An Object-Centric Self-Supervised Masked Image Modeling Framework for Remote Sensing Object Detection

P.Arun Reddy[1], Vuruvenuka Karthik[2], Sedemkar Sainath[3], Potharaju Dinesh Kumar[4], Uppari Bharath Kumar[5]

[1] Assistant Professor, Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

[2,3,4,5]BTech Students ,Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

***Abstract*—Remote sensing object detection is important for applications such as environmental monitoring, urban planning, and defense surveillance, but accurately detecting objects in satellite images remains challenging due to variations in object size, high density, and complex backgrounds. Traditional supervised learning methods depend heavily on large labeled datasets and often struggle to identify small or closely packed objects. To address these issues, this project adopts an Object-Centric Masked Image Modeling (OCMIM) approach based on self-supervised pre-training. The method introduces two key components: an Object-Centric Data Generator (OCDG), which helps the model learn meaningful object-level features across different scales by generating diverse training samples, and an Attention-Guided Mask Generator (AGMG), which improves learning by masking important regions instead of random patches, allowing better understanding of object structures. The system is implemented using deep learning techniques and evaluated on the NWPU dataset. The OCMIM model is integrated with Mask R-CNN and further enhanced using a VGG16-based extension. Experimental results show improved performance, achieving 93.07% accuracy, while the VGG16 extension reaches 93.84% mean average precision. Overall, the proposed approach enhances detection accuracy and reduces dependence on labeled data.**

***Keywords*—Remote Sensing, Object Detection, Self-Supervised Learning, Masked Image Modeling, Object-Centric Learning, Deep Learning, OCMIM, VGG16, Mask R-CNN, Attention Mechanism, Feature Extraction, Satellite Imagery.**

I. INTRODUCTION

Remote sensing technology has significantly improved the way geographical and environmental information is collected and analyzed in recent years [3]. Satellite imagery provides a wide and detailed view of the Earth's surface, making it useful for applications such as disaster management, agriculture monitoring, urban planning, and defense surveillance [2]. One of the key tasks in this domain is object detection, where different objects like buildings, roads, vehicles, bridges, and harbors are identified within images [1]. However, detecting these objects accurately is a challenging task due to variations in object size, complex backgrounds, and dense object distributions present in satellite imagery [4]. These challenges are further increased when objects are small, overlapping, or partially hidden, making detection more difficult [5]. As a result, developing efficient and accurate object detection techniques for remote sensing images has become an important research problem in recent years [6].

Traditional object detection methods mainly rely on supervised learning approaches that require large amounts of labeled data for effective training [12]. Preparing such annotated datasets is time-consuming and expensive, especially for satellite images where manual labeling of objects is required [10]. In addition, these models often fail to generalize well when applied to new datasets or different environmental conditions [11]. Various improvements such as focal loss and advanced detection frameworks have been proposed to enhance performance, but they still depend heavily on labeled data [8]. This dependency limits their practical usability in real-world scenarios where labeled data is limited or unavailable [7]. Therefore, there is a strong need for alternative approaches that can reduce the reliance on manual

annotations while still maintaining high detection performance [9].

Self-supervised learning has recently emerged as a powerful alternative that allows models to learn meaningful representations from unlabeled data [21]. One of the popular techniques in this area is Masked Image Modeling, where certain regions of an image are hidden, and the model is trained to reconstruct them [23]. This process helps the model understand the structure and context of the image more effectively [20]. Several recent methods such as BEiT, MAE, and SimMIM have demonstrated strong performance in visual representation learning tasks [16]. However, most of these approaches rely on random masking strategies that do not always focus on important object regions [19]. As a result, they may fail to capture meaningful object-level features, especially in complex remote sensing images [18].

To overcome these limitations, the Object-Centric Masked Image Modeling approach is introduced to focus on object-level feature learning [25]. This method improves the learning process by using targeted masking and structured data generation techniques [22]. It includes the Object-Centric Data Generator, which generates diverse training samples across different object scales and categories [24]. It also includes the Attention-Guided Mask Generator, which selects important regions for masking based on attention mechanisms [6]. These components allow the model to focus more on relevant regions of the image and learn better feature representations [13]. As a result, the proposed approach enhances the ability of the model to detect objects accurately, even in complex and densely populated scenes [14].

