Object Detection and Recognition in Remote Sensing Images by Employing a Hybrid Generative Adversarial Networks and Convolutional Neural Networks

Dr Araddhana Arvind Deshmukh[1], Mamta Kumari[2], Dr. V.V. Jaya Rama Krishnaiah[3], Suraj Bandhekar[4], R. Dharani[5]
Head and Associate Professor, Department of Artificial Intelligence and Data Science,
Marathwada Mitra Mandal College of Engineering-Affiliated to Savitribai Phule Pune University[1]
Assistant professor, Department of CSE-ET, Panipat Institute of Engineering and Technology (PIET), Samalakha[2]
Associate Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, India[3]
Reader, Department of Mechanical Engineering, Rungta College of Engineering and Technology, Bhilai, C.G, India[4]
Assoc. Prof, IT, Panimalar Engineering College Chennai, India, 600123[5]

***Abstract*—Due to diverse backdrops, scale fluctuations, and a lack of annotated training data, the identification and recognition of objects in remote sensing images present major problems. In order to overcome these difficulties, this work suggests a novel hybrid technique that blends GAN and CNN. The suggested approach expands the small labelled dataset by synthesising realistic training examples using the generative abilities of GANs. The samples generated capture the various variances and backgrounds found in remote sensing photos, improving the object identification and recognition model's capacity to generalise. Additionally, CNNs, which are recognised for their outstanding feature extraction skills, are incorporated into the hybrid approach, enabling precise and reliable object identification and recognition. The model's CNN component is developed using both real and synthetic data, effectively combining the advantages of both fields. Several experiments are conducted on a large dataset of satellite photos to evaluate the performance of the proposed method. The results demonstrate that the hybrid model, with accuracy 97.32%, outperforms traditional approaches and pure CNN-based approaches in terms of dependability and resilience. The model may be efficiently generalised to unknown remote sensing images thanks to the GAN-generated samples, which bridge the gap among synthetic and actual data. The hybrid methodology used in this study demonstrates the possibility of merging GANs and CNNs for item detection and recognition using deep learning in remote sensing images.**

Keywords*—Object detection; Generative Adversarial Networks (GAN); Convolutional Neural Networks (CNN); deep learning; remote sensing; satellite images; hybrid model

I. INTRODUCTION

Remote sensing images captured by satellites and aerial platforms provide a wealth of valuable information about the Earth's surface. Analysing these images for object detection and recognition tasks is of utmost importance in various domains such as environmental monitoring, urban planning, and disaster management [1]. Deep learning algorithms have the potential to significantly increase the precision and effectiveness of item recognition and detection in this field when applied to remote sensing photos. Conventional methods to identifying and recognising objects in remote sensing photos frequently depend on rule-based algorithms and hand-crafted features, which have difficulties capturing the intricate and varied aspects of the data. The identification and classification of objects based upon their visual patterns and properties is made possible by deep learning techniques, which excel at autonomously learning hierarchical representations straight from the data [2].

In recent years, deep learning-based approaches have gained traction in remote sensing applications, leveraging the power of CNNs to learn discriminative features from large-scale remote sensing datasets. These models can effectively detect and recognize various objects, such as buildings, roads, vehicles, vegetation, and water bodies, in remote sensing images. By learning from a vast amount of data, CNNs can capture intricate spatial and spectral information, enabling accurate and robust object detection and recognition [3]. The advantages of deep learning in remote sensing imagery include its ability to handle complex scenes with diverse backgrounds, variations in lighting conditions, and different sensor characteristics. Additionally, deep learning models can learn from a wide range of remote sensing data sources, including optical imagery images and multispectral/hyperspectral data, making them versatile for different remote sensing applications [4]. By employing deep learning techniques, one can anticipate significant improvements in object detection and recognition performance in remote sensing images. The automated and efficient nature of deep learning models will enable faster analysis of large-scale datasets, leading to timely and accurate decision-making in various domains [5].

Additionally, the adaptability of deep learning approaches allows for transfer learning, where models trained on one remote sensing dataset can be fine-tuned on another dataset, reducing the need for extensive annotation efforts. The effectiveness of the deep learning-based object identification and recognition system will be assessed throughout this research using benchmark satellite imagery datasets, comparing it with current state-of-the-art techniques [6]. One will consider metrics such as detection precision, recall, and computational efficiency to assess the accuracy and efficiency of our proposed approach. By advancing the state-of-the-art in deep learning-based object detection and recognition in remote sensing images, this research has the potential to greatly enhance our understanding of the Earth's surface and enable informed decision-making in a wide range of applications. The accurate identification and classification of objects in remote sensing images contribute to improved land cover mapping, infrastructure monitoring, disaster response, and environmental assessments, ultimately leading to more effective and sustainable management of our planet's resources [7].

A fundamental aspect of a computer vision job, object detection has a wide range of uses in automation, autonomous vehicles, and surveillance. By extracting discriminative characteristics from big datasets, deep learning models in particular CNN have achieved extraordinary performance in object recognition over time [8]. However, traditional CNN-based approaches often struggle with detecting objects in challenging scenarios, such as occlusions, small object sizes, and cluttered backgrounds. To address these challenges and improve object detection performance, a hybrid approach that combines the power of GANs and CNNs has gained significant attention. Generative Adversarial Networks have demonstrated their effectiveness in generating realistic synthetic data that follows the same distribution as the real data. GANs consist of a generator network and a discriminator network that engage in a competitive learning process [9]. The generator network synthesizes samples, aiming to fool the discriminator into classifying them as real, while the discriminator network tries to accurately distinguish between real and synthetic samples. This adversarial training leads to the generation of synthetic data that closely resembles the real data distribution [10].

By leveraging the generative capabilities of GANs, the hybrid approach aims to improve object detection performance by generating additional training samples. These synthetic samples provide the CNN-based object detection model with a more diverse and comprehensive understanding of object classes, augmenting the training data and enhancing the model's ability to generalize to different variations and challenging scenarios [11]. The hybrid approach involves two main stages. In the first stage, a GAN is trained on a large dataset of real object images, learning the underlying data distribution and generating synthetic samples that closely resemble real objects. These synthetic samples, combined with the real training data, create an augmented dataset for training the CNN-based object detection model. In the second stage, the CNN learns discriminative features from the augmented dataset, enabling accurate and robust object detection [12].

Fig. 1 represents the hybrid approach which offers several advantages in object detection. Firstly, it addresses the challenge of limited training data by synthesizing additional samples that capture a broader range of object variations. This augmentation leads to improved generalization and better handling of rare or underrepresented object classes. Secondly, the adversarial training process in GANs encourages the generation of realistic and diverse synthetic samples, effectively enhancing the model's ability to handle variations in object appearance, scale, and background clutter. Lastly, the hybrid approach promotes the transferability of learned features across different datasets and domains, enabling the model to adapt and generalize well to unseen data [13]. This approach focuses on developing and evaluating the hybrid approach of GANs and CNNs for object detection. This work will conduct extensive experiments using benchmark object detection datasets, comparing the performance of the hybrid approach against traditional CNN-based methods. This work will evaluate metrics such as detection accuracy, precision, recall, and robustness to challenging scenarios to assess the effectiveness of the hybrid approach [14].

CNN and GAN have revolutionized the field of object detection by providing powerful tools for accurate and robust identification of objects in images and videos. CNNs are extensively used in the early stages of object detection to extract relevant features from the input data. These deep neural networks are trained on large datasets to learn hierarchical representations of objects, enabling them to recognize patterns and objects at different levels of abstraction [15]. The convolutional layers of CNNs perform local feature extraction, while the fully connected layers analyze the extracted features and classify the objects. On the other hand, GANs play a crucial role in enhancing object detection by generating realistic and high-quality synthetic data. By training a GAN on a large dataset, it learns to generate images that closely resemble real-world objects, even in complex scenarios or rare situations. These synthetic images can be combined with the original dataset to augment the training data, thus increasing the diversity and robustness of the object detection model [16]. The improved accuracy and robustness of object detection have implications in various real-world applications, including autonomous systems, surveillance, and object recognition [17]. The findings from this research contribute to advancing the field of object detection and pave the way for more effective and reliable computer vision systems in practical applications [18]. The goal of this project is to create an effective recognition and detection of objects system for satellite or other aerial platform-derived remote sensing photos. In order to overcome issues like changing lighting circumstances and sensor noise, the project intends to automate the detection and classification of things like roads, structures, and automobiles in these photos. To increase the effectiveness of analysing remote sensing data for uses like urban planning as well as disaster assessment, a precise and scalable system is being developed.

