IJHRMOB

International Journal of HRM and Organizational Behavior

www.ijhrmob.com

editor@ijhrmob.com

International Journal of
HRM and Organizational Behavior ISSN 2454 - 5015

Remote Sensing Object Detection Based on Convolution and Swin Transformer

SHAIK JILANI, Assistant Professor, Dept of CSE, Chirala Engineering College, Chirala,
jilani.peace@gmail.com

VEMPATI ANITHA, PG Student-MCA, Dept of MCA, Chirala Engineering College, Chirala,
anithavempati65@gmail.com

Abstract: Remote sensing object detection poses significant challenges, particularly in natural environments with intricate backgrounds and small-scale targets. Addressing this, the RAST-YOLO algorithm integrates the Regin Attention (RA) mechanism with Swin Transformer as its backbone, enhancing feature extraction for improved detection accuracy amidst complex backgrounds. Additionally, the incorporation of the C3D module facilitates the fusion of deep and shallow semantic information, effectively tackling the multi-scale issue inherent in remote sensing targets, thus elevating the detection precision, particularly for smaller objects. Extensive experimentation on DIOR and TGRS-HRRSD datasets underscores the algorithm's prowess, showcasing state-of-the-art accuracy, efficiency, and robustness. Comparative analysis against baseline networks underscores RAST-YOLO's superiority, notably enhancing mean average precision (mAP) on DIOR and TGRS-HRRSD datasets. Moreover, the algorithm's lightweight architecture ensures real-time detection speeds without compromising on detection efficacy. Further exploration utilizing techniques like YOLOv5x6 and YOLOv8 exhibits promising potential, with YOLOv5x6 demonstrating a notable mAP improvement of over 0.80%, reinforcing its viability for advanced remote sensing object detection applications.

INDEX TERMS: Remote sensing images, object detection, attention mechanism, swin transformer, multiscale features.

1. INTRODUCTION

Remote sensing plays a pivotal role in various fields, including resource exploration, intelligent navigation, environmental monitoring, and target tracking. With the rapid advancement in aerospace and unmanned aerial vehicles (UAVs), there has been a surge in the creation of high-resolution and high-quality datasets for remote sensing image processing [1]-[4]. The primary objective of remote sensing target detection is to ascertain the presence of objects of interest in remote sensing images and provide their spatial coordinates. However, remote sensing object detection encounters several challenges distinct from those faced by traditional natural scene image detection methods [1]-[4].

In contrast to natural scene images, remote sensing images often exhibit a smaller data scale, leading to unique challenges in object detection. Furthermore, objects in remote sensing images can appear similar

ISSN 2454 - 5015

across different categories or exhibit significant variations within the same category. This disparity in appearance, coupled with uneven distributions of small, medium, and large targets, presents formidable obstacles. Additionally, the density and distribution of targets can vary widely, ranging from sparse to dense arrangements, further complicating detection tasks. Complex backgrounds and class imbalances further exacerbate the challenges encountered in remote sensing object detection [1]-[4].

For instance, as illustrated in Fig. 1, objects such as aircraft may appear against backgrounds of ocean or land, with significant variations in size. Similarly, targets can exhibit sparse or dense distributions, further compounded by the presence of objects belonging to different categories but sharing highly similar appearances [1]-[4]. These complexities underscore the inadequacy of employing traditional object detection methods designed for natural scene images in the context of remote sensing.

Traditional object detection algorithms typically involve multiple steps, including feature extraction, feature transformation, and classifier prediction. However, these methods often rely on manual feature selection and exhibit limited capabilities in extracting deep semantic information, leading to reduced robustness and generalization [1]-[4]. The advent of deep learning, particularly convolutional neural networks (CNNs), has revolutionized computer vision tasks, including target detection.

Deep learning-based target detection algorithms can be broadly categorized into one-stage and two-stage approaches. One-stage algorithms, such as YOLO and SSD, generate class probabilities and object coordinates directly in a single stage, eliminating the need for region proposals. In contrast, two-stage algorithms, like R-CNN and Faster RCNN, involve separate stages for generating region proposals and refining object locations [1]-[4].

Transformer architectures, initially developed for natural language processing tasks, have also gained prominence in computer vision. Vision Transformer (ViT), for instance, employs Transformer-based architectures for image classification, bypassing the need for CNNs. Transformer-based object detection algorithms can be categorized based on their network structures, with some utilizing Transformers as backbones alongside CNNs for feature extraction and prediction, while others rely solely on Transformers for both feature extraction and prediction [20]-[29].