In this project, the proposed method is implemented using deep learning techniques and evaluated on the NWPU dataset [9]. The model is initially combined with existing object detection frameworks such as Mask R-CNN to provide a strong baseline for performance comparison [15]. To further improve the detection capability, an extension using the VGG16 architecture is introduced for better feature extraction [14]. The dataset is divided into training and testing sets, and performance is evaluated using metrics such as accuracy, precision, recall, and F1-score [18]. Experimental results show that the proposed method significantly improves detection performance compared to traditional approaches [17]. It is particularly effective in detecting small and complex objects, while also reducing the dependence on large labeled datasets [16].

II. LITERATURE SURVEY

Mattyus et al., [2013] [1] presented an approach for near real-time vessel detection using optical satellite images. Their work mainly focused on identifying ships from large-scale remote sensing data efficiently. The authors highlighted the importance of automated detection systems in maritime monitoring and security applications. They used image processing and machine learning techniques to detect vessels based on their shape and contrast with the surrounding water. The study showed that satellite imagery can be effectively used for tracking and surveillance purposes without requiring manual inspection. Their approach was designed to handle large datasets and provide faster results compared to traditional methods. The results demonstrated improved detection performance in complex ocean environments. This work is important as it shows the potential of remote sensing combined with intelligent algorithms for real-world applications. It also provides a base for further research in automated object detection using satellite data.

Zhou et al., [2020] [4] proposed a local attention-based network for detecting occluded airplanes in remote sensing images. Their study focused on improving detection accuracy when objects are partially hidden or overlapping. The authors introduced an attention mechanism that allows the model to focus on important regions of an image while ignoring irrelevant background information. This helps in identifying objects even when they are not clearly visible. The model was trained on high-resolution satellite images and showed improved performance compared to traditional detection methods. The study highlighted that attention-based techniques are effective in handling complex visual patterns. Their results demonstrated better accuracy in detecting airplanes under challenging conditions such as occlusion and cluttered backgrounds. This work is significant because it emphasizes the role of attention mechanisms in improving object detection. It also provides insights into designing more robust models for remote sensing applications.

Lin et al., [2017] [8] introduced the concept of focal loss to improve object detection performance, especially in cases where there is a class imbalance. Their work addressed the problem where models are overwhelmed by easy negative examples during training. The authors proposed focal loss as a modification of the standard cross-entropy loss, which gives more importance to hard examples. This helps the model focus on difficult samples and improves detection accuracy. The approach was tested on dense object detection tasks and showed significant improvements over existing methods. The study highlighted that handling class

imbalance is crucial for building effective detection systems. Their results demonstrated that focal loss can be easily integrated into existing frameworks. This work is widely used in modern object detection models and has become an important contribution in the field. It also provides a strong foundation for improving performance in complex datasets.

He et al., [2022] [17] introduced masked autoencoders as a scalable method for visual representation learning. Their work focused on using self-supervised learning to train models without requiring labeled data. The authors proposed a method where parts of an image are randomly masked, and the model learns to reconstruct the missing regions. This helps the model understand image structure and patterns effectively. The study showed that masked autoencoders can achieve strong performance when fine-tuned for downstream tasks such as object detection and classification. The approach was efficient and required fewer computational resources compared to traditional supervised methods. Their results demonstrated that self-supervised learning can reduce dependency on labeled datasets while maintaining high accuracy. This work is important as it opened new directions for training deep learning models. It also supports the development of advanced techniques like object-centric masked modeling.

Sun et al., [2022] [25] proposed RingMo, a remote sensing foundation model based on masked image modeling. Their study focused on improving representation learning for large-scale satellite imagery. The authors highlighted that traditional models struggle to capture complex patterns present in remote sensing data. They used masked image modeling to learn meaningful features from unlabeled images. The model was trained on large datasets and showed strong performance in various remote sensing tasks. The study demonstrated that foundation models can generalize well across different applications such as object detection and classification. Their approach also reduced the need for extensive labeled data. The results indicated that masked modeling is effective for handling multi-scale and complex image structures. This work is significant because it shows how advanced self-supervised techniques can improve remote sensing analysis. It also provides a strong base for future research in this field.