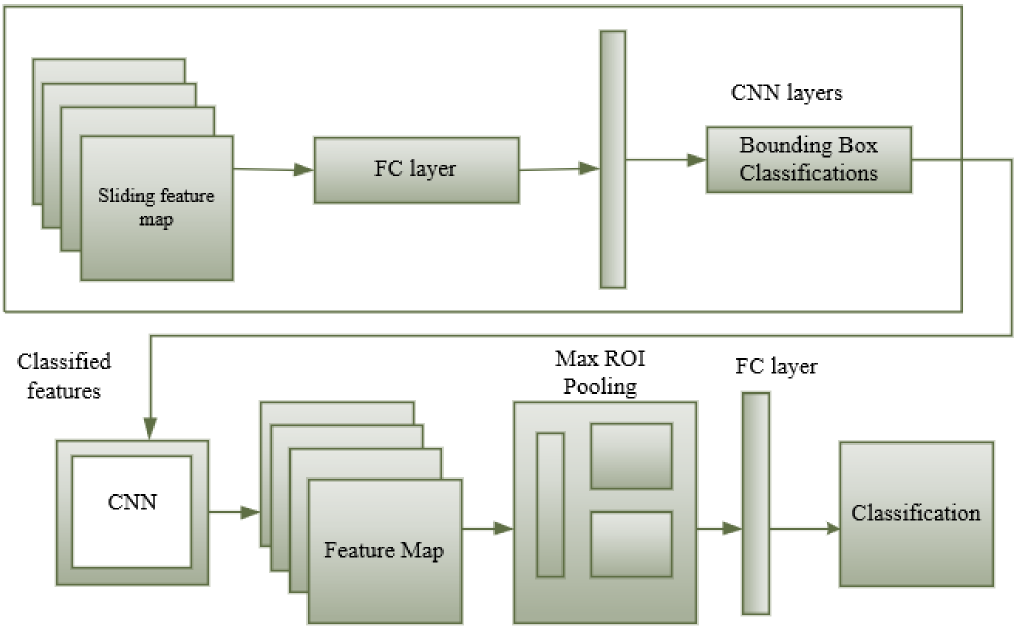

Fig. 1. CNN-GAN approach for object detection.

The following are the research's main contributions:

- The GAN-CNN approaches have been employed by other researchers, the proposed method stands out by introducing a distinctive data augmentation strategy that leverages GANs to generate authentic training examples, effectively addressing the challenges posed by diverse backgrounds, scale fluctuations, and limited annotated training data in remote sensing imagery.
- This hybrid strategy effectively augments the small annotated dataset and captures a variety of background fluctuations by harnessing the generative powers of GANs to create realistic training samples.
- The method makes the most of both domains by training the CNN components on both real and artificial data.
- The hybrid model significantly outperforms traditional and pure CNN-based approaches, according to experimental results.
- The potential of the hybrid technique for reliable and effective remote sensing item detection and recognition is demonstrated by the bridges of synthetic and actual information through GAN-generated samples, which improves the model's generalization to unknown remote sensing images.

The rest of the paper is structured as follows: Section II is described as related works while problem statement is explained in Section III. Similarly, Section IV is described as proposed methodology and Section V described as results and discussion and the conclusion is described as Section VI.

II. Related Works

Li et al. [19] proposed a novel lightweight CorrNet is an ORSI-SOD approach. In CorrNet, first a compact subnet is created for feature extraction and lessens the core network (VGG-16). Then initial crude prominence map is created using semantic features that are high-level in the correlation module, according to the coarse-to-fine technique. The granular scalar maps act as a geographical cue for low-level characteristics. Using the cross-layer association procedure the object position information is mined between high-level semantic characteristics. Finally, using low-level detailed characteristics, the coarse prominence map in the refinement subnet was refined to create the final fine saliency map. By lowering the requirements and calculations for each component, CorrNet ends up with only 4.09 million parameters and uses 21.09 gigaflops to execute. Results from tests on two open data sets demonstrate that lightweight CorrNet outperforms 26 modern techniques, including 16 huge methods based on CNN and two ultralight techniques, while saving a substantial amount of memory and runtime. Compact CorrNets are less suited to tackling difficult tasks involving the need for a greater capacity model because they are often built to contain less information. A lightweight CorrNet might not have enough capacity to catch such subtleties if the problem you're attempting to solve contains complex patterns or necessitates a lot of data presentation.

Sun et al. [20] created a part-based convolution neural network (PBNet) for integrated composite object detection in remote sensing pictures. PBNet evaluates an amalgamated object as a collection of parts and integrates component variables with contextual data to improve composite object recognition. Accurate ingredient knowledge can aid in the forecasting of an integrated item and help with problems resulting from various shapes and sizes. In order to provide accurate part information, a part placement module is developed that teaches the classification and localization of component positions using solely a boundary annotation. From a publicly available dataset, three representative categories of composite items are chosen for conducting operations to test the effectiveness and generalizability of this method's identification capabilities. This dataset includes sewage treatment facilities from seven Yangtze Valley cities, encompassing an extensive variety of geographic areas. Extensive testing on two datasets demonstrates that PBNet outperforms the current detection methods and reaches cutting-edge accuracy. Part-based models, however, primarily rely on precise part identification. The component detection process' noise or imprecision can have a negative impact on how well the PBNet performs. Because of its reliance on precise component localization, the model may be more susceptible to mistakes or noise during the part identification process, which might result in poorer robustness and generalization.

Zhang et al. [21] proposed the Feature Pyramid Network which makes use of the built-in multiple scales rounded characteristics as well as incorporates the strong-semantic, and the weak-semantic, excellent quality features simultaneously, has been proposed as an efficient region-based VHR remote sensing imagery identification framework. The DM-FPN is made up of two modules that may be trained end-to-end: a multi-scale region suggestion network and a multiple habitats object detection network. To broaden the range of training data and get beyond input image size limitations, a number of multi-scale training methodologies are presented. To improve detection performance, particularly for tiny and dense objects, multi-scale prediction techniques are presented. Extensive tests and thorough analyse on a sizable DOTA dataset show how successful the suggested architecture. DM-FPN introduces an additional level of complexity compared to the original FPN. The inclusion of double multi-scale features requires more computational resources, including memory and processing power. This increased complexity can impact training and inference times, making it less suitable for real-time or resource-constrained applications.

Chen et al. [22] proposed a CNN for object recognition that combines scene-contextual data. The environment-contextual feature pyramidal network (SCFPN), in particular, seeks to improve the bond among the objective and the scene and address issues brought on by fluctuations in target size. The network is created by repeating an accumulated remnant block in order to enhance the ability to perform extracting features. With the help of this block, the receptive field may harvest targets' deeper information and perform very well in terms of tiny object recognition. Additionally, group normalization, which separates each channel into group and determines the variance and mean for normalization within each group, is utilized to overcome the batch normalization's limitation and enhance the efficacy of the suggested model. A tough public dataset is used to validate the suggested approach. The experimental findings show that our suggested approach outperforms existing cutting-edge object identification techniques. To include scene-level contextual data, SCFPN adds further layers and calculations. This could result in higher computing demands for both inference and training. SCFPN may be less appropriate for actual time or limited in resources applications as a result of the extra complexity.

