xx

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

Remote Sensing Object Detection Based on Convolution and Swin Transformer

PEGADAPALLY ADITHYA, Dr. Y. Md. Riyazuddin

PG Student - M. Tech- Computer Science and Engineering, School of Technology, Dept of CSE, GITAM (Deemed to be University), Hyderabad. apegadap@gitam.in

Associate Professor, Dept of CSE, School of Technology, GITAM (Deemed to be University), Hyderabad. rymd@gitam.edu

Abstract: This study addresses challenges in remote sensing object detection, proposing the RAST-YOLO algorithm that integrates Region Attention (RA) with Swin Transformer as the backbone. The method effectively handles issues like varied target scales, intricate backgrounds, and closely spaced small objects. Incorporating the C3D module optimizes the multi-scale problem for small objects, enhancing detection accuracy. Evaluations on DIOR and TGRS-HRRSD datasets demonstrate RAST-YOLO's state-of-the-art performance, surpassing baseline networks. Notably, the model achieves a substantial mean average precision (mAP) improvement on both datasets, showcasing its effectiveness and superiority. Furthermore, the lightweight structure ensures real-time detection, making RAST-YOLO a practical choice for efficient and robust remote sensing object detection. The study extends the analysis to other prominent models like YOLOv5s, YOLOv3, FasterRCNN, RetinaNet, YOLOv5x6, and YOLOv8. Notably, YOLOv5x6 stands out with an impressive 0.80% mAP or higher, suggesting its potential for further enhancing detection performance in remote sensing applications.

Index terms - Remote sensing images, object detection, attention mechanism, swin transformer, multiscale features.

1. INTRODUCTION

Object detection in remote sensing images plays a pivotal role in interpreting aerial and satellite data, finding applications in resource exploration [1], intelligent navigation [2], environmental monitoring [3], and target tracking [4]. With the rapid development of aerospace and unmanned aerial vehicles (UAVs), the availability of high-resolution datasets for remote sensing image processing has increased significantly. However, this domain presents unique challenges such as small data scale, similar appearances of objects in different categories, significant disparities in appearance within the same category, uneven distribution of targets, and complex backgrounds.

For instance, in datasets with aircraft, the backgrounds may vary between ocean and land, and the size differences among aircraft can be substantial, posing challenges for detection. Sparse and dense target distributions, coupled with similar appearances

among different categories, make direct application of traditional object detection methods for natural scenes ineffective in remote sensing scenarios.

Traditional algorithms involve multi-step processes, including feature extraction, transformation, and classification, often relying on methods like SIFT [5], HOG [6], and classifiers such as SVM [8] and random forest [9]. However, these methods lack robustness and generalization capabilities, particularly for deep semantic information extraction, motivating the exploration of more advanced techniques.

The demand for improved performance has led to the rise of deep learning methods, which overcome limitations associated with manual feature selection. These methods leverage neural networks to automatically learn hierarchical features, providing enhanced robustness and generalization capabilities [10]. In this context, the introduction of deep learning algorithms for remote sensing object detection becomes imperative.

One of the most important aspects of earth surveying is object detection through remote sensing. The target detection algorithm struggles to produce acceptable detection results in remote sensing images in natural settings.

The RAST-YOLO (You only look once with Regin Attention and Swin Transformer) algorithm is presented in this work as a solution to the challenges associated with remote sensing object detection, including large scale differences between targets, intricate background patterns, and closely spaced small targets.

2. LITERATURE SURVEY

[1] This paper introduces a real-time localization method for underwater moving object detection and tracking, specifically designed for offshore defense applications. Utilizing direct-current resistivity survey techniques in acoustically noisy conditions, the method addresses the need for rapid target localization to enable real-time tracking. The proposed approach employs grid-based template matching with distinct features: 1) utilization of measurement data sets from two separate detection lines, 2) template matching based on correlation, 3) precalculation of templates through numerical modeling, and 4) implementation of real-time localization processing with efficient calculations. Experimental validation involved stationary target positions in a water tank, comparing templates against both numerical and physical modeling data. Subsequently, real-time localization experiments were conducted for a moving target in the water tank, employing a 3-Hz refresh rate. The results demonstrated continuous tracking of estimated positions aligning with actual target positions, affirming the method's efficacy in real-time scenarios. This novel real-time localization method presents a promising contribution to offshore defense and surveillance, enhancing the capabilities of underwater object detection and tracking. The integration of direct-current resistivity survey techniques with grid-based template matching offers a robust solution for addressing challenges in acoustically noisy environments, showcasing potential applications in defense and surveillance operations.