Despite the advantages offered by Transformer-based object detection algorithms, including improved detection accuracy, they suffer from drawbacks such as large model parameters, slow training and inference speeds, and dependency on large datasets. Moreover, their computational costs increase exponentially with image resolution, rendering them less suitable for processing high-resolution images [20]-[29].

2. LITERATURE SURVEY

The field of object detection in various domains, including construction automation, environmental monitoring, and exploration, has witnessed significant advancements driven by data-driven approaches and innovative techniques. Muhammad et al. [2] proposed a robot-assisted object detection method for construction automation, emphasizing a data and information-driven approach. Their work focused on leveraging robotic systems for efficient

ISSN 2454 - 5015

object detection tasks, facilitating automation in construction processes.

Zurowietz and Nattkemper [3] introduced an unsupervised knowledge transfer method for object detection in marine environmental monitoring and exploration. Their approach emphasized the transfer of knowledge from unsupervised data sources to improve object detection accuracy in marine environments, showcasing the potential for leveraging unsupervised learning techniques in specialized domains like marine monitoring.

In the realm of feature extraction, Zhao and Ngo [5] presented a flip-invariant SIFT method tailored for copy and object detection tasks. Their work addressed the challenge of image variations due to flips, enhancing the robustness of object detection algorithms. Gao et al. [6] proposed a combined object detection method with applications in pedestrian detection. Their approach integrated multiple detection techniques to improve overall detection performance, demonstrating the effectiveness of fusion strategies in object detection tasks.

Tang et al. [7] introduced a weakly supervised learning approach for deformable part-based models in object detection via region proposals. Their method aimed to alleviate the reliance on annotated training data by leveraging weak supervision, highlighting the importance of innovative learning paradigms in advancing object detection capabilities.

In terms of classifier algorithms, Lad et al. [9] presented a boundary-preserved salient object detection method using a guided filter-based hybridization approach. Their work focused on enhancing object detection performance while preserving boundary information, contributing to the development of robust detection algorithms for salient objects.

The landscape of object detection algorithms includes both one-stage and two-stage approaches, each with its unique strengths and applications. Redmon et al. [12] introduced the YOLO (You Only Look Once) algorithm, a unified real-time object detection method that directly generates class probabilities and object coordinates in a single stage. YOLO's efficiency and real-time performance have made it a popular choice for various applications requiring rapid object detection.

Lin et al. [14] proposed RetinaNet, a one-stage object detection algorithm that addresses the challenge of dense object detection. RetinaNet introduced the focal loss function to mitigate the class imbalance issue inherent in dense object detection tasks, demonstrating superior performance in accurately detecting densely packed objects.

The literature also encompasses advancements in transformer-based object detection algorithms, leveraging transformer architectures for improved detection accuracy. Notably, Dosovitskiy et al. [20] introduced the Vision Transformer (ViT), which revolutionized image classification tasks by employing transformer-based architectures without relying on convolutional neural networks. ViT's success paved the way for transformer-based object detection algorithms, which can be categorized based on their network structures.

Transformer-based object detection algorithms with transformers as backbones alongside CNNs for feature extraction and prediction have shown

ISSN 2454 - 5015

promising results. For instance, Wang et al. [28] proposed the Flexible Pyramid Transformer (FPT), which combines the flexibility of transformer architectures with the hierarchical feature representation of pyramid structures, achieving state-of-the-art performance in object detection tasks.

Swin Transformer [29], introduced by Liu et al., presents a hierarchical transformer architecture that leverages shifted windows for efficient computation and enhanced feature representation. Swin Transformer's innovative design addresses the scalability and efficiency challenges associated with transformer-based object detection algorithms, making it well-suited for processing high-resolution images.

Despite the advancements in transformer-based object detection algorithms, challenges remain, including large model parameters, slow training and inference speeds, and dependency on large datasets. Future research efforts aim to address these challenges and further enhance the capabilities of transformer-based object detection algorithms for a wide range of applications.These studies collectively demonstrate the diverse range of methodologies employed in object detection research, spanning traditional feature extraction techniques to state-of-the-art deep learning algorithms. The choice of method often depends on the specific requirements of the application domain, with each approach offering unique advantages and trade-offs.