Yan et al. [23] Developed the full-scale object detection network (FSoD-Net), which is comprised of a suggested multiscale enrichment network backbone transmitted with scale-invariant regression layers (SIRLs), is a one-stage scanner. First, by integrating the Laplace kernels with less concurrent multiscale layers of convolution, MSE-Net offers the multiscale characterization improvement. Second, because SIRLs have three distinct independent extrapolation branch layers (small, medium, and large scales), full-scale object information is covered by the default discrete scale bounding boxes (bboxes) in the regression technique. A further approach employs an oval-specific scale joint loss that combines a strong L1 norm restriction with the soft max function for each regression branch layer. It can also hasten convergence and boost anticipated b-box classification results. The findings of extensive research conducted on challenging sets of data over identifying objects in aerial images (DOTA) and object identification in visual imagery from remote sensing (DIOR), which include numerous examples from various imaging platforms, show that FSoD-Net is capable of performing better than other cutting-edge one-stage detectors. FSoD-Net, tend to have a higher computational complexity compared to simpler tasks like image classification. Object detection involves not only classifying objects but also accurately localizing their positions and generating bounding box predictions. This increased complexity can result in longer training and inference times and require more computational resources.

Ming et al. [24] proposed a Critical Feature Capturing Network (CFCNet) enhancing the accuracy of detection by focusing on three areas: developing robust visualization of features, fine-tuning pre-anchored patterns, and label assignment optimization. For instance, while constructing robust key features specific to a task, researchers first isolate the classification and recurrence elements using the Polarization Attention Module (PAM). The Rotation Anchor Refinement Module (R-ARM) performs localization improvement on preset perpendicular anchors to create superior rotation anchors using the retrieved selective regression characteristics. After that, high-quality anchors are adaptively chosen using the Dynamic Anchor Learning (DAL) technique based on their capacity to capture crucial information. The proposed system achieves outstanding performance immediate time object recognition and more potent conceptual representations for structures in remote sensing pictures. Experimental finding on three remote sensing datasets which demonstrate that this technique outperforms numerous state-of-the-art methodologies in terms of detection performance. Attention mechanisms often require additional memory to store the attention maps or weights. If PAM generates attention maps with high spatial resolution, it can significantly increase the memory usage of the network, making it less suitable for memory-constrained environments or large-scale applications.

Lu et al. [25] proposed a feature-fusion SSD and an end-to-end network called attention. First, a complex feature fusion framework is created to improve the shallow features' semantic information. The feature information is then screened by the introduction of a dual-path attention module. The background noise is muted and the main feature is highlighted in this module using spatial focus and channel attention. A multiscale responsive field module follows, which improves the network's capacity for feature representation even more. In order to correct the imbalance among both the positive and negative samples, the loss function is lastly optimized. The results of this method's experiments on the data sets demonstrate its efficacy. Integrating attention mechanisms and feature fusion techniques into the SSD framework can introduce additional computational overhead. This may lead to increased training and inference times, making it less suitable for real-time or resource-constrained applications. SSD is designed to handle objects of different scales using a set of predefined anchor boxes. The introduction of attention and feature fusion mechanisms may introduce additional challenges in effectively handling scale variation. If not properly designed, the architecture may struggle to accurately detect objects at various scales, leading to potential detection errors. Attention mechanisms and feature fusion techniques introduce additional learnable parameters into the model. This increased parameter count may make the model more prone to overfitting, especially when the training dataset is limited or regularization techniques are not effectively employed. Careful parameter initialization and regularization strategies are required to mitigate these issues.

Fu et al. [26] suggested a feature-fusion architecture that uses a top-down pathway to add semantic descriptions to depth layer characteristics and an upward pathway to integrate top layer map features with low-level data to produce a multiple scales feature hierarchy. It is possible to create a potent representation of characteristics for numerous scales objects by mixing features from many levels. Axis-aligned boxes, which may include nearby instances and backdrop regions, have been used by the majority of prior approaches to find objects with variable directions and dense spatial distributions. This method creates a rotation-aware entity detector that locates items in remote sensing pictures by using oriented boxes. The region suggestion network adds more default angles to the anchors to better cover orientated objects. Instead of using horizontal proposals, which only imperfectly locate oriented objects, it uses oriented suggestion boxes to contain objects. For obtaining the characteristic maps of oriented suggestions for the next R-CNN subnetwork, the orientation-based RoI pooling procedure is implemented. On a public dataset, extensive tests are run for oriented object recognition in remote sensing photos. Feature-fusion architectures typically rely on having access to multiple modalities or sources of information. If one or more of these modalities are missing or inaccessible, the model may not be able to effectively perform feature fusion, limiting its performance or applicability.

III. Problem Statement

The problem addressed in this research is the development of a robust and efficient object detection and recognition system for remote sensing images. Remote sensing images, acquired from satellites or aerial platforms, provide valuable information for applications such as land cover mappings, urban planning, and disaster assessment. However, manually analysing these images is time-consuming and impractical, necessitating automated methods to identify and classify objects of interest [27]. The objective of this study is to design an accurate and scalable system that can detect and localize various objects in remote sensing imagery, such as buildings, roads, vehicles, and natural features, and subsequently recognize and categorize them into relevant classes. The system should be able to handle the challenges associated with remote sensing data, such as varying lighting conditions, sensor noise, and the large-scale nature of the datasets. By addressing these challenges, the proposed system aims to enhancing the efficiency and accuracy of object detection and recognition in remote sensing images, facilitating the analysis and interpretation of these critical data sources.

IV. Proposed GAN-CNN Approach

For object detection and recognition in remote sensing images, the suggested methodology combines GANs with CNNs. The studies make use of two datasets, DOTA and UCAS-AOD, totalling 2900 and 1500 aerial photos with labelled items, respectively. By producing fake remote sensing photos, GAN-based data augmentation is used to broaden the dataset's variety and generalisation. An object generation network plus an image interpretation network makes up the GAN model known as RDAGAN. The image interpretation network makes sure that the generated pictures approximate the specified domain while the object generation network creates realistic objects. In order to handle multiscale objects, the CNN-based object identification employs a Faster R-CNN architecture with multilayer Region Proposal Networks (RPNs). The RPNs improve the detection of both tiny and large objects by using various CNN levels to create object suggestions. Additionally, CNN feature map fusion is included in the suggested approach to improve the representation of tiny objects without the need of up sampling. Overall, to accomplish precise and reliable object detection and recognition, the hybrid strategy includes GAN-based data augmentation, CNN-based object detection, and specialised algorithms for remote sensing images. Fig. 2 shows the Overall architecture of the proposed methodology.

A. Dataset Collection

DOTA and UCAS-AOD are two datasets used in the study's experiments [28]. The responsibilities for oriented (OBB) and horizontal bounding boxes (HBB) are included in both. The DOTA dataset, which now comprises 2900 aerial images with pixel sizes ranging from 800 x 800 to 4000 x 4000 and objects belonging to fifteen different groups with an overall number of 196171 occurrences, is the biggest dataset for object recognition in aerial imagery. It is divided into three sets: training (1/2), validation (1/6), and testing (1/3). UCAS-AOD includes 15683 occurrences of each of the two classifications (Plane and Car) and 1500 aerial images, each measuring roughly 1000 by 1000 pixels. For training and assessment, the research randomly chose 1220 images. Employing the authorized development kit for DOTA, the study divided images into 1024 × 1024 squares with 512 pixels of overlaps. The datasets for identifying objects and recognition using remote sensing are displayed in Table I.

