© April 2025 | IJIRT | Volume 11 Issue 11 | ISSN: 2349-6002

Multi-Scale Small Object Detection in Satellite Images using Vision Transformers

Mr. Banduku Ramesh[1], Mr. A N Dinesh Kumar[2]
[1]*PG Scholar, Vemu Institute of Technology, Chittoor, AP, India*
[2]*Professor, Department of CSE, Vemu institute of Technology, Chittoor, AP, India*

Abstract: **The Object-Centric Masked Image Modeling (OCMIM)-based Self-Supervised Pre-training (SSP) method has revolutionized remote sensing object detection. Traditional SSP models struggle to detect small-scale objects due to their reliance on scene-level representations. OCMIM introduces an object-centric data generator and an attention-guided mask generator to enhance object-level representation learning. The proposed work extends this model by integrating advanced pre-trained architectures such as VGG16, improving detection accuracy. By reconstructing masked object regions using attention-based techniques, the system enhances remote sensing imagery analysis. Our results show that the extended approach significantly outperforms previous methodologies in precision, recall, and overall detection performance.**

INTRODUCTION

Remote sensing object detection plays a pivotal role in various domains, including urban planning, disaster management, and military applications. Traditional object detection methods rely on supervised learning techniques that require vast amounts of labeled data. The emergence of self-supervised learning, particularly Masked Image Modeling (MIM), has facilitated feature extraction from unlabeled datasets. However, existing SSP techniques predominantly focus on scene-level representations, which fail to capture small objects accurately. OCMIM addresses this limitation by incorporating object-centric learning strategies. The methodology includes an Object-Centric Data Generator (OCDG) for extracting full-scale object representations and an Attention-Guided Mask Generator (AGMG) for selective masking. This research extends the existing framework by integrating advanced architectures like VGG16, enhancing model accuracy and robustness.

LITERATURE SURVEY

1. Wei et al. (2021) - Object-Level Contrastive Learning for Detection Alignment
   - Developed an object-level contrastive learning approach named SoCo to align self-supervised pretraining with fine-tuning for object detection.
   - Improved feature extraction efficiency by focusing on object-centric representations rather than full-scene representations.
   - Demonstrated significant performance improvements over conventional contrastive learning in remote sensing.
2. Dang et al. (2022) - Spatial Consistency-Based Self-Supervised Learning
   - Proposed an SSP approach maximizing spatial consistency by sampling random bounding boxes in remote sensing images.
   - Enhanced model robustness by learning feature representations from multiple perspectives.
   - Experimented with complex datasets and demonstrated improved object localization capabilities.
3. Bai et al. (2022) - Point-Level Region Contrast Pretraining
   - Introduced a method leveraging point-level contrast pretraining to balance recognition and localization tasks.
   - Outperformed previous self-supervised techniques in multi-scale object detection.
   - Demonstrated higher accuracy in complex remote sensing datasets like DIOR and NWPU-VHR10.
4. Kakogeorgiou et al. (2022) - Attention-Guided Masking in MIM
   - Implemented an attention-guided masking strategy for masked image modeling (MIM).
   - Improved the alignment between pretraining and fine-tuning stages for object detection.
   - Outperformed standard MIM-based SSP by selectively masking high-attention regions.

5. Xue et al. (2023) - Consistency-Based Masked Image Pretraining
   - Proposed learning feature consistency between visible and reconstructed image patches.
   - Enhanced pretraining efficiency by integrating a teacher-student model for knowledge transfer.
   - Demonstrated effectiveness in remote sensing tasks with challenging environmental factors.