3. METHODOLOGY

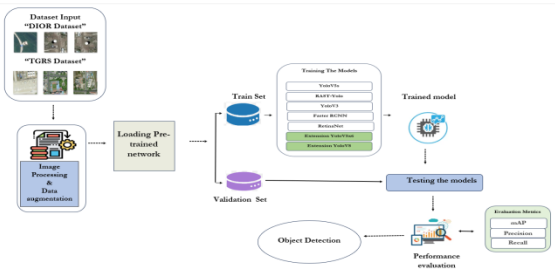
a) Proposed Work:

The proposed system introduces the RAST-YOLO (You Only Look Once with Region Attention and Swin Transformer) algorithm, designed specifically for remote sensing object detection tasks. By leveraging the Region Attention mechanism in conjunction with the Swin Transformer as its backbone architecture, the system aims to enhance feature extraction capabilities and improve detection accuracy in challenging remote sensing environments.

Extensive experimentation utilizing the DIOR and TGRS-HRRSD datasets serves as a comprehensive evaluation of the proposed algorithm's performance. Through this experimentation, the system evaluates the effectiveness of the RAST-YOLO[12] algorithm in accurately detecting objects of interest amidst complex backgrounds and varying target sizes.

By combining cutting-edge techniques such as Region Attention and Swin Transformer, the proposed system addresses the unique challenges posed by remote sensing object detection, including small data scales, complex backgrounds, and disparities in target appearances. Through rigorous experimentation and evaluation, the system aims to demonstrate the superiority of the RAST-YOLO algorithm in achieving state-of-the-art performance in terms of detection accuracy and efficiency, thereby contributing to advancements in remote sensing technology.

b) System Architecture:



ISSN 2454 - 5015

Fig1 Proposed Architecture

The system architecture begins with the input of two datasets: the DIOR Dataset and the TGRS Dataset, which serve as the basis for training and evaluating the object detection models. Image processing and data augmentation techniques are applied to enhance the diversity and quality of the training data, thereby improving the robustness of the models.

Pretrained networks, including YOLOv5s, RAST-YOLO[12], YOLOv3[32], Faster RCNN[18], and RetinaNet[14], are loaded into the system to leverage pre-existing knowledge and accelerate the training process. The datasets are split into train and validation sets to facilitate model training and evaluation.

The models are then trained using the training set, where they learn to identify and localize objects within the images. During training, the performance of each model is continuously evaluated on the validation set to monitor progress and prevent overfitting.

After training, the trained models undergo testing using a separate testing dataset to assess their performance in real-world scenarios. Performance evaluation metrics such as mean Average Precision (mAP), precision, and recall are calculated to quantify the effectiveness of each model in detecting objects accurately and efficiently.

Finally, the trained models are deployed for object detection tasks, where they analyze new input images and identify objects of interest based on the learned patterns and features. Through this systematic approach, the system architecture ensures the development of robust and accurate object detection models capable of effectively analyzing remote sensing images.

c) Dataset Collection:

Reading The Image: The DIOR dataset, released by Northwestern Polytechnic University, comprises 23,463 high-quality optical remote sensing images and 192,472 instance objects, spanning 20 common remote sensing categories such as airplanes, airports, bridges, dams, ships, and vehicles. This dataset offers an extensive range of object sizes, rich images, high inter-class similarity, and intra-class diversity. Instances within categories are unevenly distributed, presenting a diverse and challenging dataset for object detection tasks.

Similarly, the TGRS-HRRSD dataset, released by the University of Chinese Academy of Sciences, consists of 21,761 images and 55,740 instance objects obtained from Google Earth and Baidu maps. It features 13 categories including airplanes, bridges, harbors, and vehicles. Each category in TGRS-HRRSD contains approximately 4,000 instances, ensuring a balanced distribution across classes. This characteristic enhances the dataset's suitability for training and evaluating object detection models, making it valuable for research and development in remote sensing applications.

Plotting the Image: Exploring the dataset involves reading and visualizing the images to gain insights into the data's characteristics and structure. By plotting the images, researchers can examine the diversity of objects, variations in backgrounds, and overall image quality. This step is crucial for understanding the dataset's complexity and identifying potential challenges for object detection

ISSN 2454 - 5015

algorithms. Additionally, image plotting facilitates data preprocessing and augmentation, contributing to the development of robust and accurate object detection models trained on the DIOR and TGRS-HRRSD datasets.