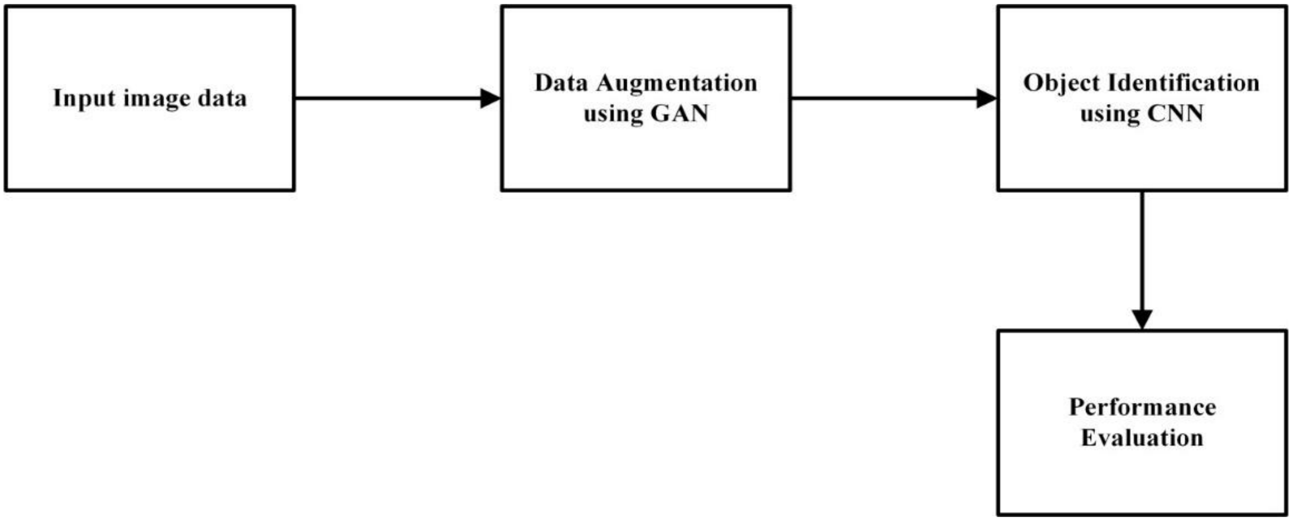

Fig. 2. Overall architecture of the proposed methodology.

TABLE I. REMOTE SENSING OBJECT DETECTION AND RECOGNITION DATASETS

Datasets	Total number of images	Categories	Instances	Image Sizes (Pixels)	Image Type
DOTA	2900	15	196171	800x800 to 4000x4000	RGB
UCAS-AOD	1500	2	15683	1000x1000	RGB

B. GAN based Data Augmentation

In several industries, including remote sensing and medical imaging, an image data augmentation technique based on GAN is frequently employed. Because neural networks in these domains need a lot of training data, it might be challenging to collect enough of it. It is simple for models to over fit or fall victim to the class imbalance problem when there are few data points. By creating fresh samples from a data distribution, the GAN-based picture data augmentation techniques can solve these issues. Fig. 3 shows the overall design of RDAGAN.

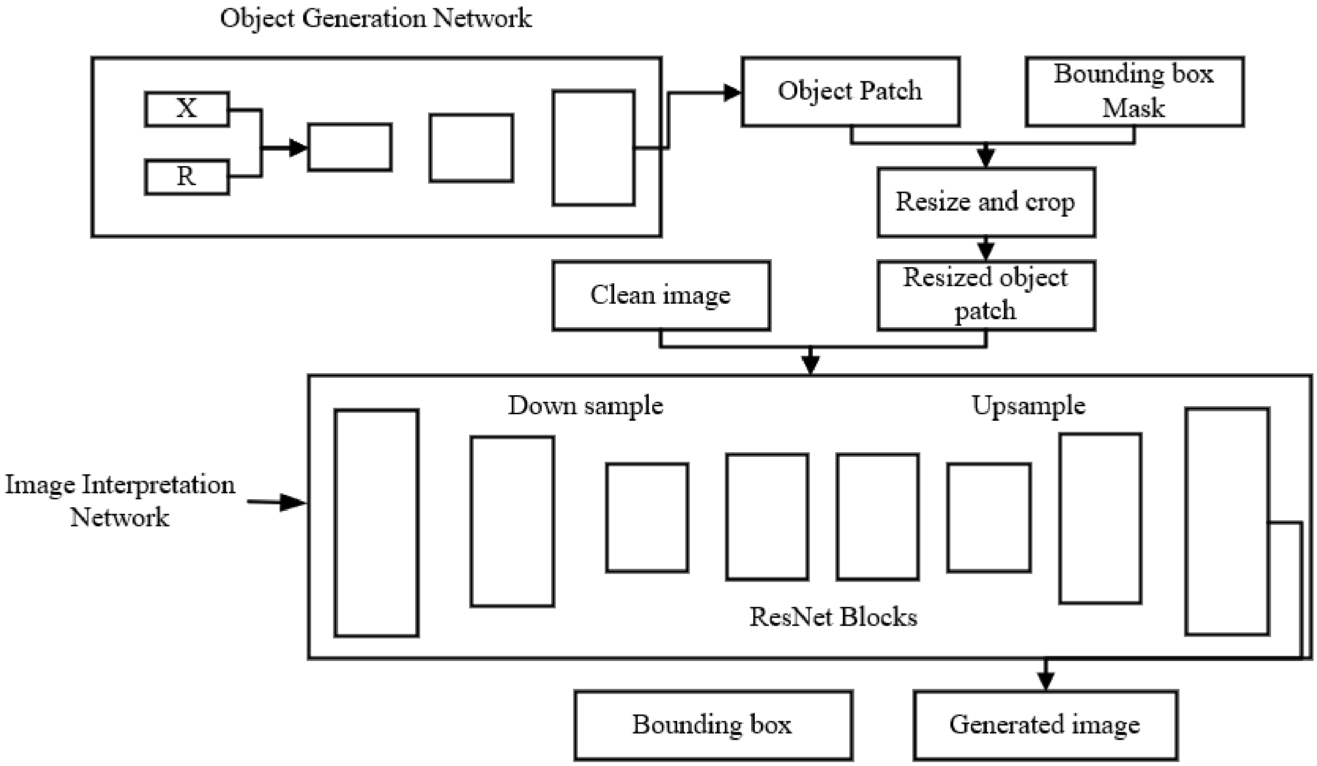

Fig. 3. The overall design of RDAGAN.

RDAGAN operates by training a GAN to generate synthetic data samples that closely resemble the true data distribution. This process significantly expands the available training data, mitigating the risk of over fitting and enhancing model generalization. The suggested robust data augmentation GAN (RDAGAN) model does the data augmentation. The objective was to create a framework that maps targeted images in the desired image domain $(i_s \in S)$ to cleaned images in the image cleanliness domain$(i_r \in R)$. The proposed algorithm was trained employing a dataset for object recognition with few images, most of which contained occlusions. The proposed framework uses the divide-and-conquer strategy, splitting the framework into two networks. The model tries to add realistic objects to the image (i_r), but it also tries to change the overall image to seem like it belongs in the domain that it is targeting *(S).* A single GAN structure makes it difficult to accomplish these objectives since the training process becomes unpredictable.

To be placed intoi_r, the target object's image is created by the object creation network. The visualization created by the object creation network is sent into the image interpretation network as a source of input. Due to the objectives of object formation and image interpreting, the imagery reduces the training instabilities in the image interpretation network. In order to generate a separated illustration of the target object, the network utilizes the InfoGAN architecture. The image interpretation network creates a loss function using the separated representations it received from the object development network. The crop and resize modules *C* were used to crop and resize object images $C(i_s)$ from image i_r in order to train the algorithm. The hidden code *r* and indestructible noise *x*, which are obtained by sampling from a normal distribution, are inputs to the generatorE_{obj}. In addition to validating the input images, the discriminator M_{obj}also forecasts the input hidden code *r′*.

The framework's goal N_{obj}includes two losses since the object formation network utilizes the InfoGAN design: an adversarial loss N_{GAN}^{obj} and an information lossN_{Info}^{obj}.