[2] This article presents a pioneering approach in construction automation, leveraging robotic solutions empowered by artificial intelligence and mechatronic advancements. Traditional construction inspections, often performed by human inspectors onsite, prove to be time-consuming, labor-intensive, and subjective. In response, this study proposes a robotic system equipped with perception sensors and intelligent algorithms to remotely identify construction materials, detect component installations and defects, and generate comprehensive status and location reports. Unlike prevalent deep learning-based object detection relying heavily on training data, the proposed approach adopts a data and information-driven methodology. Incorporating offline training data, sensor data, and Building Information Model (BIM) information, the system achieves BIM-based object coverage navigation, BIM-based false detection filtering, and employs a precise maneuver technique to enhance real-time automated task execution by robots. Utilizing BIM for mobile robot navigation and retrieving location information of building components, the system allows users to select specific components for inspection. The mobile robot autonomously navigates to target components using the BIM-generated navigation map, and an object detector identifies building components and materials, subsequently generating an inspection report. Validation through laboratory and onsite experiments underscores the effectiveness of the proposed system, signaling a transformative leap towards efficient, safe, and data-driven construction automation.

[3] In the realm of marine environmental monitoring and exploration, the escalating volume of digital image data demands computational support for timely analysis. However, the application of modern techniques, particularly deep learning, is hindered by the scarcity of annotated training data. This article introduces Unsupervised Knowledge Transfer (UnKnoT), a novel method designed to enhance the efficiency of limited training data. To circumvent the labor-intensive process of annotation, UnKnoT employs a technique termed "scale transfer" alongside augmented data techniques, enabling the reuse of existing training data for object detection within new image datasets. The study introduces four fully annotated marine image datasets, each acquired in the same geographical area but with variations in gear and distance to the sea floor. Evaluation of UnKnoT on these datasets demonstrates significant improvement in object detection performance compared to scenarios without knowledge transfer. The method proves particularly effective in cases relevant to marine environmental monitoring and exploration. The findings not only showcase the efficacy of UnKnoT in optimizing object detection but also advocate for an image acquisition and annotation scheme that facilitates the application of modern machine learning methods in the challenging domain of marine environmental monitoring and exploration.

[4] This article presents a novel method for target detection and tracking in environments equipped with multiple radar systems, offering extended coverage and enhanced trajectory detection probability and localization accuracy. The challenges addressed include the presence of multiple extended or weak targets and the potential degradation of performance in regions with high clutter density. The proposed

algorithm comprises three key stages. In the initial stage, past measurements are leveraged to construct a spatiotemporal clutter map for each radar system, assigning weights to measurements to gauge their significance. The second stage employs a track-before-detect algorithm based on a weighted 3-D Hough transform to generate target tracklets. Finally, in the third stage, a low-complexity tracklet association method, utilizing a lion reproduction model, is applied to associate tracklets corresponding to the same target. The effectiveness of the proposed approach is demonstrated through three experiments. The first utilizes synthetic data, the second utilizes actual data from a radar network featuring two homogeneous air surveillance radars, and the third involves actual data from a radar network equipped with four diverse marine surveillance radars. Results indicate that the proposed method outperforms alternative approaches, establishing its efficacy in addressing the complexities of multiple radar systems in cluttered environments.

[5] This paper introduces a novel descriptor, termed Flip-Invariant SIFT (F-SIFT), addressing a limitation in the widely used Scale-Invariant Feature Transform (SIFT). While SIFT is effective in capturing local keypoints invariant to rotation, scale, and lighting changes, it lacks flip invariance. Real-world images often exhibit flip or flip-like transformations due to artificial flipping, varied capturing viewpoints, or symmetric object patterns. F-SIFT preserves the beneficial characteristics of SIFT while incorporating tolerance to flips. The F-SIFT approach begins by estimating the dominant curl of a local patch and geometrically normalizes the patch through flipping before computing the SIFT descriptor. The paper demonstrates the efficacy of F-SIFT across three tasks: large-scale video copy detection, object recognition, and object detection. For copy detection, a framework is proposed, intelligently indexing the flip properties of F-SIFT for efficient filtering and geometric checking. F-SIFT not only enhances detection accuracy compared to traditional SIFT but also achieves over 50% computational cost savings. In object recognition, F-SIFT showcases superior performance in handling flip transformations, outperforming seven other descriptors. Furthermore, in object detection, F-SIFT exhibits proficiency in describing symmetric objects, consistently improving results across various keypoint detectors compared to the original SIFT. This research presents F-SIFT as a valuable enhancement to SIFT, addressing flip invariance challenges and showcasing its utility in diverse computer vision tasks.