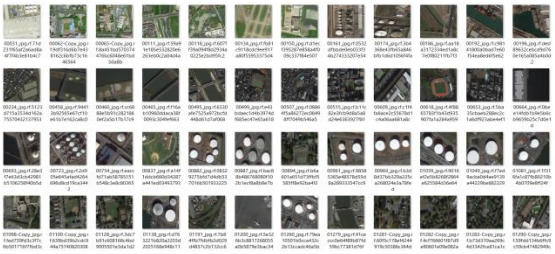


Fig 2 TGRS Dataset

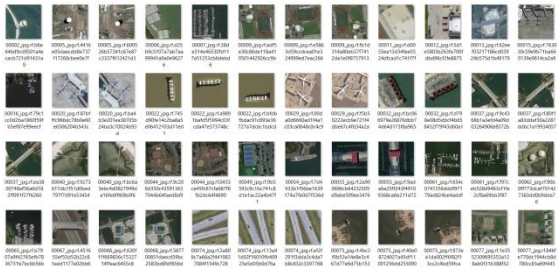


Fig 3 DIOR Dataset

d) Image Processing:

Converting to Blob Object: The first step in image processing involves converting the input image into a blob object, which is a specific format required by deep learning frameworks for model inference. This process typically involves resizing the image to match the input dimensions expected by the pre-trained model and normalizing pixel values.

Defining the Class and Declaring the Bounding Box: Next, the class labels for the objects present in the image are defined, along with the corresponding bounding box coordinates. This information is crucial for annotating the detected objects in the image and providing context for subsequent analysis.

Converting the Array to a NumPy Array: Once the class labels and bounding box coordinates are defined, the image and annotation data are converted into NumPy arrays for efficient manipulation and processing. NumPy arrays offer a versatile and high-performance data structure for handling multidimensional data, making them well-suited for image processing tasks.

Loading the Pre-Trained Model: The pre-trained object detection model is loaded into memory, allowing for inference on new input images. This step involves reading the network layers of the model and extracting the output layers responsible for predicting object classes and bounding box coordinates.

Image Processing: In this stage, various image processing techniques are applied to prepare the input image for object detection. This includes appending the image with its corresponding annotation file, converting the image from BGR to RGB color space, creating a mask to highlight regions of interest, and resizing the image to match the input dimensions expected by the pre-trained model.

Data Augmentation:Data augmentation techniques are employed to enhance the diversity and robustness of the training data. This may involve randomly transforming the image by applying geometric operations such as rotation, translation, and scaling. Data augmentation helps prevent overfitting and improves the generalization capabilities of the object detection model.

e) Algorithms:

ISSN 2454 - 5015

YOLOv5: YOLOv5 is an object detection algorithm that builds upon the You Only Look Once (YOLO) framework. It utilizes a single neural network to simultaneously predict bounding boxes and class probabilities for multiple objects within an image. YOLOv5 improves upon previous versions by introducing a more efficient architecture and training methodology, resulting in faster inference speeds and higher accuracy.

RAST YOLO: RAST YOLO (Region Attention Swin Transformer YOLO) is an object detection algorithm that integrates the Region Attention mechanism with the Swin Transformer architecture as the backbone network for feature extraction. By combining these components, RAST YOLO[12] aims to enhance feature extraction capabilities and improve detection accuracy, particularly in challenging remote sensing environments with complex backgrounds and small-scale targets.

YOLOv3: YOLOv3 is another variant of the YOLO object detection algorithm. It divides the input image into a grid and predicts bounding boxes and class probabilities for each grid cell. YOLOv3[32] introduces improvements such as feature pyramid networks and multi-scale predictions, resulting in better detection performance across different object scales.

Faster R-CNN: Faster R-CNN is a two-stage object detection algorithm that consists of a Region Proposal Network (RPN) for generating candidate object bounding boxes and a subsequent object detection network for refining the proposals and classifying objects. [18]It achieves high accuracy by leveraging region-based convolutional neural networks (R-CNN) and advances in region proposal techniques.

RetinaNet: RetinaNet is a one-stage object detection algorithm that addresses the class imbalance problem inherent in object detection tasks by introducing a focal loss function.[14] This loss function assigns higher weights to hard-to-detect examples during training, thereby improving the model's ability to focus on challenging cases. RetinaNet achieves state-of-the-art performance in terms of accuracy and efficiency.

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)

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

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.

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

ISSN 2454 - 5015

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.