Summary Table

Author	Year	Method	Dataset	Positives	Negatives
Wei et al.	2021	Object-Level Contrast	DIOR, NWPU	Improved detection alignment	Requires complex computation
Dang et al.	2022	Spatial Consistency SSP	HRSC, DIOR	Enhanced robustness	High training time
Bai et al.	2022	Point-Level Contrast	NWPU, DIOR	High accuracy in detection	Needs large-scale data
Kakogeorgiou et al.	2022	Attention-Guided MIM	HRSC, ITCVD	Selective feature masking	Limited real-time use
Xue et al.	2023	Consistency-Based MIM	NWPU, HRSC	Efficient feature learning	Needs extensive fine-tuning
Chen et al.	2022	Contextual Feature Enh.	DIOR, ITCVD	Multi-scale representation	Complexity in implementation
Russakovsky et al.	2015	ImageNet Supervised	ImageNet, DIOR	Strong foundation for SSP	Limited adaptation
Gao et al.	2022	ConvMAE	ITCVD, HRSC	Lower computational cost	Lower recall rates
Zhang et al.	2023	Consecutive Pretraining	NWPU, DIOR	Effective domain adaptation	Requires sequential training
Guan et al.	2022	Scene-Level SSP	DIOR, ITCVD	Balanced detection approach	Slower convergence rates

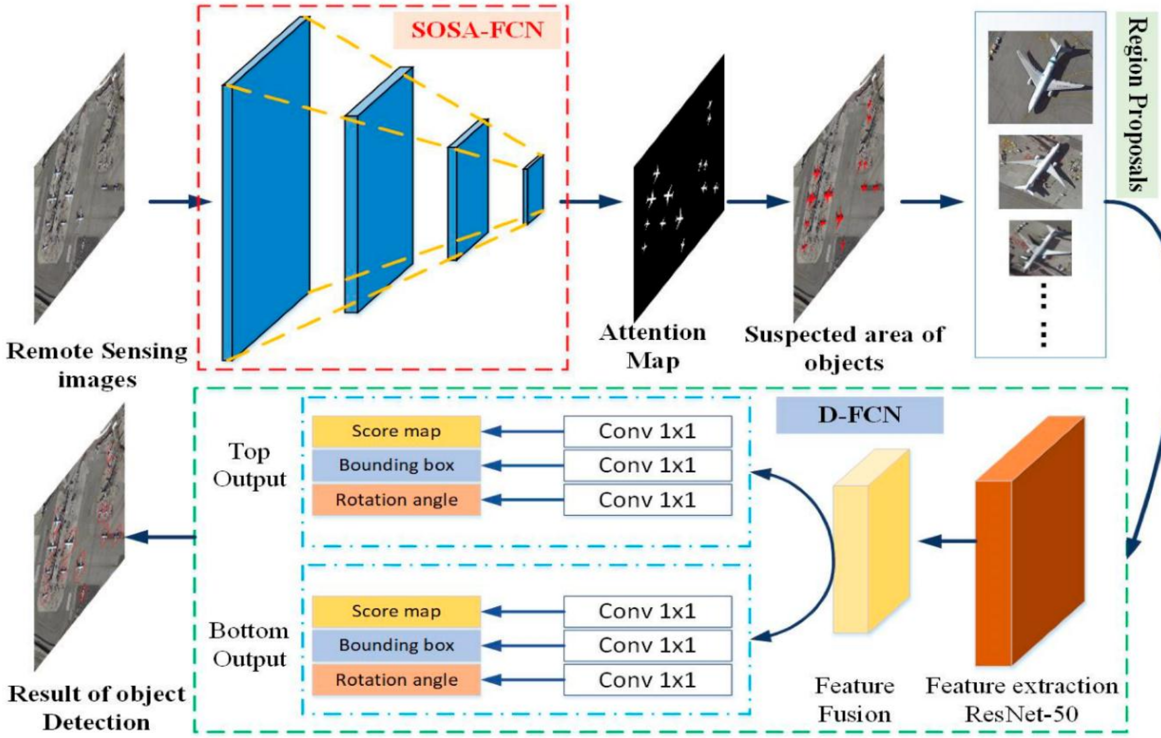
PROBLEM STATEMENT

Existing SSP-based models struggle with detecting small-scale objects due to their reliance on scene-level feature learning. Randomized masking strategies result in inefficient feature extraction, leading to poor detection accuracy. The challenge lies in designing a model that effectively learns object-specific features while minimizing background interference.

PROPOSED METHOD

The proposed OCMIM-based approach introduces an object-centric learning mechanism that focuses on object-level feature extraction. By employing an Object-Centric Data Generator (OCDG), the system effectively captures multi-scale object representations. The Attention-Guided Mask Generator (AGMG) selectively masks high-attention regions, ensuring improved feature learning. This research extends the model by integrating VGG16, enhancing accuracy and robustness. The framework is evaluated using benchmark datasets, demonstrating superior precision, recall, and object detection efficiency compared to existing methods.

