Interactive Masked Image Modeling for Multimodal Object Detection in Remote Sensing

Minh-Duc VU, Zuheng MING, Fangchen FENG, Bissmella BAHADURI, Anissa MOKRAOUI

L2TI Laboratory, University Sorbonne Paris Nord, Villetaneuse, France

minh-duc.vu@epita.fr, {zuheng.ming, fangchen.feng, bissmella.bahaduri, anissa.mokraoui}@univ-paris13.fr

Abstract—Object detection in remote sensing imagery plays a vital role in various Earth observation applications. However, unlike object detection in natural scene images, this task is particularly challenging due to the abundance of small, often barely visible objects across diverse terrains. To address these challenges, multimodal learning can be used to integrate features from different data modalities, thereby improving detection accuracy. Nonetheless, the performance of multimodal learning is often constrained by the limited size of labeled datasets. In this paper, we propose to use Masked Image Modeling (MIM) as a pre-training technique, leveraging self-supervised learning on unlabeled data to enhance detection performance. However, conventional MIM such as MAE which uses masked tokens without any contextual information, struggles to capture the fine-grained details due to a lack of interactions with other parts of image. To address this, we propose a new interactive MIM method that can establish interactions between different tokens, which is particularly beneficial for object detection in remote sensing. The extensive ablation studies and evluation demonstrate the effectiveness of our approach.

Index Terms—Self-supervised learning, multimodal learning, object detection, remote sensing image.

Fig. 1. Interactive masked image modeling for self-supervised pre-training. The top is the conventional masked image modeling such as MAE [6]. The bottom is the interactive masked image modeling, in which a cross-attention is introduced to create the interaction between unmasked tokens and masked tokens. The features of unmasked token from encoder (green squares) are merged with the features of masked token from the cross-attention module (orange squares) to reconstruct the masked images.

I. INTRODUCTION

Object detection in Remote Sensing Images (RSI) including aerial images is a critical task enabling the identification and localization of objects within satellite or aerial imagery. It has numerous applications for Earth Observation (EO) such as environmental monitoring, climate change, urban planning, and military surveillance [1]. Object detection in aerial imagery is a complex task, made even more challenging by several critical factors. These include the limited availability of annotated data compared to standard imagery [2], the smaller size of objects in aerial environments [3], and the distinct top-down perspective inherent to aerial observations.

To overcome these challenges, complementary information from alternative modalities, such as infrared (IR) which allows us to see through certain obstructions like smoke or fog, can be leveraged [4]. Both IR and RGB sensors are widely used in aerial vehicles and Earth observation satellites. By combining these spectral channels, multispectral images are generated, providing far richer information than RGB images alone [5]. This enhanced data stream has the potential to significantly improve the precision and accuracy of object detection in aerial imagery. However, detection performance in multimodal scenarios is often limited by the small size of available datasets, as collecting multimodal data is inherently challenging. Additionally, multimodal architectures are gener-ally more complex than single-modal ones and require larger datasets for effective training.

To mitigate this issue, self-supervised learning (SSL) has gained prominence as a powerful deep learning approach, particularly for tasks where labeled data is scarce or expensive to obtain. SSL leverages unlabeled data by creating pretext tasks that enables the model to learn generic and comprehensive features of the images [7]. A model pretrained on a pretext task can be more efficiently adapted to downstream tasks through fine-tuning with limited domain-specific labeled data. As one of the most widely used SSL method, Masked Image Modeling [8] (MIM) allows the model to understand the underlying concepts of an image by predicting the masked portions of an input image (see Figure 1), thereby forcing the model to develop a deeper understanding of the image structure. This approach is particularly effective in scenarios where images contain significant amounts of contextual information, such as aerial imagery or remote sensing [9].