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

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

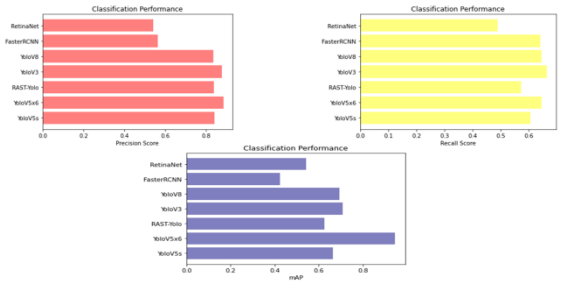


Fig 4 Comparison Graphs of DIOR Dataset

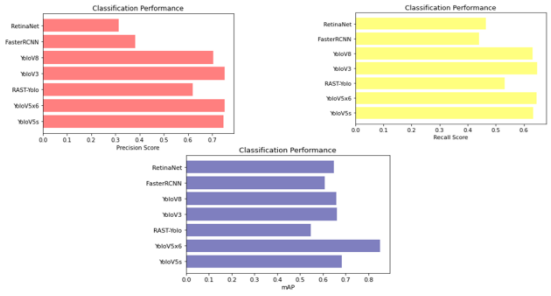


Fig 5 Comparison Graphs of TGRS Dataset

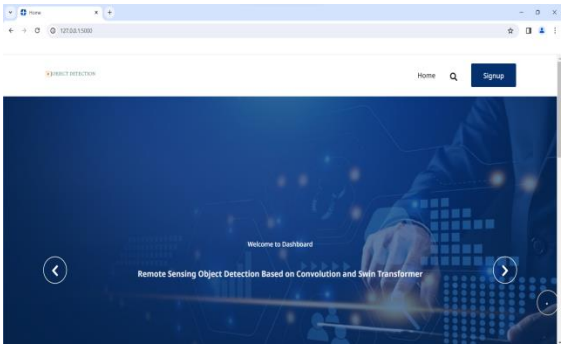


Fig 6 Home Page

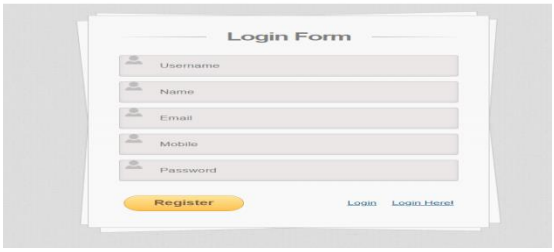


Fig 7 Registration Page

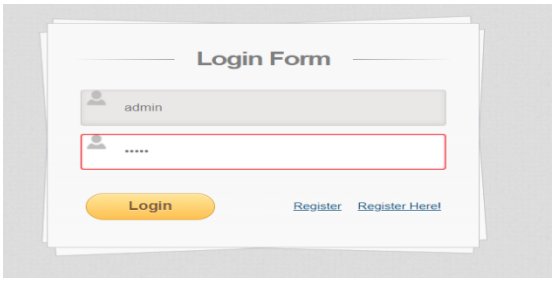


Fig 8 Login Page

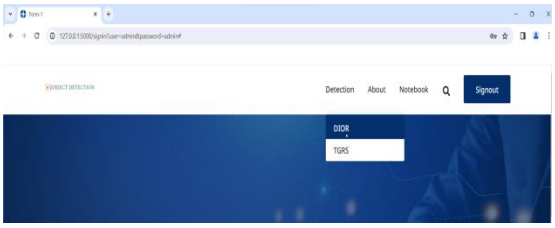


Fig 9 for DIOR

ISSN 2454 - 5015

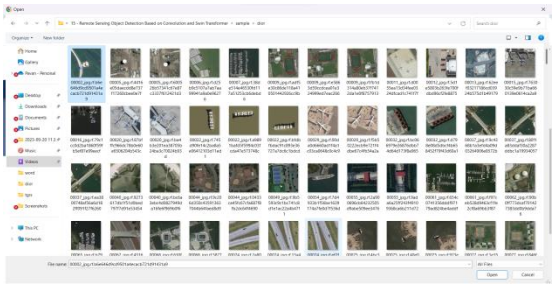


Fig 10 Upload Input Image

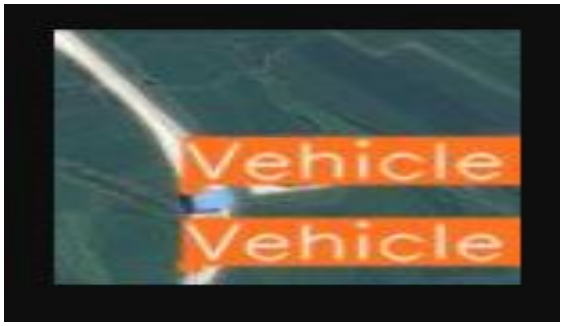


Fig 11 Final Outcome

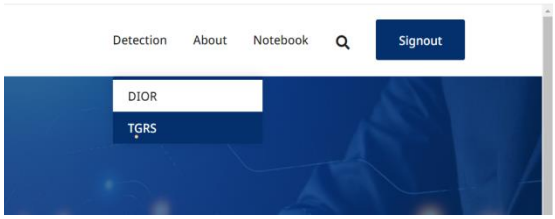


Fig 12 for TGRS

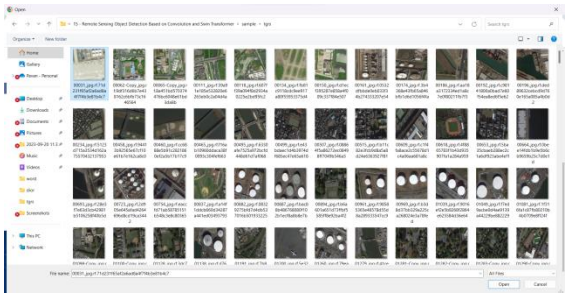


Fig 13 Upload Input Image

Fig 14 Final Outcome

5. CONCLUSION

In conclusion, the integration of YOLOv5[46], RAST YOLO[12], YOLOv3[32], Faster R-CNN[18], and RetinaNet[14] in this project offers a comprehensive exploration of object detection algorithms for remote sensing. Addressing challenges such as complex backgrounds, small-scale targets, and multi-scale targets, the proposed RAST YOLO algorithm demonstrates significant advancements by leveraging the Region Attention mechanism and Swin Transformer backbone. Additionally, the inclusion of the C3D module enhances the detection accuracy of multi-scale and small targets, further improving performance.

The project's findings underscore the importance of exploring diverse algorithmic approaches to identify the most suitable models for remote sensing object detection tasks. Through benchmarking against mainstream natural scene detection algorithms, RAST YOLO emerges as a promising solution, showcasing superior performance and adaptability. Furthermore, the extension algorithm YOLOv5x6 demonstrates state-of-the-art accuracy, reinforcing its efficacy in remote sensing applications.

ISSN 2454 - 5015

Overall, this project provides valuable insights and serves as a foundation for future research endeavors in remote sensing image interpretation. By guiding the development of robust models tailored to the unique challenges of aerial and satellite imagery, it contributes to advancements in remote sensing technology and its applications across various fields.