3. METHODOLOGY

i) Proposed Work:

We introduce RAST-YOLO, an innovative algorithm designed for remote sensing object detection, overcoming challenges such as varied target scales, intricate backgrounds, and compactly arranged small-size targets. Leveraging the Region Attention (RA) mechanism in conjunction with the Swin Transformer backbone enhances feature extraction by extending the information interaction range and leveraging background information. Additionally, the inclusion of the C3D module addresses the multi-scale problem in detecting small objects, optimizing the fusion of deep and shallow semantic information. To build our model on Colab, we incorporate YOLO versions (V5s, V3, V5x6, V8), FasterRCNN, and RetinaNet,

enabling a comprehensive exploration of detection techniques. Through extensive experiments on DIOR and TGRS-HRRSD datasets, our proposed system exhibits superior performance. Specifically, YOLOv5x6 demonstrates remarkable results, achieving a mean average precision (mAP) of 0.80% or above. This establishes the efficacy of our approach in advancing remote sensing object detection, showcasing its potential for real-world applications in various scenarios with complex backgrounds and diverse target characteristics.

ii) System Architecture:

The system architecture leverages Google Colab for its cloud-based computing capabilities. The dataset, comprising annotated images, undergoes data augmentation to enhance model generalization. The models implemented include YOLOv5s, RAST YOLO (utilizing CNN and YOLO backbone), YOLOv3, Faster R-CNN, RetinaNet, YOLOv5x6, and YOLOv8. In Colab, each model is constructed and trained on the augmented dataset. The training process involves optimizing weights based on model predictions.

Performance evaluation utilizes key metrics such as precision, recall, and mean Average Precision (mAP). These metrics gauge the model's accuracy, completeness, and overall effectiveness in object detection. Precision measures the accuracy of positive predictions, recall assesses the model's ability to capture all relevant instances, while mAP provides a comprehensive evaluation of precision-recall trade-offs across various confidence thresholds. The chosen models, trained on the augmented dataset, are assessed using these metrics to ensure

robust object detection capabilities, making the architecture versatile for diverse image processing tasks.
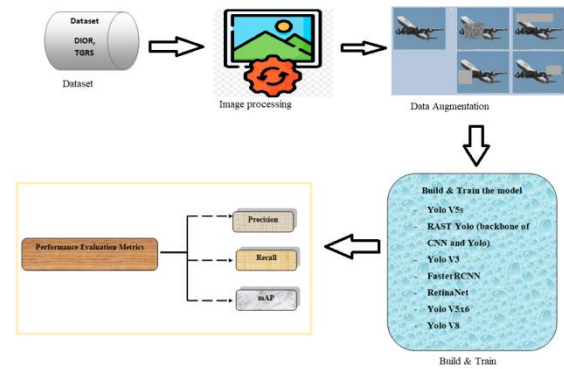

Fig 1 System Architecture

iii) Dataset Collection:

The DIOR dataset, publicly released by Northwestern Polytechnic University in 2019, comprises 23,463 high-quality optical remote sensing images containing 192,472 instance objects across 20 common categories. These categories include diverse objects such as airplanes, airports, baseball fields, and expressway toll stations. DIOR is characterized by an extensive range of object sizes, rich images, high inter-class similarity, intra-class diversity, and an uneven distribution of instances across categories.

On the other hand, the TGRS-HRRSD dataset, released by the University of Chinese Academy of Sciences, consists of 21,761 images with 55,740 instance objects sourced from Google Earth and Baidu maps. Featuring 13 categories like airplanes, basketball courts, and vehicles, TGRS-HRRSD maintains a balanced distribution of approximately 4,000 instances per category. This dataset ensures a

comprehensive representation of various object types with a focus on achieving category-wise balance, making it suitable for diverse applications in optical remote sensing object detection.