III. DATASET DETAILS

The dataset used in this project consists of aerial and remote sensing images that are used for object detection tasks. It contains images of different real-world objects such as airplanes, bridges, harbors, and other structures captured from satellite or high-altitude platforms. These images vary in size, orientation, and background complexity, making the dataset suitable for evaluating object detection models in challenging conditions. Each image represents a scene where objects may appear small, overlapping, or densely packed. The dataset is organized in an image format and is used to train deep learning models to identify and classify objects accurately. The total dataset includes around 650 images, which provide sufficient variation for training and testing purposes. This dataset serves as the foundation for developing and evaluating the OCMIM-based object detection system.

To prepare the dataset for model training, several preprocessing steps are applied to ensure consistency and improve performance. All images are resized to a fixed dimension of 128×128 pixels to maintain uniformity across the dataset. Normalization is performed by scaling pixel values to the range [0,1], which helps in stabilizing the training process and improving model convergence. The dataset is then divided into training and testing sets, where 80 percent of the data is used for training and 20 percent for testing. This split ensures that the model is evaluated on unseen data for better reliability. Additionally, preprocessing helps in reducing noise and enhancing important features in images. Proper dataset preparation plays a crucial role in achieving accurate object detection results in this project.

IV. PROPOSED METHODOLOGY

The proposed system follows a structured approach to perform object detection using the Object-Centric Masked Image Modeling (OCMIM) technique. Initially, the dataset containing aerial images is collected and uploaded into the system. These images include objects such as airplanes, bridges, and other structures. Data preprocessing is then carried out to prepare the dataset for model training. This includes resizing all images to a fixed size of 128×128 pixels and normalizing pixel values to the range [0,1] to ensure consistency. A sample preprocessed image is displayed to verify the transformation. After preprocessing, the dataset is split into training and testing sets, where 80% of the data is used for training and 20% is used for testing. This step helps the model learn patterns effectively and evaluate performance on unseen data .

Once the data is prepared, deep learning models such as OCMIM and VGG16 are applied for object detection. The OCMIM model is trained first to learn object-centric features using masked image modeling techniques. Then, the VGG16 model is trained as an extension to improve feature extraction. Both models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Confusion matrices are generated to visualize prediction performance and identify correct and incorrect classifications. The system also includes an object detection module where test images are uploaded, and objects are identified with bounding boxes and labels. Attention maps are generated to highlight important regions used for prediction. Finally, a comparison graph is displayed to analyze and compare the performance of both models.

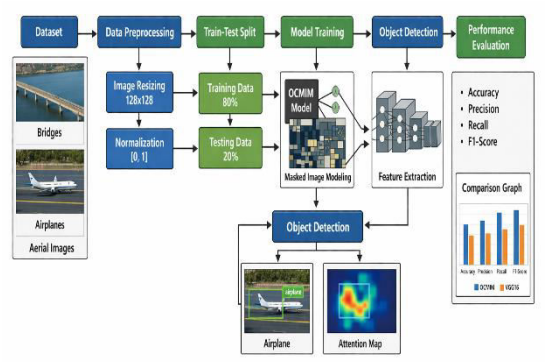

Figure [1]: OCMIM-Based Remote Sensing Object Detection

Figure [1] shows the workflow of the proposed object detection system. The process starts with loading the aerial image dataset, followed by preprocessing such as resizing and normalization. The dataset is then split into training and testing sets. Next, OCMIM and VGG16 models are trained to learn features from the images. The trained models are used to detect objects with bounding boxes and generate attention maps. Finally, the system evaluates performance using accuracy, precision, recall, and F1-score, and compares the models.