6. FUTURE SCOPE

The exploration of object detection algorithms for remote sensing in this project presents a broad range of opportunities for future research and development. Firstly, there is considerable potential for further refinement and optimization of the RAST YOLO algorithm. Fine-tuning hyperparameters, exploring alternative network architectures, and incorporating advanced attention mechanisms could enhance its performance, particularly in addressing specific challenges such as highly cluttered backgrounds or very small-scale targets. Secondly, the integration of more advanced feature extraction techniques, such as graph convolutional networks or self-attention mechanisms, holds promise for better capturing spatial relationships and context within remote sensing images. Expanding the dataset used for training and evaluation to include more diverse environments and object categories would allow for more comprehensive testing and validation of the algorithms across different scenarios and conditions, thereby improving their robustness and generalization capability. Furthermore, exploring real-time implementation and deployment of the algorithms on edge devices or embedded systems could enable on-the-fly object detection for applications such as autonomous vehicles, environmental monitoring drones, or disaster response systems. Overall, the future scope of this project lies in continued innovation and refinement of object detection algorithms for remote sensing, with a focus on improving accuracy, efficiency, and applicability in real-world scenarios.

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.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[17] R. Girshick, ‘‘Fast R-CNN,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, ‘‘Faster R-CNN: Towards realtime object detection with region proposal networks,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ‘‘Mask R-CNN,’’ Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ 2017, arXiv:1706.03762.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ 2018, arXiv:1810.04805.