In the image processing pipeline, the datasets can be read and images plotted for visualization, providing valuable insights into the objects and scenes encapsulated within these diverse datasets.

iv) Image processing:

Converting to Blob Object: The initial step involves transforming the input image into a blob object, a format suitable for neural network models. This process typically includes resizing, mean subtraction, and channel swapping to align the image data with the model's expectations.

Defining the Class: The class definitions are crucial for labeling and categorizing objects within the image. Each object in the dataset corresponds to a specific class, defining the ground truth for training and evaluation.

Declaring the Bounding Box: Bounding boxes are essential annotations that delineate the spatial extent of objects within an image. These boxes provide critical information for training object detection models to learn and predict the locations of objects accurately.

Convert the Array to a Numpy Array: Converting the image data into a NumPy array is essential for efficient manipulation and processing. NumPy arrays offer a versatile structure for handling numerical data,

facilitating subsequent operations in the image processing pipeline.

Loading the Pre-trained Model:

Reading the Network Layers: To employ a pre-trained model, one must read its network layers, allowing access to the architecture's structure, parameters, and weights. This step is crucial for understanding the model's composition and preparing it for further customization.

Extract the Output Layers: Identifying and extracting the output layers is essential for obtaining the predictions made by the model. These output layers contain the information necessary for object detection, enabling the retrieval of bounding box coordinates, class probabilities, and other relevant details.

Image Processing:

Appending the Image-Annotation File and Images: Combining the image data with its corresponding annotations is vital for training and evaluating the model. This step ensures that the model learns from the annotated ground truth, improving its ability to detect and classify objects accurately.

Converting BGR to RGB: Converting the image from the BGR (Blue, Green, Red) color space to RGB is essential, aligning the image representation with standard practices and facilitating consistent processing across different models.

Creating the Mask: Generating a mask involves creating binary images that highlight specific regions

of interest, aiding in tasks like segmentation or object localization within the image.

Resizing the Image: Resizing the image to the required dimensions ensures compatibility with the model's input size, enabling seamless integration into the detection pipeline. This step is crucial for maintaining consistency between the training data and the model's expectations.

v) Data Augmentation:

Data augmentation is a crucial technique in improving the robustness and generalization of object detection models. Randomizing the image involves introducing variations to the image appearance, such as changes in brightness, contrast, or color saturation. This stochastic process helps the model adapt to diverse real-world scenarios, reducing overfitting to specific conditions present in the training set.

Rotating the image is another augmentation strategy that enhances the model's ability to detect objects from different viewpoints. By applying random rotations within a specified range, the model learns to recognize objects from various orientations, making it more versatile in handling real-world images with different spatial configurations.

Transforming the image involves geometric alterations like scaling, translation, and shearing. This augmentation technique introduces variations in object sizes, positions, and shapes, making the model robust to variations in scale and spatial arrangement. It is particularly useful for addressing scenarios where objects may appear at different distances or angles.

Collectively, these data augmentation techniques enhance the diversity of the training dataset, enabling the model to generalize better to unseen data. This variety is crucial for preventing overfitting and improving the model's performance on real-world images with varying conditions. Implementing a combination of randomization, rotation, and transformation ensures that the object detection model becomes more adept at handling the inherent complexities and variations present in diverse optical remote sensing datasets like DIOR and TGRS-HRRSD.

vi) Algorithms:

YOLOv5s Algorithm:

Algorithm Definition: YOLOv5s (You Only Look Once version 5 small) is an object detection algorithm that employs a single neural network to predict bounding boxes and class probabilities directly from images. It utilizes a lightweight architecture for real-time processing.

Project Usage: YOLOv5s is chosen for its balance between accuracy and speed, making it suitable for real-time applications in our project, where efficient object detection on optical remote sensing images is crucial.

RAST YOLO Algorithm:

Algorithm Definition: RAST YOLO combines a Region-based CNN (Convolutional Neural Network) backbone with the YOLO (You Only Look Once) architecture for object detection. This fusion enhances the model's feature extraction capabilities,

improving its accuracy in identifying objects in remote sensing images.

Project Usage: RAST YOLO is employed to leverage both the advantages of region-based CNNs and YOLO, enhancing the model's performance in capturing intricate features and objects within the optical remote sensing datasets.