Nevertheless, conventional MIM methods such as MAE [6] reconstruct masked images using the features of both unmasked and masked tokens. The masked tokens are typically set to null, containing no information. Consequently, the interaction between different parts of the image, such as between unmasked and masked tokens, is often interrupted during reconstruction. While this approach can help the model learn the global context, it may not always capture fine-grained details that are crucial for detecting small or densely packed objects in remote sensing images, as the surround-

ing contextual information of an unmasked token cannot be provided by neighboring masked tokens. Due to the lack of interaction between unmasked and masked tokens, standard MIM methods may also struggle with variations in object size and contextual complexity, such as diverse terrains in remote sensing images, as they may not fully capture these variations in their representations. In this work, we propose an interactive MIM, in which a cross-attention module is introduced to the standard MIM to create dependencies between unmasked and masked tokens. The generated features derived from the masked tokens via the cross-attention module are then merged with the features of the unmasked tokens to reconstruct the masked images. This approach encourages the encoder to learn a more holistic understanding of the image content, which benefits downstream object detection tasks in remote sensing images.

To summarize, the main contributions of this work are:

1) We propose using multimodal SSL for object detection in remote sensing images to address data scarcity by leveraging the rich unlabeled data from different sources.

2) To address the limitations of traditional MIM, we propose a new interactive masked image modeling method that better supports downstream object detection tasks in remote sensing images.

3) The extensive experiments including ablation studies and overall evaluations demonstrate the effectiveness of the proposed approach in both single-modality and multimodal settings.

II. RELATED WORKS

Since its introduction, Masked Image Modeling (MIM) has gained significant popularity as a pre-training strategy, leading to extensive research aimed at improving its effectiveness. Several studies have focused on enhancing the masking process. In distillation-based MIM, for example, the authors of [10] introduced a teacher transformer encoder that generates an attention map, guiding the masking process for the student model. Similarly, the authors of [11] used a self-attention mechanism to extract semantic information during training, which informs the masking strategy. A comparable approach is found in [12], specifically for remote sensing images, where the authors developed an object-centric data generator to automatically configure pre-training data based on objects within the imagery. By masking regions centered around objects, this method encourages models to learn richer, object-specific features, outperforming traditional MIM techniques.

In addition to masking strategies, some works focus on the relationship between masked and unmasked tokens. The authors of [13] argue that it is beneficial to constrain predicted representations to the encoded representation space. Meanwhile, the authors of [14] proposed a cross-attention decoder to combine masked and unmasked tokens—this work serves as the inspiration for our paper. Building on this approach, we extend its application to the more complex multimodal setting. One key improvement we introduce, as demonstrated through experiments, is that the query in the cross-attention decoder

can be either masked or unmasked tokens, adding flexibility to the model. Further details are provided in the next section. Other relevant research includes foundation models for remote sensing images, also based on MIM [9].

III. METHODS

A. Overall architecture

Fig. 2. Overview of our framework. Our proposed framework consists of two stages: pre-training (top) on the AVIID or DIOR datasets, and fine-tuning (bottom) on VEDAI. During pre-training, the output features of unmasked tokens from the encoder, merged with the output features of masked tokens from the cross-attention module, are used to reconstruct the multimodal images. After pre-training, the decoder is discarded, and the pre-trained encoder serves as the image encoder for fine-tuning on VEDAI.

Figure 2 shows the overall architecture of our proposed multimodal cross-attention Masked Image Modeling (MIM) for object detection in remote sensing images. Similar to conventional SSL-based modeling [15], our framework consists of two stages: pre-training and downstream task fine-tuning. The pretrained encoder in the first stage is adapted to downstream tasks, such as object detection, through fine-tuning. In the pre-training stage, the images of different modalities are first concatenated and masked. The unmasked tokens are fed to the encoder to obtain their features. Unlike the MAE [6], one of the most widely used MIM methods for pre-training, we introduce a cross-attention module that uses the features of unmasked tokens as anchors to infer meaningful features of masked tokens that contains contextual information. The features of unmasked tokens and those derived from masked tokens are then input to the decoder to reconstruct the masked multimodal images, such as RGB-IR images. The details of the proposed cross-attention MIM will be elaborated in the following section. Here, we use the Swin Transformer [16] as the encoder, and YOLOv5 [17] as the detection neck/head in the downstream object detection task.