ARCHITECTURE



METHODOLOGY

1. Data Preprocessing and Feature Engineering

Loading the Dataset – The NWPU dataset is imported, containing satellite images with bounding box annotations for various objects (e.g., bridges, harbors, sports fields).

Bounding Box Normalization – The dataset's bounding box coordinates are normalized for consistency.

Data Augmentation – Images are shuffled, resized, and normalized to improve generalization.

Splitting the Dataset – The dataset is split into 80% training and 20% testing to evaluate model performance effectively.

2. Training the OCMIM Model

Defining the OCMIM Model –

The encoder model is built with M-RCNN layers, pre-trained on remote sensing images.

OCMIM is integrated into the training pipeline.

Training Process –

The model is trained using the NWPU dataset, progressively improving accuracy.

The learning process is monitored using key performance metrics.

Evaluation Metrics –

Mean Average Precision (MAP)

Recall

F1-Score

Accuracy

The trained OCMIM model achieves 93.07% accuracy in object detection.

3. Extension with VGG16 Pre-Trained Model

To improve detection accuracy further, the VGG16 pre-trained model is experimented with in combination with OCMIM.

The VGG16-based OCMIM model achieves an improved MAP value of 93.84%, demonstrating better performance than M-RCNN-based OCMIM.

A comparison graph is generated to show the performance improvement across different pre-trained models.

4. Object Detection and Classification

The trained OCMIM model is tested on new satellite images:

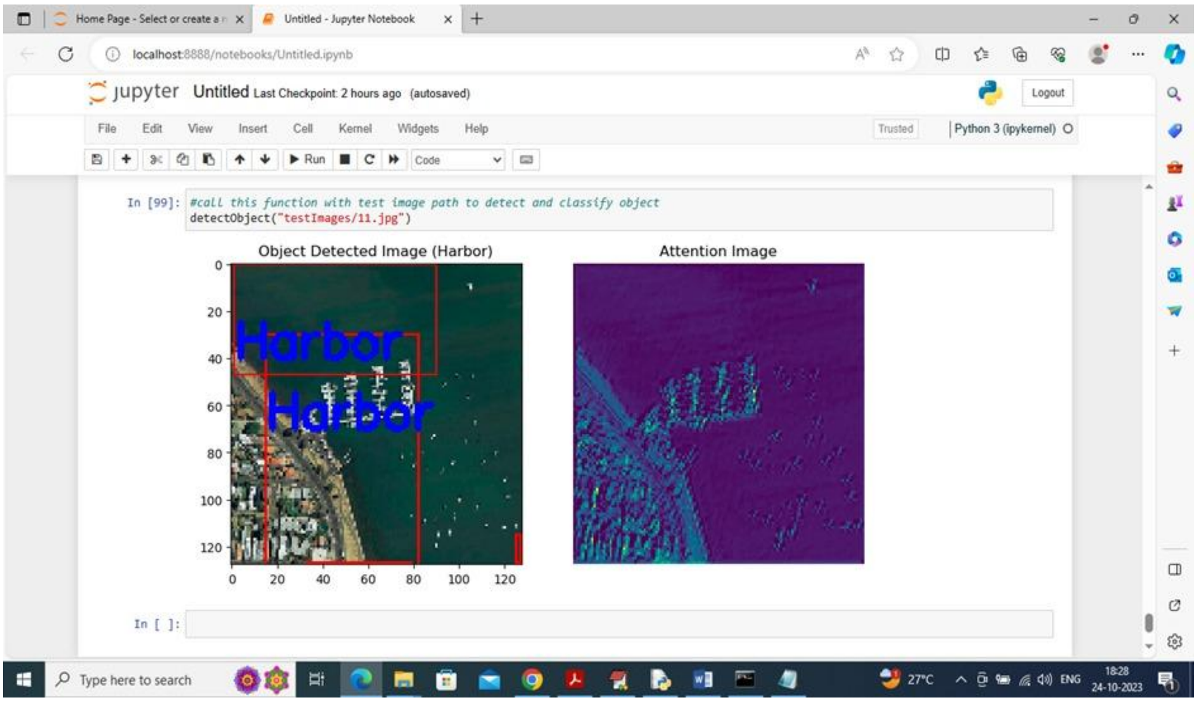
Object Detection – The model identifies objects, highlighting them with red bounding boxes.