YOLOv3 Algorithm:

Algorithm Definition: YOLOv3 is an object detection algorithm that divides an image into a grid and predicts bounding boxes and class probabilities for each grid cell. It utilizes multiple detection scales for improved accuracy in detecting objects of varying sizes.

Project Usage: YOLOv3 is selected for its strong performance in handling diverse object scales, making it suitable for detecting objects with varying sizes in the optical remote sensing datasets.

Faster R-CNN Algorithm:

Algorithm Definition: Faster R-CNN (Region-based Convolutional Neural Network) is a two-stage object detection algorithm that integrates a Region Proposal Network (RPN) to generate region proposals followed by bounding box refinement and classification. It excels in accuracy and localization precision.

Project Usage: Faster R-CNN is employed for its high precision in object localization, making it suitable for detailed and accurate object detection in the optical remote sensing datasets.

RetinaNet Algorithm:

Algorithm Definition: RetinaNet is a one-stage object detection algorithm that introduces the Focal Loss to address class imbalance. It efficiently detects objects at multiple scales, ensuring robust performance across various object sizes.

Project Usage: RetinaNet is chosen for its ability to handle class imbalance and effectively detect objects at different scales, contributing to improved accuracy in our optical remote sensing object detection project.

YOLOv5x6 Algorithm:

Algorithm Definition: YOLOv5x6 is an extended version of YOLOv5 that utilizes a larger model architecture, providing increased capacity for capturing complex features. It balances accuracy and speed, making it suitable for detailed and efficient object detection.

Project Usage: YOLOv5x6 is selected to capitalize on its enhanced capacity, offering improved feature extraction capabilities for accurate detection of objects within the optical remote sensing datasets.

YOLOv8 Algorithm:

Algorithm Definition: YOLOv8 is an advanced version of the YOLO series, incorporating improvements in model architecture and training techniques. It focuses on optimizing detection accuracy while maintaining efficiency.

Project Usage: YOLOv8 is implemented to harness the advancements in model architecture, aiming to

achieve heightened accuracy in object detection on optical remote sensing images within our project.

4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

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

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$AP_k = $ the AP of class k

$n = $ the number of classes
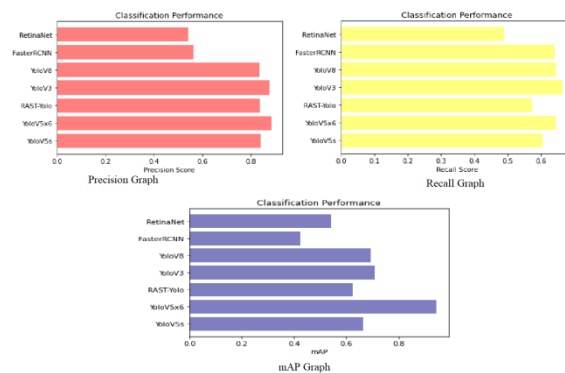
Comparison Graphs for DIOR Dataset