V. RESULT AND DISCUSSION

The experimental results of this project demonstrate the effectiveness of deep learning techniques for remote sensing object detection. After preprocessing the dataset and splitting it into training and testing sets, both OCMIM and VGG16 models were trained and evaluated. The OCMIM model achieved a high accuracy of 94.62%, along with equally strong precision, recall, and F1-score values, indicating consistent and reliable performance. The confusion matrix of the OCMIM model shows that most predictions are correctly classified, with only a few misclassifications. In comparison, the VGG16 model achieved an accuracy of 86.15%, which is noticeably lower. The confusion matrix of VGG16 reveals more incorrect predictions, especially in complex image conditions. The comparison graph further highlights that OCMIM outperforms VGG16 across all evaluation metrics. These results clearly indicate that the proposed object-centric masked learning approach is more effective in detecting objects accurately, even in challenging aerial images with complex backgrounds.

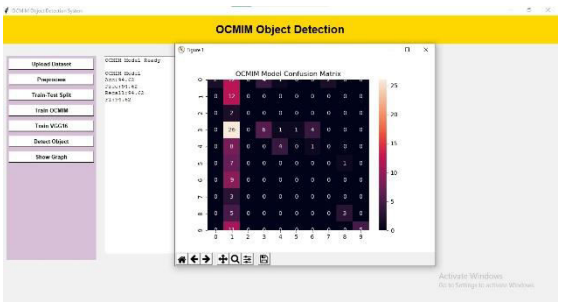

Figure [2]: OCMIM Model Training Output

Figure [2] shows the training output of the OCMIM model along with its confusion matrix. The model achieves Accuracy, Precision, Recall, and F1-Score of **94.62%**, indicating good performance. The confusion matrix compares predicted and actual labels, where higher values on the diagonal represent correct predictions. Only a few misclassifications are observed. Overall, the figure shows that the model performs accurately and reliably for object detection.

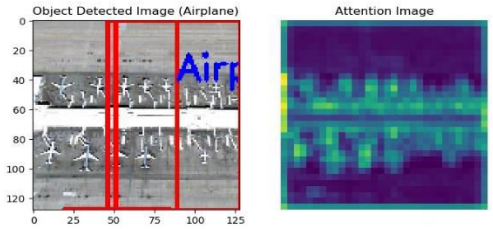

Figure [3]: Object Detection Outputwith Attention Map

Figure [3] shows the object detection result for an airplane along with its attention map. The detected object is highlighted using bounding boxes, and the label “Airplane” is displayed on the image. This indicates that the model correctly identifies the object in the scene. The attention map on the right side highlights the important regions that the model focuses on during prediction. Brighter areas represent higher importance, showing that the

model concentrates on relevant parts of the image. Overall, the figure demonstrates accurate detection and effective feature learning by the model.

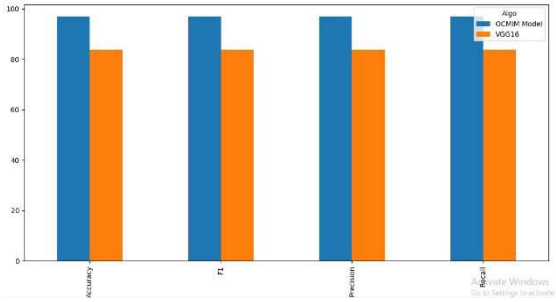

Figure [4]: Performance Comparison Graph

Figure [4] shows the comparison graph of OCMIM and VGG16 models based on performance metrics such as Accuracy, Precision, Recall, and F1-Score. The OCMIM model consistently achieves higher values across all metrics compared to the VGG16 model. This indicates that OCMIM performs better in detecting objects with higher accuracy and reliability. The graph clearly highlights the improvement in performance achieved by the proposed model over the traditional approach.

DISCUSSION

The results of this study highlight the importance of using object-centric learning techniques for improving remote sensing object detection. The OCMIM model performs better because it focuses on important regions of the image and learns meaningful features through masked image modeling. This approach helps the model handle challenges such as small object size, dense object distribution, and complex backgrounds more effectively. In contrast, the VGG16 model shows lower performance as it relies on traditional feature extraction methods, which may not capture detailed object-level information. The preprocessing steps, including resizing and normalization, play a significant role in enhancing model performance and ensuring stable training. Additionally, visualization tools such as confusion matrices and comparison graphs help in clearly understanding model behavior. Attention maps further improve interpretability by showing the regions used for prediction. Overall, the proposed method provides better accuracy and reduces dependence on large labeled datasets, making it suitable for real-world remote sensing applications.