B. Cross Attention MIM

As shown in Figure 1, we designed a cross-attention module to generate features from the masked tokens by creating dependencies between the masked and unmasked areas. Unlike conventional MIM using masked tokens (represented by the gray squares in Figure 1), the proposed cross-attention MIM merges the output features of unmasked tokens (the green

squares in Figure 1) from the encoder with the features of masked tokens from the cross-attention module (the orange squares in Figure 1) to reconstruct the image. The features from the masked tokens can capture the surrounding context of the unmasked tokens, which is lost in the original masked tokens set as null. This approach allows the encoder to be trained more effectively, capturing more generic and comprehensive features of the image, which benefits downstream tasks.

The cross-attention mechanism can be formulated as follows:

$$\text{Attention}(Q_m, K_u, V_u) = \text{softmax}\left(\frac{Q_m K_u^T}{\sqrt{d_k}}\right) V_u \quad (1)$$

Where:
- Q_m: Query matrix for masked tokens
- K_u: Key matrix for unmasked tokens
- V_u: Value matrix for unmasked tokens
- d_k: Dimension of the key vectors

The obtained attention scores are used as the features of masked tokens to reconstruct the masked images.

C. Reconstruction loss

We employ the l_2-norm distance between the pixels of the original image and the reconstructed image to as the reconstruction loss, as described by the following equation:

$$L_{W_e, W_d} = \frac{1}{N} \cdot \sum_{j=1}^{N} ||\mathbf{x} - f^c(\mathbf{x})||^2, \quad (2)$$

where W_e and W_d represent the weights of the encoder and decoder respectively. Here, N denotes the total number of samples, $\mathbf{x}$ is the original image, and $f^c(\mathbf{x})$ is the reconstructed image generated by the model $f^c(\cdot)$. When inputting multimodal images such as RGB-IR image pairs, the IR image is concatenated with RGB image in a channel-wise manner to form the new input for the model, which can be given by:

$$\mathbf{x} = \mathbf{x}_{rgb} \oplus \mathbf{x}_{ir}, \quad (3)$$

where $\mathbf{x}_{rgb}$ is RGB image, $\mathbf{x}_{ir}$ is IR image and $\mathbf{x}$ is the multimodal input for the model, $\oplus$ denotes the concatenation of images in a channel-wise manner.

IV. EXPERIMENTS

A. Experiments settings

We apply our proposed self-supervised methods to three remote sensing image datasets for pre-training: VEDAI [18], DIOR [19], and AVIID [20]. VEDAI consists of 1,246 images in both 1024×1024 and 512×512 resolutions. Each image has 4 channels: RGB and IR. DIOR is a large-scale remote sensing dataset containing over 23,000 high-resolution RGB images with approximately 200,000 annotated object instances across 20 different classes. AVIID is composed of 3 parts, but only part 3 consists of aerial images.AVIID-3 contains 1,280 pairs of RGB-IR images at a 512×512 resolution and has been used in this work for pre-training. For the downstream task, we only use the VEDAI dataset for training and validation.The dataset is divided into 10 folds and we use the first folder for the ablation studies, and all 10 folders for the overall evaluation. The mAP (mean Average Precision) at 0.5 IoU, i.e. MAP@.5, is used as accuracy metrics to evaluate the object detection performance. We have used an AdamW optimizer with a base learning rate of 1e-5 and 0.005 weight decay. 8 Nvidia Tesla V100 GPUs were used for the training.