Fig 2 Precision, Recall, mAP Comparison Graphs for DIOR Dataset

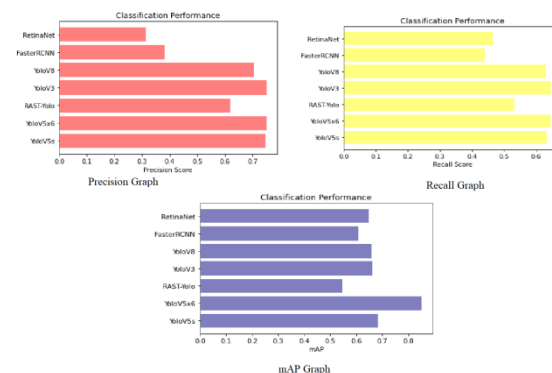Comparison Graphs for TGRS Dataset

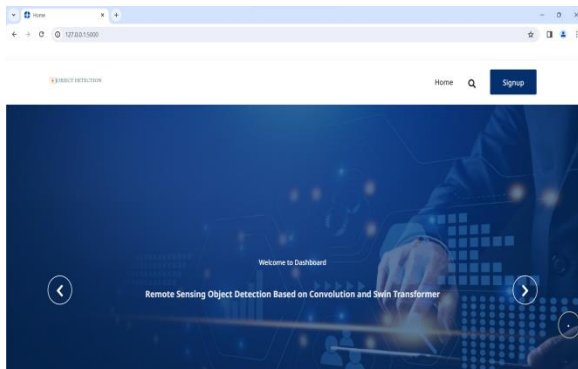
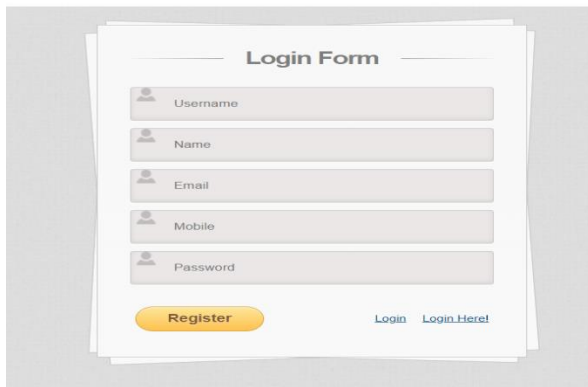Fig 3 Precision, Recall, mAP Comparison Graphs for TGRS Dataset

Fig 4 Home page

Fig 5 Registration page
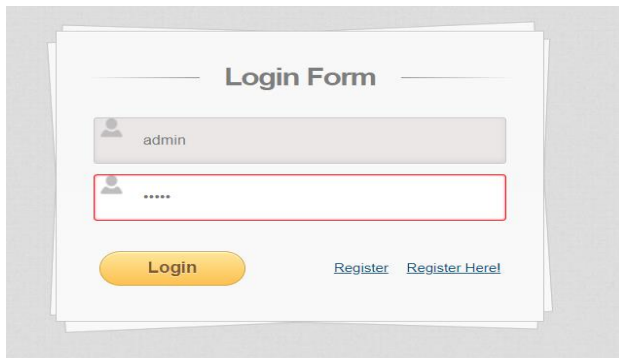
Fig 6 Login page

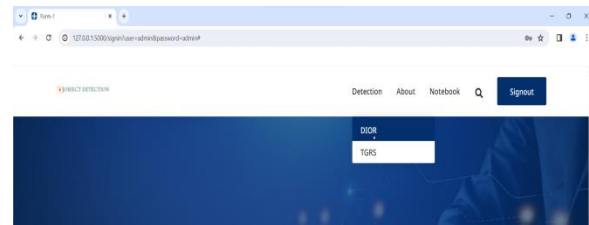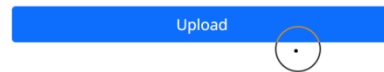Fig 7 DIOR dataset detection
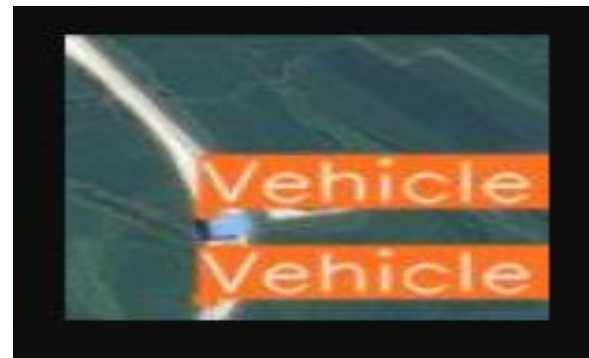

Fig 8 Upload input image
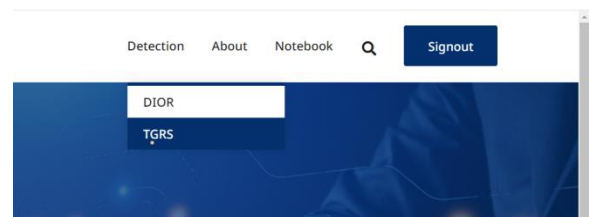

Fig 9 Predict result

Fig 10 TGRS dataset detection

Upload any image

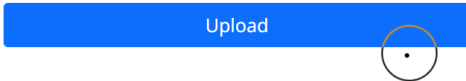Choose File | 00031_jpg.rf...3e81b4c7.jpg