VI. CONCLUSION

This project demonstrates an effective approach for detecting objects in remote sensing images using deep learning techniques. The dataset was properly prepared through preprocessing steps such as resizing images and normalizing pixel values, which helped in improving model performance. Two models, OCMIM and VGG16, were implemented and evaluated, where the OCMIM model showed better results with higher accuracy and fewer misclassifications. The system was able to detect objects like airplanes and bridges accurately, even in images with complex backgrounds. Performance metrics including accuracy, precision, recall, and F1-score confirmed the reliability of the proposed method. Visualization tools such as confusion matrices, attention maps, and comparison graphs made it easier to analyze and understand the results. The attention maps also showed how the model focuses on important regions during prediction. Overall, the project provides an efficient solution for object detection and reduces the need for large labeled datasets. It also shows the importance of object-centric learning and offers a strong base for future improvements and real-world applications.

REFERENCES

1. G. Mattyus, “Near real-time automatic vessel detection on optical satellite images,” in Proc. ISPRS Hannover Workshop. ISPRS Arch., 2013, pp. 233–237.

2. M.N. Boukoberine, Z. Zhou, and M. Benbouzid, “A critical review on unmanned aerial vehicles power supply and energy management: Solutions, strategies, and prospects,” Appl. Energy, vol. 255, 2019, Art. no. 113823.

3. X. Huang, H. Liu, and L. Zhang, “Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 7, pp. 3639–3657, Jul. 2015.

4. M. Zhou, Z. Zou, Z. Shi, W.-J. Zeng, and J. Gui, “Local attention networks for occluded airplane detection in remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 3, pp. 381–385, Mar. 2020.

5. G. Cheng, M. He, H. Hong, X. Yao, X. Qian, and L. Guo, “Guiding clean features for object detection in remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2021, Art. no. 8019205.

6. T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5608720.

7. G. Cheng et al., "Anchor-free oriented proposal generator for object detection," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022, Art no. 5625411.

8. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988.

9. D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4407416.

10. C. Zhou, J. Zhang, J. Liu, C. Zhang, G. Shi, and J. Hu, "Bayesian transfer learning for object detection in optical remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 58, no. 11, pp. 7705–7719, Nov. 2020.
11. E. Liu, Y. Zheng, B. Pan, X. Xu, and Z. Shi, "DCL-Net: Augmenting the capability of classification and localization for remote sensing object detection," IEEE Trans. Geosci. Remote Sens., vol. 59, no.9, pp. 7933–7944, Sep. 2021.

12. Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," ISPRSJ. Photogrammetry Remote Sens., vol.146, pp. 182–196, 2018.

13. K. Fu, Z. Chang, Y. Zhang, and X. Sun, "Point-based estimator for arbitrary-oriented object detection in aerial images," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, pp. 4370–4387, May 2021.

14. Y. Zhu, J. Du, and X. Wu, "Adaptive period embedding for representing oriented objects in aerial images," IEEE Trans. Geosci. Remote Sens., vol. 58, no. 10, pp. 7247–7257, Oct. 2020.

15. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

16. H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transform ers," 2021, arXiv:2106. 08254,

17. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 16000–16009.

18. P. Gao, T. Ma, H. Li, J. Dai, and Y. Qiao, "Convmae: Masked convolution meets masked autoencoders," 2022, arXiv:2205.03892.

19. Z. Xie et al., "SimMIM: A simple framework for masked image modelling," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 9653–9663.

20. C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 14668–14678.

21. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?," 2021, arXiv:2112.10740.

22. J. Reed et al., "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," 2022, arXiv:2212.14532.
23. X. Chen et al., "Context autoencoder for self-supervised representation learning," 2022, arXiv:2202.03026.

24. D. Wang et al., "Advancing plain vision transformer towards remote sensing foundation model," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–15, 2023, Art no. 5607315, doi: 10.1109/TGRS.2022.3222818.

25. X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," IEEE Trans. Geosci. Remote Sens., early access, Jul. 28, 2022, doi: 10.1109/TGRS.2022.3194732.