplines, including medical imaging, robotics, autonomous driving, and more, that depend on reliable and consistent object localization. With the increasing complexity of real-world environments, the hybrid design in conjunction with attention mechanisms has opened up new and intriguing possibilities to enhance the adaptability and reliability of computer vision systems. Further study in this hybrid design in several domains and situations is a potential avenue. Particular difficulties arise with applications in industrial automation, agriculture, and underwater research. More substantial advancements in computer vision could result from evaluating the architecture's resilience and flexibility under these conditions. This forthcoming study aims to scrutinize the versatility of the hybrid technique, exploring its capacity to redefine and revolutionize object localization across diverse real-world scenarios.

REFERENCES

[1] Z. Yang, X. Yang, M. Li, and W. Li, "Small-sample learning with salient-region detection and center neighbor loss for insect recognition in real-world complex scenarios," Computers and Electronics in Agriculture, vol. 185, p. 106122, 2021.

[2] A. Abdelreheem, U. Upadhyay, I. Skorokhodov, R. Al Yahya, J. Chen, and M. Elhoseiny, "3dreftransformer: Fine-grained object identification in real-world scenes using natural language," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3941–3950.

[3] S. Xiong, G. Tziafas, and H. Kasaei, "Enhancing Fine-Grained 3D Object Recognition using Hybrid Multi-Modal Vision Transformer-CNN Models," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023), 2023.

[4] Y. Xi, W. Jia, Q. Miao, X. Liu, X. Fan, and H. Li, "FiFoNet: Fine-Grained Target Focusing Network for Object Detection in UAV Images," Remote Sensing, vol. 14, no. 16, p. 3919, 2022.

[5] Y. Chu et al., "A Fine-Grained Attention Model for High Accuracy Operational Robot Guidance," IEEE Internet of Things Journal, vol. 10, no. 2, pp. 1066–1081, 2022.

[6] S. Ye et al., "CDLT: A Dataset with Concept Drift and Long-Tailed Distribution for Fine-Grained Visual Categorization," arXiv preprint arXiv:2306.02346, 2023.

[7] J. Song, L. Miao, Q. Ming, Z. Zhou, and Y. Dong, "Fine-Grained Object Detection in Remote Sensing Images via Adaptive Label Assignment and Refined-Balanced Feature Pyramid Network," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 71–82, 2022.

[8] T. A. Abir, E. Kuantama, R. Han, J. Dawes, R. Mildren, and P. Nguyen, "Towards Robust Lidar-based 3D Detection and Tracking of UAVs," in Proceedings of the Ninth Workshop on Micro Aerial Vehicle Networks, Systems, and Applications, 2023, pp. 1–7.

[9] L. Dodds, I. Perper, A. Eid, and F. Adib, "A Handheld Fine-Grained RFID Localization System with Complex-Controlled Polarization," arXiv preprint arXiv:2302.13501, 2023.

[10] D. Rathnayake, M. Radhakrishnan, I. Hwang, and A. Misra, "LILOC: Enabling precise 3D localization in dynamic indoor environments using LiDARs," 2023.

[11] M. Cimdins, S. O. Schmidt, F. John, M. Constapel, and H. Hellbrück, "MA-RTI: Design and Evaluation of a Real-World Multipath-Assisted Device-Free Localization System," Sensors, vol. 23, no. 4, p. 2199, 2023.

[12] I. Kar, S. Mukhopadhyay, and B. Guha, "A Dual Fine Grained Rotated Neural Network for Aerial Solar Panel Health Monitoring and Classification," in International Conference on Data Management, Analytics & Innovation, Springer, 2023, pp. 457–477.

[13] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," arXiv preprint arXiv:2304.13705, 2023.

[14] "Deep learning-based object identification with instance segmentation and pseudo-LiDAR point cloud for work zone safety - Shen - 2021 - Computer-Aided Civil and Infrastructure Engineering - Wiley Online Library." https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12749 (accessed Aug. 03, 2023).

[15] E. Xie et al., "Segmenting Transparent Object in the Wild with Transformer." arXiv, Feb. 23, 2021. doi: 10.48550/arXiv.2101.08461.

[16] W. Ding, M. Abdel-Basset, H. Hawash, and O. M. ELkomy, "MT-nCov-Net: A Multitask Deep-Learning Framework for Efficient Diagnosis of COVID-19 Using Tomography Scans," IEEE Transactions on Cybernetics, vol. 53, no. 2, pp. 1285–1298, Feb. 2023, doi: 10.1109/TCYB.2021.3123173.

[17] S. Mukhamadiev, S. Nesteruk, S. Illarionova, and A. Somov, "Enabling Multi-Part Plant Segmentation with Instance-Level Augmentation Using Weak Annotations," Information, vol. 14, no. 7, Art. no. 7, Jul. 2023, doi: 10.3390/info14070380.

[18] F. X. Viana, G. M. Araujo, M. F. Pinto, J. Colares, and D. B. haddad, "Aerial Image Instance Segmentation Through Synthetic Data Using Deep Learning," Learn. Nonlin. Mod., vol. 18, no. 1, pp. 35–46, Sep. 2020, doi: 10.21528/lnlm-vol18-no1-art3.

[19] C. Chen, L. Z. Fragonara, and A. Tsourdos, "GAPointNet: Graph attention based point neural network for exploiting local feature of point cloud," Neurocomputing, vol. 438, pp. 122–132, May 2021, doi: 10.1016/j.neucom.2021.01.095.

[20] X. Wang et al., "Self-Paced Feature Attention Fusion Network for Concealed Object Detection in Millimeter-Wave Image," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 224–239, Jan. 2022, doi: 10.1109/TCSVT.2021.3058246.

[21] D. Tian et al., "SA-YOLOv3: An Efficient and Accurate Object Detector Using Self-Attention Mechanism for Autonomous Driving," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 5, pp. 4099–4110, May 2022, doi: 10.1109/TITS.2020.3041278.

[22] Q. Wang, J. Wang, H. Deng, X. Wu, Y. Wang, and G. Hao, "AA-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification," Pattern Recognition, vol. 140, p. 109547, Aug. 2023, doi: 10.1016/j.patcog.2023.109547.

[23] A. R. Shahid and H. Yan, "SqueezExpNet: Dual-stage convolutional neural network for accurate facial expression recognition with attention mechanism," Knowledge-Based Systems, vol. 269, p. 110451, Jun. 2023, doi: 10.1016/j.knosys.2023.110451.

[24] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 422–440.

[25] X. Sun et al., "FAIR1M: A Benchmark Dataset for Fine-grained Object Recognition in High-Resolution Remote Sensing Imagery." arXiv, Mar. 24, 2021. Accessed: Aug. 07, 2023. [Online]. Available: http://arxiv.org/abs/2103.05569.

[26] K. R., H. M., S. Anand, P. Mathikshara, A. Johnson, and M. R., "Attention embedded residual CNN for disease detection in tomato leaves," Applied Soft Computing, vol. 86, p. 105933, Jan. 2020, doi: 10.1016/j.asoc.2019.105933.

[27] C. Yin, X. Cheng, X. Liu, and M. Zhao, "Identification and Classification of Atmospheric Particles Based on SEM Images Using Convolutional Neural Network with Attention Mechanism," Complexity, vol. 2020, p. e9673724, Sep. 2020, doi: 10.1155/2020/9673724.