B. Experiment results

1) Ablation study: We verify the effectiveness of our proposed method by designing a series of ablation experiments conducted on the first fold of the validation set of VEDAI.

a) Effectiveness of cross-attention MIM Table II shows the effectiveness of the proposed interactive MIM using cross-attention for pre-training on RGB images. To provide a baseline for comparison, we first trained a model from scratch on VEDAI for object detection without any pre-training, achieving a score of 0.50 for mAP@.5. Next, we applied conventional MIM such as MAE, to pre-train the encoder on VEDAI and fine-tune the pre-trained encoder for the downstream task. This resulted in a subpar score of 0.42, confirming that SSL requires a large dataset to be effective [15]. The score improved to 0.52 when pre-training the encoder on the larger DIOR dataset (20x larger than VEDAI). The performance improves from 0.52 to 0.62, when we use the proposed interactive MIM introducing a cross-attention module into the standard MIM framework. This improvement demonstrates that the interaction between unmasked tokens and masked tokens enables the encoder to learn fine-grained details in the surrounding context that are crucial for detecting small or densely packed objects in remote sensing images. Additionally, we also use unmasked tokens as the query matrix Q to generate its features through the cross-attention module. As expected, the performance was lower than the previous approach since partial information is missing when using the theirs features instead of original unmasked tokens.

b) Effectiveness of multimodal interactive MIM Table III also shows the effectiveness of the proposed interactive MIM when using multimodal images. Compared to the baseline only uses RGB images, the new baseline using multiple modalities significantly improves from 0.50 to 0.63 in mAP@.5, demonstrating the effectiveness of multimodal learning for the object detection in remote sensing. Moreover, the performance further improves to 0.64 when using proposed interactive MIM. However, the AVIID dataset contains only 1246 images, which limits the power of the proposed method. We augmented the data by resizing the images from 480x480 pixels to 512x512 pixels and then splitting them into 4256 images. Then the performance improves to 0.68, demonstrating that our proposed method is also effective on multimodal images.

c) Impact of mask size The varying size of objects, especially small objects, is a major challenge in object detection within remote sensing images. The impact of applying different mask sizes in pre-training, based on our proposed method, on the downstream object detection task is shown in Table

TABLE I

CLASS-WISE OVERALL EVALUATION OF THE PROPOSED METHODS FOR OBJECT DETECTION IN REMOTE SENSING IMAGES.

Model	Modality	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP.5
YOLOv3 [21]	RGB	0.83	0.71	0.69	0.59	0.48	0.67	0.33	0.55	0.61
	RGB + IR	0.84	0.72	0.67	0.62	0.43	0.65	0.37	0.58	0.61
YOLOv4 [22]	RGB	0.84	0.73	0.71	0.59	0.52	0.66	0.34	0.60	0.62
	RGB + IR	0.85	0.73	**0.72**	0.63	0.49	0.69	0.34	0.55	0.62
YOLOv5s [23]	RGB	0.80	0.68	0.66	0.51	0.46	0.64	0.22	0.41	0.55
	RGB + IR	0.81	0.68	0.69	0.55	0.47	0.64	0.24	0.46	0.57
YOLOv5m [23]	RGB	0.81	0.70	0.66	0.54	0.47	0.67	0.36	0.50	0.59
	RGB + IR	0.83	0.72	0.68	0.59	0.46	0.66	0.33	0.57	0.61
YOLOv5l [23]	RGB	0.81	0.72	0.68	0.57	0.46	0.71	0.36	0.55	0.61
	RGB + IR	0.83	0.72	0.70	0.64	0.48	0.63	0.40	0.56	0.62
YOLOv5x [23]	RGB	0.82	0.72	0.68	0.59	0.48	0.66	0.39	0.62	0.62
	RGB + IR	0.84	0.73	0.70	0.61	0.50	0.67	0.39	0.57	0.62
YOLOrs [4]	RGB	**0.85**	0.73	0.70	0.51	0.43	**0.77**	0.19	0.39	0.57
	RGB +IR	0.84	0.78	0.69	0.53	0.47	0.68	0.21	0.58	0.60
Ours	RGB	0.81	0.74	0.64	0.63	**0.53**	0.63	0.53	0.61	0.64
	RGB + IR	0.80	**0.79**	0.70	**0.73**	0.45	0.72	**0.53**	**0.68**	**0.68**

TABLE II

ABLATION STUDY ON THE EFFECTIVENESS OF INTERACTIVE MIM USING CROSS-ATTENTION AS PRE-TRAINING FOR THE OBJECT DETECTION TASK.