The negative outcome Eq. (1) explains how to utilize N_{GAN}^{obj} to ensure that the created patch $E_{obj}(x,r)$ resemble the domain of the intended images$C(i_s)$.

$$N_{GAN}^{obj} = G_{C(i_s)\sim S} \log M_{obj}\big(C(i_s)\big) + G_{C(i_r)\sim R} \log(1 - M_{obj}\left(E_{obj}(x,r)\right)) \quad (1)$$

The interaction of data between the produced image $G_{(C(i_r))}$ and the hidden code c is measured by loss of dataN_{Info}^{obj}. Eq. (2) explains how to compute it employing the mean squared error of the projected code *r′* from the discriminator M_{obj} and the input hidden code *c*.

$$N_{Info}^{obj} = G_{r\sim L(0,1),r'\sim} M_{obj}\left(E_{obj}(x,r)\right)^{(\|r-r'\|2)} \quad (2)$$

Eq. (3) states that the total goal, N_{obj}, is the aggregate of prior losses.

$$N_{obj} = N_{GAN}^{obj} + \lambda N_{Info}^{obj} \quad (3)$$

Where the extent of the data loss is represented by minimizing the overall aim, a simulation was trained.

The intended image $i_s \in S$ is created by combining the freshly processed images $i_r \in R$ and the object patch $E_{obj}(x,r)$ produced by the object patch network using the image interpretation network i_r. However, utilizing the standard GAN model and a single adversarial loss, it is difficult to carry out these difficult tasks at once. To lessen the load of difficult jobs, the suggested model contains a local discriminator and extra loss functions.

1) Generator: Similar to the generator employed in CycleGAN, the image interpretation network generator $\mathrm{E_{sc}}$ features encoder-decoder architecture with blocks from the residual network (ResNet) in the center. However, because every characteristic are down and up sampled, the generator has adaptability in the form variance of the output imagery.

The generator needs a bounding box maskd_a, which identifies the place of flame insertion, and to produce the image. Eq. (4) demonstrate that the place where the mask's value is 0 denotes the background and the position where its value is 1 denotes the flame's location. The bounding box region is determined using no special techniques. The height and breadth of the images are used to sample discrete uniform randomness at each location in the bounding box region.

$$d_a = \begin{cases} 1\, for\, flame \\ 0\, for\, background \end{cases} \quad (4)$$

By resizing the object patch and placing it in the region where the integer value of the bounding box mask is one, the resized object patch $i_q \coloneqq Resize\left(E_{obj}(x,r)\right)$ is created. The generator input is created by concatenating the scaled object patch with a clean imagery. By automatically combining the six-channel combined imagery and interpreting them such that they resemble the intended domain image$i_s \in S$, the generator produces the produced image$E_{sc}\left(i_q, i_r\right)$.

2) Discriminator: The global $\mathrm{M_{sc}^{global}}$ and the local $\mathrm{M_{sc}^{local}}$discriminators make up the image interpretation network. The image interpretation network responsibilities of image interpretation and natural merging are carried out by these discriminators.

The images produced by the generator, $E_{sc}\left(i_q, i_r\right)$, are evaluated by the global discriminator, M_{sc}^{global}. The PatchGAN discriminator, which analyses portions of the image rather than the entire one, serves as the foundation for its construction. It determines if the imagery is comparable to the intended domain image *S*. An adversarial loss results from this assessment outcome.

When utilizing the mask of the created image$E_{sc}\left(i_q, i_r\right)$, the local discriminator M_{sc}^{local}decides if the object patch $C(E_{sc}\left(i_q, i_r\right))$ is realistic and whether it can be acquired through the scaling and cropping operation *R*. The local discriminator's architecture is comparable to that of the global discriminators. However, similar to the InfoGAN discriminator, it also has a separate auxiliary layer that generates the anticipated code *r′* from the image's map of characteristics. The adversarial loss contains the local discriminator's authentic assessment outcome.

3) Adversarial loss: In order to illustrate the generator for the mapping from R to S, the study employed adversarial loss$\mathrm{M}_{\mathrm{GAN}}^{\mathrm{sc}}$. Eq. (5) represents the goal as follows:

$$M_{GAN}^{sc} = G_{i_s \sim q_S} \log M_{sc}^{global}(i_s) + G_{i_q \sim Egen(x,r), i_r \sim Q_r} log M_{sc}^{local}(E_{sc}(i_q, i_r)) + G_{i_s \sim q_S} \log(1 - M_{sc}^{global}(C(i_s))) + G_{i_q \sim Egen(x,r), i_r \sim Q_r} \log(1 - M_{sc}^{local}(C\left(E_{sc}(i_q, i_r)\right)) \tag{5}$$

Where the global discriminant M_{sc}^{global} seeks to separate the produced image $E_{sc}(i_q, i_r)$ from the images acquired from the intended domain S, whereas E_{sc} attempts to produce images identical to those received from the targeted domain S and object targets look as genuine objects. In order to distinguish the created object $C\left(E_{sc}(i_q, i_r)\right)$ from the object acquired from *S*, the local discriminator M_{sc}^{local} makes a determination.

C. CNN-Based Object Detection

The foundation for CNN-based object detection is introduced in this section. In contrast to categorization, the problem of object detection requires the prediction of both the precise location and labelling of numerous objects inside an image. R-CNN was initially a very effective method of object detection in the fields of computer vision. Three processes make up R-CNN: categorization, representation of characteristics obtained by CNN, and area proposal produced by selective search. Without having to calculate each ROI, Fast R-CNN can speed up object identification. In order to create fixed-dimensional characteristics from every ROI, it applies a ROI-pooling layer. The Hyper Region Proposal Network (RPN) now includes the production of object proposals because of the Faster R-CNN. Improved accuracy is achieved via faster R-CNN, which unifies object identification and recognition into a single network. It has influenced other creative and profitable item detectors for special instances.

The objective is to develop a unique detection network that can recognize both tiny and large items by utilizing the quicker R-CNN, which identifies objects employing high-level semantics. Faster R-CNN cannot be immediately implemented in remote sensing objects recognition because to recognize the distinctions between natural and remote sensing imagery. There are other optimization techniques suggested, such as multilayer RPNs and detecting subnetworks. The characteristic representation of the image is recovered using a sequence of convolution layers in the quicker R-CNN framework. RPN employs a number of anchors with predetermined sizes and aspect ratios over the map of characteristics to generate object proposals. Convolutional characteristics along with object suggestions are used in the categorization step to determine the labelling's and bounding box of various objects. Due to the distinctions among natural and remote sensing images, it is difficult to recognize certain tiny objects in large remote sensing images, such as vehicles and ships, and it is also important for balancing these multiscale objects because certain large objects, such as ground track fields, must be identified. The CNN built on quicker R-CNN along fails to operate well on remote sensing information in regard to all the difficulties.

1) Multilayer RPNs: In the attempts, the bases are raised first taking into account the RPN principles. The original CNN bases employ three ratios of aspect and three scales, $\{128^2, 256^2, 512^2\}$. Extremely small bases have been included to the collection of bases because small objects can be seen in remote sensing images. The study finally uses five scales $\{32^2, 64^2, 128^2, 256^2, 512^2\}$to accurately fit the ground truth after multiple failed tries. There are now fifteen bases instead of the previous nine bases. The reliability of particular small object detection has increased as a result of this improvement. Although adding additional bases is an easy and basic technique to find more small components, the precision still cannot be improved upon. The size of the characteristic map gets smaller as the CNN advances, and typically the last layer characteristics are input into the RPN. This causes smaller components in a big image to lose information. There may be no information about this object in the characteristic map of the previous layer. The study assumes that lowering network levels have reduced receptive fields and therefore better suited for tiny item identification. On the other hand, larger objects can be detected better at higher levels.

The VGG16 model and ResNet-101 framework are the foundations of the proposed SAPNet. There are 13 convolution layers in the VGG16. The four pooling layers can split all of the convolution layers into five segments. Faster R-CNN generates proposals using the conv5_3 layer, although it is challenging to include the characteristics of small objects. By creating a second RPN network, the study employs conv4_3 to forecast ROIs, in contrast to earlier techniques that exclusively used conv5_3 to create recommendations. Considering that conv5_3 in VGG16 acquires more characteristics to obtain huge objects whereas conv4_3 in VGG16 has additional characteristics concerning smaller objects. These two layers are suggested for adoption by multilayer RPNs.