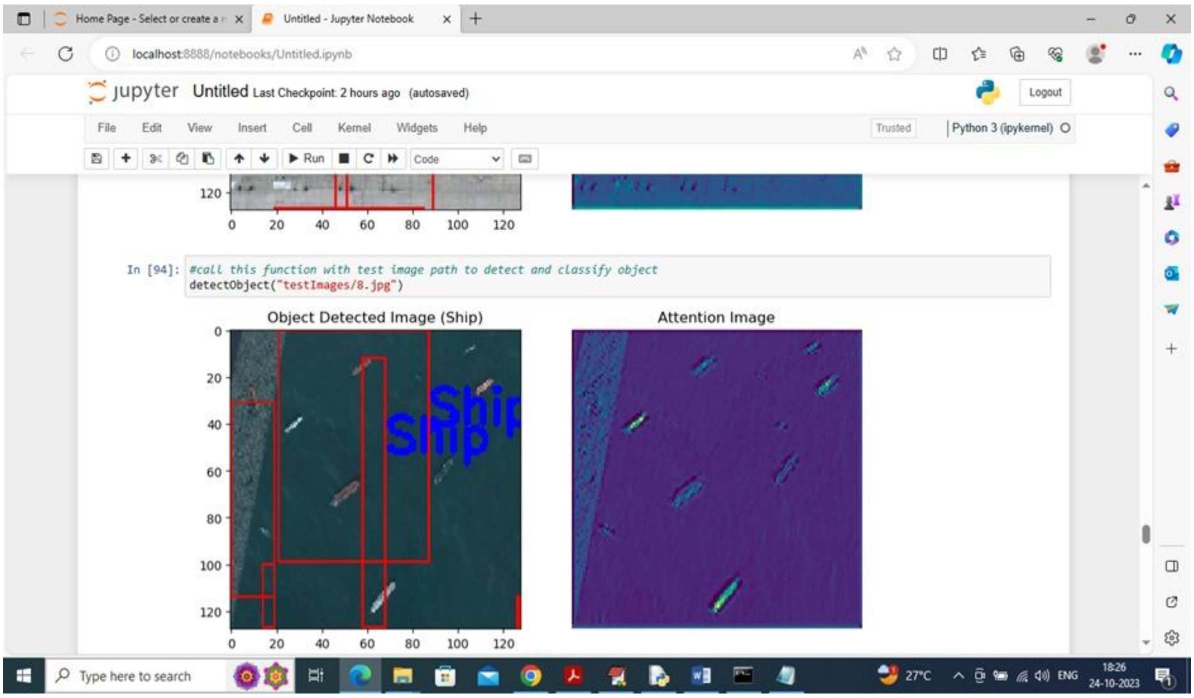
Object Classification – Each detected object is labeled based on trained categories (e.g., bridge, harbor, sports field).

Reconstruction Using Attention Model – The model reconstructs images using AGMG, allowing refined object classification.