[22] A. Radford and K. Narasimhan, ‘‘Improving language understanding by generative pre-training,’’ in Proc. Int. Conf. Learn. Represent. (ICLR), vol. 1, 2018, pp. 5–12.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16×16 words: Transformers for image recognition at scale,’’ 2020, arXiv:2010.11929.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, ‘‘End-to-end object detection with transformers,’’ Proc. 16th Eur. Conf. Comput. Vis., Glasgow, U.K.: Springer, Aug. 2020, pp. 213–229.

[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, ‘‘Deformable DETR: Deformable transformers for end-to-end object detection,’’ 2020, arXiv:2010.04159.

[26] M. Zheng, P. Gao, R. Zhang, K. Li, X. Wang, H. Li, and H. Dong, ‘‘End-to-end object detection with adaptive clustering transformer,’’ 2020, arXiv:2011.09315.

[27] Z. Dai, B. Cai, Y. Lin, and J. Chen, ‘‘UP-DETR: Unsupervised pre-training for object detection with transformers,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1601–1610.

[28] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, ‘‘Feature pyramid transformer,’’ in Proc. 16th Eur. Conf., Glasgow, U.K.: Springer, Aug. 2020, pp. 323–339.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ‘‘Swin transformer: Hierarchical vision transformer using shifted windows,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 9992–10002.

[30] Y. Xu, Y. Yang, and L. Zhang, ‘‘DeMT: Deformable mixer transformer for multi-task learning of dense prediction,’’ 2023, arXiv:2301.03461.

[31] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, ‘‘The Pascal visual object classes challenge: A retrospective,’’ Int. J. Comput. Vis., vol. 111, no. 1, pp. 98–136, 2015.

[32] A. Farhadi and J. Redmon, ‘‘YOLOv3: An incremental improvement,’’ in Proc. Comput. Vis. pattern Recognit., vol. 1804. Berlin, Germany: Springer, 2018, pp. 1–6.

[33] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, ‘‘Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation,’’ IEEE Trans. Multimedia, vol. 24, pp. 1735–1749, 2022.

[34] X. Wu, D. Hong, and J. Chanussot, ‘‘UIU-Net: U-Net in U-Net for infrared small object detection,’’ IEEE Trans. Image Process., vol. 32, pp. 364–376, 2023.

[35] H. Zhao, C. Wang, R. Guo, X. Rong, J. Guo, Q. Yang, L. Yang, Y. Zhao, and Y. Li, ‘‘Autonomous live working robot navigation with real-time detection and motion planning system on distribution line,’’ High Voltage, vol. 7, no. 6, pp. 1204–1216, Dec. 2022.

[36] T. Ye, C. Ren, X. Zhang, G. Zhai, and R. Wang, ‘‘Application of lightweight railway transit object detector,’’ IEEE Trans. Ind. Electron., vol. 68, no. 10, pp. 10269–10280, Oct. 2021.

[37] H. Wang, H. Pei, and J. Zhang, ‘‘Detection of locomotive signal lights and pedestrians on railway tracks using improved YOLOv4,’’ IEEE Access, vol. 10, pp. 15495–15505, 2022.

[38] K. Li, G. Cheng, S. Bu, and X. You, ‘‘Rotation-insensitive and contextaugmented object detection in remote sensing images,’’ IEEE Trans. Geosci. Remote Sens., vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[39] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, ‘‘Structure-guided feature transform hybrid residual network for remote sensing object detection,’’ IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5610713.

[40] Z. Tian, C. Shen, H. Chen, and T. He, ‘‘FCOS: Fully convolutional onestage object detection,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9626–9635.

[41] H. Lv, W. Qian, T. Chen, H. Yang, and X. Zhou, ‘‘Multiscale feature adaptive fusion for object detection in optical remote sensing images,’’ IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

[42] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ‘‘YOLOv4: Optimal speed and accuracy of object detection,’’ 2020, arXiv:2004.10934.

[43] X. Yang, J. Zhao, H. Zhang, C. Dai, L. Zhao, Z. Ji, and I. Ganchev, ‘‘Remote sensing image detection based on YOLOv4 improvements,’’ IEEE Access, vol. 10, pp. 95527–95538, 2022.

[44] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, ‘‘CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection,’’ Remote Sens., vol. 13, no. 5, p. 847, Feb. 2021.

[45] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, ‘‘A contextual bidirectional enhancement method for remote sensing image object detection,’’ IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 4518–4531, 2020.

Dataset Link:

https://roboflow.com/convert/labelbox-json-to-yolov5-pytorch-txt