There are two RPNs in the proposed network, as seen in Fig. 4. The first, RPN1, utilizes the conv5_3 layer, whereas the second, RPN2, employs the conv4_3 layer. While RPN1 concentrates on large proposals, RPN2 concentrates on modest proposals. The multilayer RPNs may provide multiscale remote sensing object suggestions through two RPN branches. When fitting huge objects in RPN1, the scale set $\{128^2, 256^2, 512^2\}$ is used. When generating tiny object suggestions in RPN2, the scale set $\{32^2, 64^2, 128^2\}$is used. RPN1 utilizes the characteristics map produced by block 5 for ResNet-101, whereas RPN2 employs the characteristic map produced by block 4.

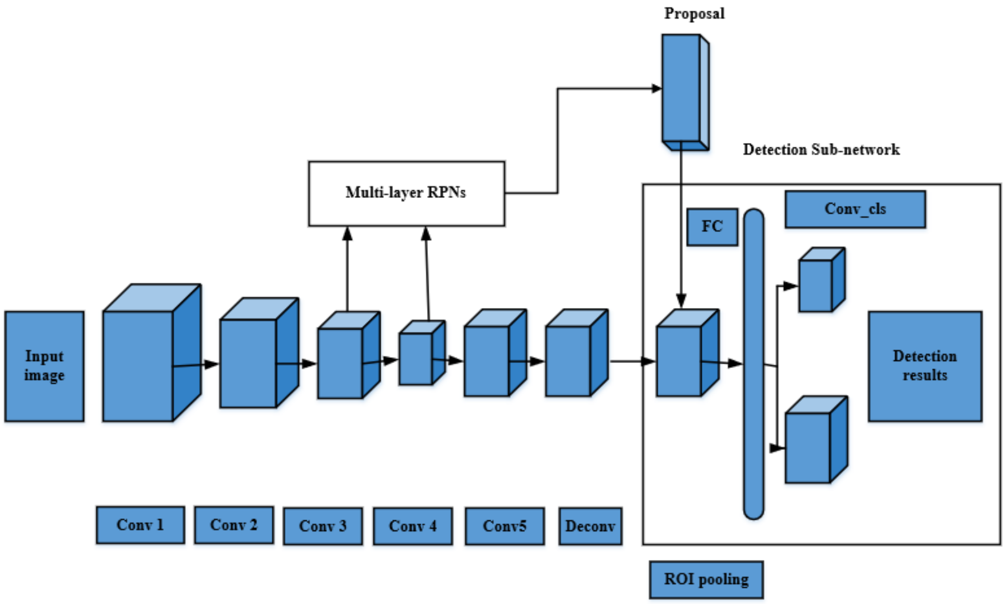

Fig. 4. Multilayered RPNs and a detection sub network.

The final identification subnetwork is where the anticipated bounding boxes and the labels of items originate from. A discrete probability, $q = (q_0, q_1, \ldots, q_k)$, over *K+1* classifications (K object classes and one background class), is present in each ROI during the training phase. A ground truth label y is assigned to each ROI. In this case, the first category has been configured as the background, while classes' *1-K* corresponds to the classes of the ground truth. The bounding box regression target is denoted by the expression$\hat{t} = \left(\widehat{t_u}, \widehat{t_v}, \widehat{t_x}, \hat{t}_y\right)$, the regressed bounding box by $t = \left(t_u, t_v, t_x, t_y\right)$, and the loss of every recognition layer by the expression is given in Eq. (6) and (7).

$$N(q, x, t, \hat{t}) = N_{els}(q, x) + \lambda[x \geq 1]N_{loc}(t, \widehat{t}) \tag{6}$$

Where,

$$N_{loc}(t, \widehat{t}) = \sum_{i \in [u,v,x,y]} smooth_{N1}\,(t - \widehat{t}) \tag{7}$$

In which,

$$smooth_{N1(y)} = \begin{cases} 0.5y^2 \\ |y| - 0.5, \quad otherwise \end{cases}$$

The bounding box loss and classification loss are counterbalanced by the hyperparameter λ. During the test,λ is set to 1.

2) CNN feature map fusion: Certain methods simply exaggerate the input images before feeding them into the network because the pretrained CNN model only accepts input with a predetermined size (224 × 224 in VGG16, for example). These methods have an impact on the effectiveness of the detection, particularly for tiny items. Some techniques up sample the input images to correct for scale inconsistencies, but this uses more memory and slows down processing. The quicker R-CNN network's ROI pooling layer is still being studied. The network can analyse pictures of any size thanks to its structure, which pools proposal regions into a fixed resolution of 7 ×7. Utilizing low level features is an effective approach to boost the information of tiny objects rather than up sampling the input images. The higher-level characteristics must be up sampled before being merged with the low-level characteristics since the size of the high-level characteristics is lower than that of the low-level characteristics.

V. Results and Discussion

A. Evaluation metrics

Having into consideration the needs for practical engineering purposes, the approach was assessed using accuracy, average precision (AP), recall, frames per second (FPS), and the precision-recall (PR) curve. To clearly illustrate the results, a distinct matching rule precision-recall (PR) curve was created, and a new PR was created according to it.

1) Precision and Recall (PR): PRC is a commonly employed metric utilised in numerous studies on object detection. Both the recall and accuracy measures can be written below, given that TP, FN, and FP stand for the amount of true positives, false negatives, and false positives, namely in Eq. (8)-(10)

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP} \tag{8}$$

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN}. \tag{9}$$

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

If a detection results has an intersecting over union (IOU) to a baseline value of at least 0.5, it is projected to be a true positive; alternatively, it is regarded as a false positive [29].

IoU refers to the intersection area over the union region between the two boxes with boundaries in the context of object detection. A projected boxes is considered a true positive (TP) if the IoU corresponding to the ground-truth box (G_0) and forecasted box (D_0) exceeds than the standard threshold, and a false positive (FP) else. IoU is characterised by Eq. (11)

$$\text{IoU} = G_0 \cap D_0 / G_0 \cup D_0 \tag{11}$$

A ground-truth box is said to be false negative (FN) when it is unable to locate the corresponding anticipated container. One may create a PR curve using these numbers for TP, FP, and FN, accuracy, and recall following calculating dynamic recall and precision at various scoring thresholds.

B. *Average precision (AP)*

The region underneath the PR curve is known as AP. They assess the detection outcomes of the suggested strategy using the mean average precision (mAP) in Eq.. (12) [30].

$$mAP = \frac{1}{N_0}\sum_{i=1}^{N} AP_{0_I} \tag{12}$$

1) Intersection-over-Detection (IoD): They create an additional challenging matched rule to compute TP in order to test the capacity to forecast the entire composites object. The intersection across the region of the outcome of detection is characterised by a new IoD, that is indicated by the following:

$$\text{IoD} = G_0 \ \cap D_0 \ / D_0 \tag{13}$$

IoD is more capable to show the superior performance of PBNet than IoU. As an outcome, when IoD > 0.5, an additional PR curve (dubbed PR-IoD) can be built.

TABLE II. PERFORMANCE METRICS OF PROPOSED GAN-CNN

Metrics	Proposed Method GAN-CNN
Accuracy	97.32
Precision	96.53
Recall	94.42
F1 score	92.27

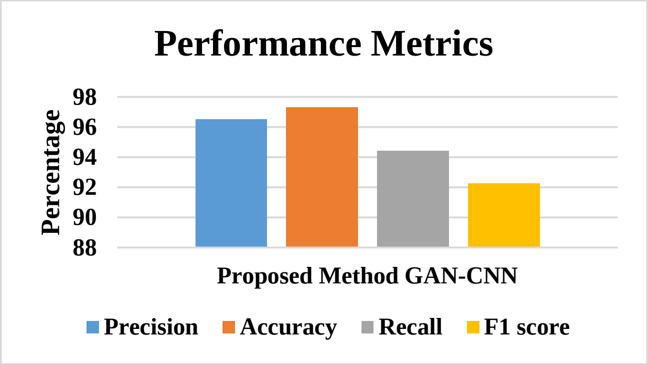

Fig. 5. Performance metrics of proposed GAN-CNN.