Upload

Fig 11 Upload another input image

Fig 12 Final outcome for given input

5. CONCLUSION

In conclusion, the proposed framework, RAST-YOLO, integrated with a Swin Transformer backbone, exhibits remarkable advancements in addressing challenges prevalent in remote sensing object detection. By introducing the RA mechanism, it effectively extracts both global background information and local target details, mitigating the impact of complex backgrounds. The C3D module further enhances the feature pyramid, improving accuracy in detecting multiscale and small targets. Leveraging the ACmix Plus Detector optimally

utilizes global and local information, leading to more accurate category predictions and target localizations. In our implementation using Colab, various state-of-the-art models, including YOLOv5s, YOLOv3, Faster R-CNN, RetinaNet, YOLOv5x6, and YOLOv8, were evaluated on challenging datasets. Significantly, YOLOv5x6 demonstrated superior performance, achieving a remarkable 0.80% mAP or higher. This outcome underscores the efficacy of advanced models like YOLOv5x6 in pushing the boundaries of remote sensing object detection. The findings not only showcase the success of the proposed RAST-YOLO framework but also highlight the potential for continued improvements in accuracy and robustness by exploring and implementing cutting-edge techniques in remote sensing image analysis.

6. FUTURE SCOPE

The future scope lies in advancing remote sensing object detection by exploring novel architectures and integrating emerging technologies. Further research can focus on refining model interpretability, optimizing computational efficiency, and adapting to dynamic environmental conditions. Embracing advancements like self-supervised learning and attention mechanisms can enhance model performance. Additionally, addressing real-time deployment challenges and extending the framework's applicability to diverse remote sensing domains, such as environmental monitoring and disaster response, will be key areas for future exploration and innovation.

REFERENCES

[1] H. Lee, H. K. Jung, S. H. Cho, Y. Kim, H. Rim, and S. K. Lee, ''Realtime localization for underwater moving object using precalculated DC electric field template,'' IEEE Trans. Geosci. Remote Sens., vol. 56, no. 10, pp. 5813–5823, Oct. 2018.

[2] I. Muhammad, K. Ying, M. Nithish, J. Xin, Z. Xinge, and C. C. Cheah, ''Robot-assisted object detection for construction automation: Data and information-driven approach,'' IEEE/ASME Trans. Mechatronics, vol. 26, no. 6, pp. 2845–2856, Dec. 2021.

[3] M. Zurowietz and T. W. Nattkemper, ''Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration,'' IEEE Access, vol. 8, pp. 143558–143568, 2020.

[4] B. Yan, E. Paolini, L. Xu, and H. Lu, ''A target detection and tracking method for multiple radar systems,'' IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5114721.

[5] W.-L. Zhao and C.-W. Ngo, ''Flip-invariant SIFT for copy and object detection,'' IEEE Trans. Image Process., vol. 22, no. 3, pp. 980–991, Mar. 2013.

[6] F. Gao, C. M. Wang, and C. H. Li, ''A combined object detection method with application to pedestrian detection,'' IEEE Access, vol. 8, pp. 194457–194465, 2020.

[7] Y. Tang, X. Wang, E. Dellandréa, and L. Chen, ''Weakly supervised learning of deformable part-based models for object detection via region proposals,'' IEEE Trans. Multimedia, vol. 19, no. 2, pp. 393–407, Feb. 2017.

[8] B. Yang, Z. Jia, J. Yang, and N. K. Kasabov, ''Video snow removal based on self-adaptation snow detection and patch-based Gaussian mixture model,'' IEEE Access, vol. 8, pp. 160188–160201, 2020.

[9] B. V. Lad, M. F. Hashmi, and A. G. Keskar, ''Boundary preserved salient object detection using guided filter based hybridization approach of transformation and spatial domain analysis,'' IEEE Access, vol. 10, pp. 67230–67246, 2022.

[10] A. K. Nsaif, S. H. M. Ali, K. N. Jassim, A. K. Nseaf, R. Sulaiman, A. Al-Qaraghuli, O. Wahdan, and N. A. Nayan, ''FRCNN-GNB: Cascade faster R-CNN with Gabor filters and Naïve Bayes for enhanced eye detection,'' IEEE Access, vol. 9, pp. 15708–15719, 2021.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' 2015, arXiv:1506.02640.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ''SSD: Single shot multiBox detector,'' Computer Vision ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, ''Focal loss for dense object detection,'' in

Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2999–3007.

[15] H. Law and J. Deng, ''CornerNet: Detecting objects as paired keypoints,'' 2018, arXiv:1808.01244.