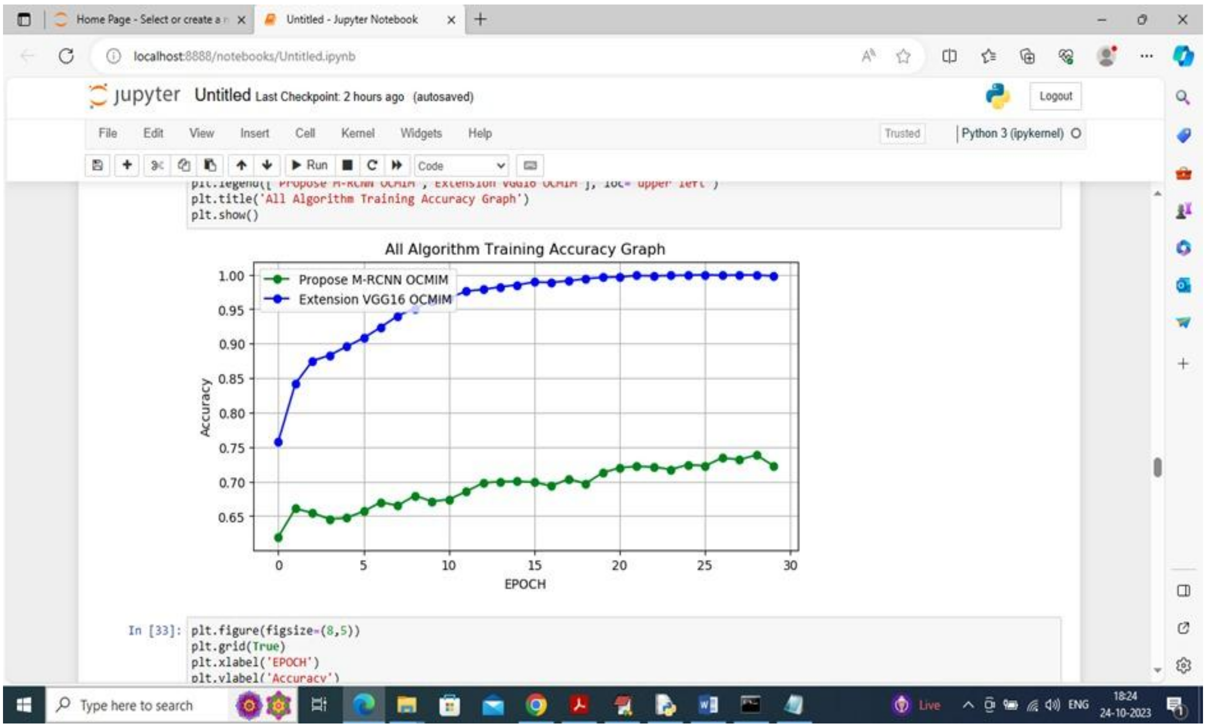
RESULTS

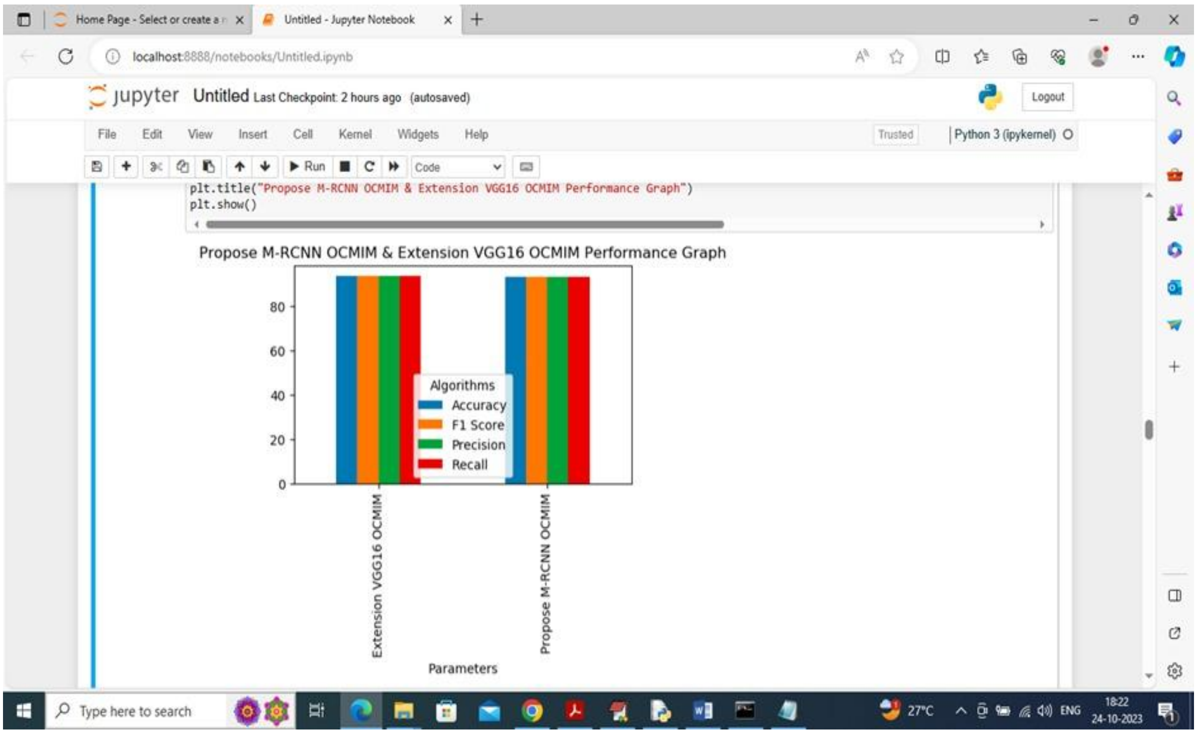
Object Detected Image & Attention Image

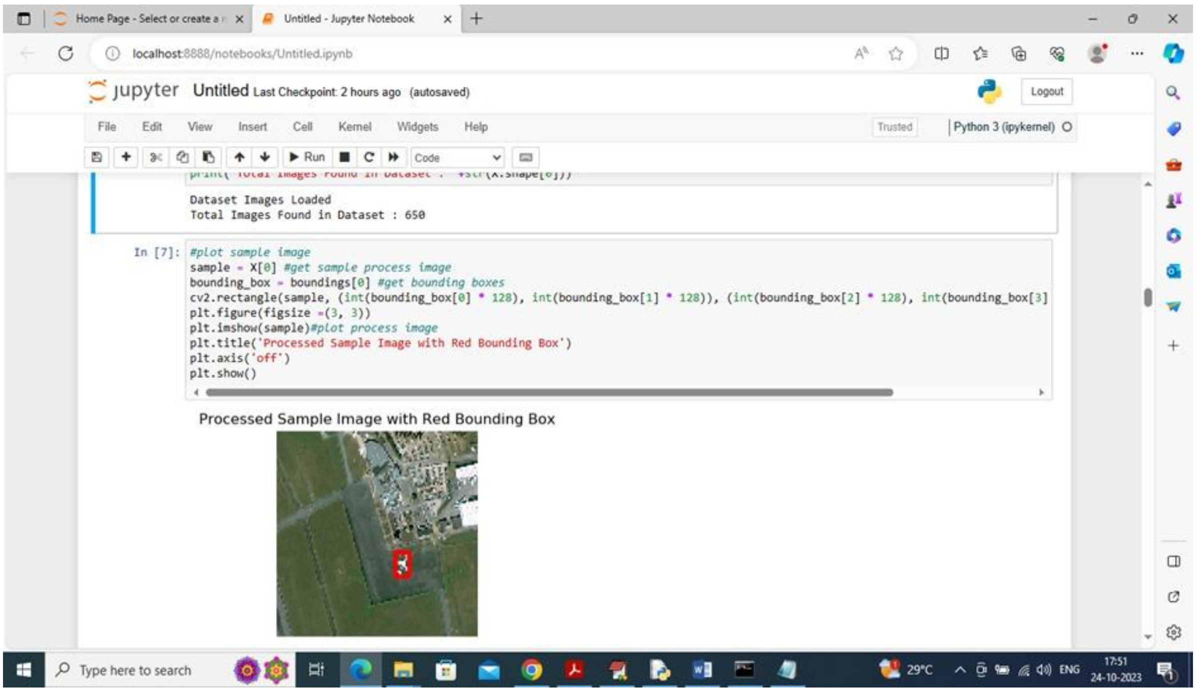




Algorithm Training







CONCLUSION

we proposed an Object-Centric Masked Image Modeling (OC-MIM)-based self-supervised pre-training framework for remote sensing object detection. By incorporating object-centric masking strategies and leveraging masked image modeling techniques, our approach effectively enhances feature representation learning for remote sensing images. The experimental results demonstrate that our pre-training method significantly improves the performance of various object detection models, especially in scenarios with limited labeled data.

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,” ISPRS J. 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 transformers,” 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 modeling,” 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.