On the remote sensing picture dataset, the suggested GAN-CNN hybrid technique displayed outstanding results across key evaluation measures. The hybrid approach in Table II demonstrated its exceptional capacity to produce high accuracy, precision, and recall, culminating in a strong F1 score, with an accuracy of 97.32%, precision of 96.53%, recall of 94.42%, and an F1 score of 92.27% which is shown in Fig. 5. The metrics presented in Table II represent the average performance of the proposed GAN-CNN method across both the UCAS-AOD dataset and the DOTA dataset. These findings highlight the hybrid methodology's ability to greatly increase the accuracy of object detection and recognition, giving it an attractive option for strengthening the interpretation and analysis of remote sensing data in a variety of applications.

TABLE III. PRECISION, RECALL, F1-SCORE OF EXISTING METHODS AND PROPOSED GAN-CNN [31]

Methods	Recall	Precision	F1 score	Accuracy
YOLO v3	78.09	84.62	81.22	84.86
SSD	77.35	83.36	80.24	85.94
CFF-SDN	87.23	93.11	90.07	94.68
Faster R-CNN	83.32	89.65	86.37	87.64
GAN-CNN	94.42	96.53	92.27	97.32

The YOLO v3 model demonstrated a recall of 78.09%, precision of 84.62%, F1 score of 81.22%, as well as accuracy of 84.86% in the evaluation of object recognition methods using the specified metrics on a remote sensing image dataset in Table III. Similar results were shown by the SSD model, which had an accuracy of 85.94%, a recall of 77.35%, and precision of 83.36%. With a recall of 87.23%, precision of 93.11%, F1 score of 90.07%, as well as accuracy of 94.68%, the CFF-SDN approach in particular produced better results. Recall was 83.32%, precision was 89.65%, F1 score was 86.37%, and accuracy was 87.64% for the Faster R-CNN model. The proposed GAN-CNN hybrid strategy, with recall of 94.42%, precision of 96.53%, F1 score of 92.27%, as well as accuracy of 97.32%, however, Fig. 6 demonstrated the most astounding performance across all parameters. These results highlight the hybrid approach's clear superiority over more traditional approaches, emphasising its potential to achieve extraordinarily accurate and dependable item recognition and detection in remote sensing images.

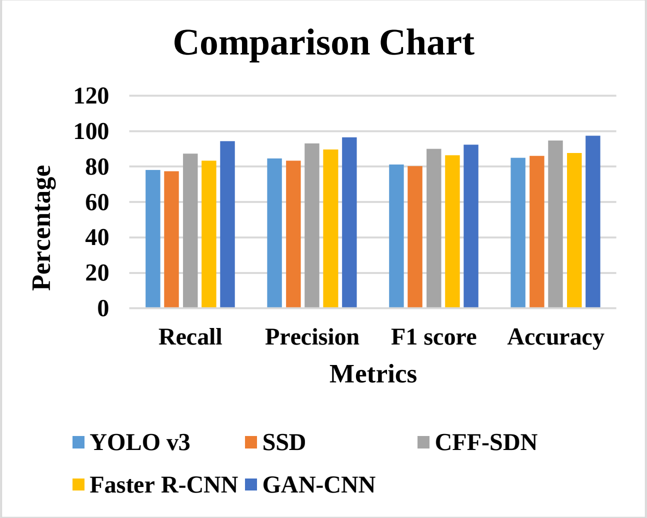

Fig. 6. Comparison chart of Precision, Recall, Fl score, Accuracy.

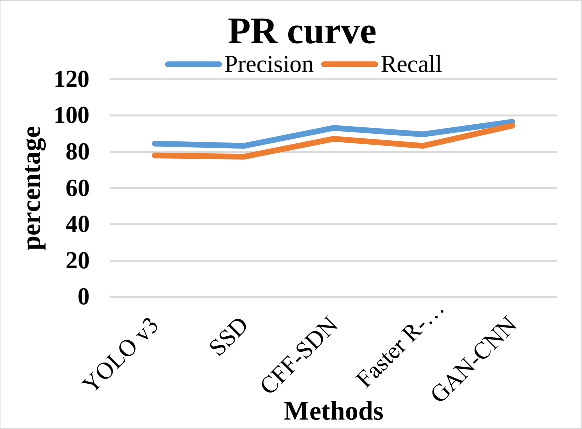

Fig. 7. PR curve of existing vs. proposed methods.

Variable performance across several approaches was demonstrated by a precision-recall curve assessment of the examined object detection algorithms on the remote sensing picture dataset in Fig. 7. While SSD demonstrated a precision of 83.36% and a recall of 77.35%, YOLO v3 attained an accuracy rate of 84.62% at a recall rate of 78.09%. The precision and recall numbers for the CFF-SDN technique were noticeably higher, with a 93.11% precision translating to an 87.23% recall percentage. Faster R-CNN achieved a recall of 83.32% and a precision rate that was 89.65%. With a precision of 96.53% and a phenomenal recall of 94.42%, the proposed GAN-CNN hybrid technique stood out as having the highest precision and recall rates. These trade-offs between high precision and recall, which are crucial for successful recognition and detection of objects activities in remote sensing images, offer insightful information about how well the models perform across various thresholds.

YOLO v3 scored a mAP of 82.73 among the investigated object detection techniques on the remote sensor image dataset in Table IV. SSD came in second with an overall rating of 81.53. With a mAP of 87.81, faster R-CNN displayed excellent performance. Notably, CFF-SDN fared better than the other approaches, obtaining a noteworthy mAP of 91.51. The maximum mAP of 94.32 was demonstrated by the suggested GAN-CNN hybrid strategy, outperforming all other approaches in Fig. 8. These findings emphasise the efficacy of the hybrid strategy and demonstrate its potential to greatly outperform both conventional single-model CNN methods and other cutting-edge methods in item recognition and detection in remote sensing images.

TABLE IV. mAP VALUES OF DIFFERENT METHODS

Methods	mAP
YOLO v3	82.73
SSD	81.53
CFF-SDN	91.51
Faster R-CNN	87.81
GAN-CNN	94.32

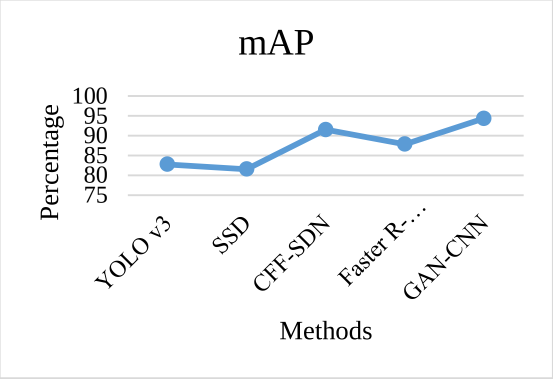

Fig. 8. Mean AP Curve for Different Methods.

VI. CONCLUSION

The authors propose a novel hybrid GAN-CNN approach for object detection and recognition in remote sensing images, aiming to address the challenges of diverse backgrounds, scale fluctuations, and limited annotated training data. Their method stands out through a data augmentation strategy that leverages GANs to generate realistic training examples capturing remote sensing photo variations, effectively expanding the labeled dataset. Notably, they integrate both real and synthetic data into their CNN component, combining the strengths of both domains. Their approach achieves superior performance, with an accuracy of 97.32%, surpassing traditional and pure CNN-based methods, while also showcasing the ability to generalize to unknown remote sensing images, bridging the gap between synthetic and actual data and demonstrating the potential of merging GANs and CNNs for remote sensing object detection and recognition. The evaluation's findings show how this hybrid approach can improve efficiency in comparison to more conventional CNN-based techniques. The hybrid technique solves issues particular to remote sensing images, including a lack of data annotations, unbalanced class distributions, and complicated backdrops, by introducing GANs into the learning pipeline. The GAN element creates artificial examples that accurately reflect the geographic distribution of targeted objects, enhancing the variety of the information and enhancing the generalisation abilities of the CNN component. Researchers found increased object detection precision, higher identification rates, and greater adaptability to difficult backdrops through empirical assessments. The combined methodology demonstrated its supremacy in remote sensing recognition and detection of objects tests by outperforming state-of-the-art techniques. The combined technique lowers the dependency on large-scale labelled datasets, which are frequently difficult to get in the satellite imagery area, by producing artificial data using GANs. This characteristic makes the technique realistic and adaptable to real-world circumstances by enabling more effective inference and training. Although the hybrid strategy has produced encouraging results, more study is needed in several areas. The accuracy and realistic nature of samples produced might be improved by adjusting the GAN design and investigating various GAN versions. Exploring various CNN designs, hyper parameters, and training methods would also offer insightful information for enhancing the efficiency of the hybrid technique. New opportunities for effective and precise analysis of remote sensing imagery are made possible by its capacity to handle issues unique to remote sensing data, enhance performance, and lessen the reliance on data with annotations. The use of accurate item identification and recognition in decision-making processes is crucial in many programmes, such as urban planning, agriculture, environmental monitoring, and disaster management.

REFERENCES

[1] S. N. Shivappriya, M. J. P. Priyadarsini, A. Stateczny, C. Puttamadappa, and B. D. Parameshachari, "Cascade Object Detection and Remote Sensing Object Detection Method Based on Trainable Activation Function," Remote Sensing, vol. 13, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/rs13020200.

[2] X. Qian, S. Lin, G. Cheng, X. Yao, H. Ren, and W. Wang, "Object Detection in Remote Sensing Images Based on Improved Bounding Box Regression and Multi-Level Features Fusion," Remote Sensing, vol. 12, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/rs12010143.

[3] H. Ma, Y. Liu, Y. Ren, and J. Yu, "Detection of Collapsed Buildings in Post-Earthquake Remote Sensing Images Based on the Improved YOLOv3," Remote Sensing, vol. 12, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/rs12010044.

[4] A. Mohan, A. K. Singh, B. Kumar, and R. Dwivedi, "Review on remote sensing methods for landslide detection using machine and deep learning," Transactions on Emerging Telecommunications Technologies, vol. 32, no. 7, p. e3998, 2021, doi: 10.1002/ett.3998.

[5] M.-T. Pham, L. Courtrai, C. Friguet, S. Lefèvre, and A. Baussard, "YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images," Remote Sensing, vol. 12, no. 15, Art. no. 15, Jan. 2020, doi: 10.3390/rs12152501.

[6] Y. Yu, J. Zhao, Q. Gong, C. Huang, G. Zheng, and J. Ma, "Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5," Remote Sensing, vol. 13, no. 18, Art. no. 18, Jan. 2021, doi: 10.3390/rs13183555.

[7] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network," Remote Sensing, vol. 12, no. 9, Art. no. 9, Jan. 2020, doi: 10.3390/rs12091432.

[8] S. S. Ismail, R. F. Mansour, A. El-Aziz, M. Rasha, A. I. Taloba, and others, "Efficient E-mail spam detection strategy using genetic decision tree processing with NLP features," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[9] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, "RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images," Remote Sensing, vol. 12, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/rs12030389.

[10] T. Hoeser and C. Kuenzer, "Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends," Remote Sensing, vol. 12, no. 10, Art. no 10, Jan. 2020, doi: 10.3390/rs12101667.

[11] K. Ravikumar, P. Chiranjeevi, N. M. Devarajan, C. Kaur, and A. I. Taloba, "Challenges in internet of things towards the security using deep learning techniques," Measurement: Sensors, vol. 24, p. 100473, 2022.

[12] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," Remote Sensing, vol. 12, no. 10, Art. no. 10, Jan. 2020, doi: 10.3390/rs12101662.

[13] K. Lambers, W. B. Verschoof-van der Vaart, and Q. P. J. Bourgeois, "Integrating Remote Sensing, Machine Learning, and Citizen Science in Dutch Archaeological Prospection," Remote Sensing, vol. 11, no. 7, Art. no. 7, Jan. 2019, doi: 10.3390/rs11070794.

[14] U. Alganci, M. Soydas, and E. Sertel, "Comparative Research on Deep Learning Approaches for Airplane Detection from Very High-Resolution Satellite Images," Remote Sensing, vol. 12, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/rs12030458.

[15] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui, "Deep Learning-Based Classification Methods for Remote Sensing Images in Urban Built-Up Areas," IEEE Access, vol. 7, pp. 36274–36284, 2019, doi: 10.1109/ACCESS.2019.2903127.

[16] N. Omer, A. H. Samak, A. I. Taloba, and R. M. Abd El-Aziz, "A novel optimized probabilistic neural network approach for intrusion detection and categorization," Alexandria Engineering Journal, vol. 72, pp. 351–361, 2023.

[17] H. Su et al., "HQ-ISNet: High-Quality Instance Segmentation for Remote Sensing Imagery," Remote Sensing, vol. 12, no. 6, Art. no. 6, Jan. 2020, doi: 10.3390/rs12060989.

[18] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: A survey," Image and Vision Computing, vol. 104, p. 104046, Dec. 2020, doi: 10.1016/j.imavis.2020.104046.

[19] X. Li, J. Deng, and Y. Fang, "Few-Shot Object Detection on Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3051383.

[20] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 173, pp. 50–65, Mar. 2021, doi: 10.1016/j.isprsjprs.2020.12.015.

[21] X. Zhang et al., "Geospatial Object Detection on High Resolution Remote Sensing Imagery Based on Double Multi-Scale Feature Pyramid Network," Remote Sensing, vol. 11, no. 7, Art. no. 7, Jan. 2019, doi: 10.3390/rs11070755.

[22] C. Chen, W. Gong, Y. Chen, and W. Li, "Object Detection in Remote Sensing Images Based on a Scene-Contextual Feature Pyramid Network," Remote Sensing, vol. 11, no. 3, Art. no. 3, Jan. 2019, doi: 10.3390/rs11030339.

[23] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, and H. Li, "IoU-Adaptive Deformable R-CNN: Make Full Use of IoU for Multi-Class Object Detection in Remote Sensing Imagery," Remote Sensing, vol. 11, no. 3, Art. no. 3, Jan. 2019, doi: 10.3390/rs11030286.

[24] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A Critical Feature Capturing Network for Arbitrary-Oriented Object Detection in Remote-Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3095186.

[25] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and Feature Fusion SSD for Remote Sensing Object Detection," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2021.3052575.

[26] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 161, pp. 294–308, Mar. 2020, doi: 10.1016/j.isprsjprs.2020.01.025.

[27] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "\mathcalR^2 -CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 8, pp. 5512–5524, Aug. 2019, doi: 10.1109/TGRS.2019.2899955.

[28] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, "Feature-Attentioned Object Detection in Remote Sensing Imagery," in 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan: IEEE, Sep. 2019, pp. 3886–3890. doi: 10.1109/ICIP.2019.8803521.

[29] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 173, pp. 50–65, Mar. 2021, doi: 10.1016/j.isprsjprs.2020.12.015.

[30] X. Jie, S. U. O. S. A. Technology, Y. Zheng, C. Dong-Ye, P. Wang, and M. Yasir, "Improved YOLOv5 Network Method for Remote Sensing Image Based Ground Objects Recognition," In Review, preprint, Feb. 2022. doi: 10.21203/rs.3.rs-1224458/v1.

[31] Y. Zhang, L. Guo, Z. Wang, Y. Yu, X. Liu, and F. Xu, "Intelligent Ship Detection in Remote Sensing Images Based on Multi-Layer Convolutional Feature Fusion," Remote Sensing, vol. 12, no. 20, p. 3316, Oct. 2020, doi: 10.3390/rs12203316.