Modality	mAP@.5	pre-training datasets		Cross Attention	
		VEDAI	DIOR	Q masked	Q unmasked
RGB	0.50	(Training from scratch, w/o pre-training)			
	0.42	✓			
	0.52		✓		
	0.62		✓	✓	
	0.60		✓		✓

TABLE III

ABLATION STUDY OF USING INTERACTIVE MIM AS PRE-TRAINING FOR THE MULTIMODAL OBJECT DETECTION TASK.

Modality	mAP@.5	AVIID		Cross Attention
		original	data expansion	Q masked
RGB + IR	0.63	(Training from scratch, w/o pre-training)		✓
	0.64	✓		✓
	0.68		✓	✓

(a) Ground truth (b) RGB w/o MIM (c) RGB with MIM (d) RGB+IR with MIM (e) RGB+IR with MIM+CA

Fig. 3. Visual illustration of the effectiveness of the proposed interactive MIM for multimodal object detection in remote sensing images.

3) Visualization: Figure 3 visually demonstrates the effectiveness of the multimodal cross-attention MIM for the object detection task. In (b), compared to the ground truth shown in (a), several objects, such as the pickup and cars, are missing when using only RGB images without pre-training. With pre-training based on MIM, the pickup and more cars are detected, as shown in (c). When using multimodal images (RGB+IR) with conventional MIM, all the cars are detected, but the car and pickup are still confused. Finally, when using multimodal RGB+IR images with cross-attention MIM, all objects are successfully detected with high accuracy.

V. CONCLUSION

The results demonstrate that integrating self-supervised learning via MIM and multimodal data fusion significantly improves the performance of object detection in remote sensing images. Our experiments show that incorporating IR data alongside RGB enhances the model's capability, especially for small and occluded objects. The proposed interactive MIM, which introduces a cross-attention module to establishe the interaction between unmasked and masked tokens, overcomes the shortcomings of conventional MIM. This approach encourages the encoder to learn a more holistic understanding of the image content, benefiting downstream object detection tasks in remote sensing images. Although our work is initially proposed for remote sensing, it can also be extended to other domains, applying to more scenarios with limited resources.

IV. The results indicate that the optimal performance was achieved with medium-sized masks of 32x32 pixels, offering a balance between capturing global context with large masks and localized fine-grained feature learning with small masks. The 32x32 pixel is the default mask size used in our experiments.

TABLE IV

IMPACT OF APPLYING MASKS WITH DIFFERENT SIZES IN PRE-TRAINING ON DOWNSTREAM OBJECT DETECTION TASK.

Mask size	16	32	64
mAP@.5	0.63	**0.68**	0.61

2) Overall evaluation: Table I compares our proposed method with other methods on the object detection task on the benchmark VEDAI dataset. Our model demonstrates superiority in both single modality using RGB images and multimodality using RGB+IR images. Notably, our model also proves effective in detecting pickups both in single moality, which are easily confused with trucks, and in detecting relatively small objects, such as boats.

REFERENCES

[1] S. Gui, S. Song, R. Qin, and Y. Tang, "Remote sensing object detection in the deep learning era—a review," *Remote Sensing*, vol. 16, no. 2, p. 327, 2024.

[2] P. Le Jeune and A. Mokraoui, "Improving few-shot object detection through a performance analysis on aerial and natural images," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 513–517, IEEE, 2022.

[3] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 79–93, 2022.

[4] M. Sharma *et al.*, "Yolors: Object detection in multimodal remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1497–1508, 2020.

[5] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[7] J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum, "A cookbook of self-supervised learning," *arXiv preprint arXiv:2304.12210*, 2023.

[8] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

[9] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2022.

[10] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, "What to hide from your students: Attention-guided masked image modeling," in *European Conference on Computer Vision*, pp. 300–318, Springer, 2022.

[11] Z. Liu, J. Gui, and H. Luo, "Good helper is around you: Attention-driven masked image modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 1799–1807, 2023.

[12] T. Zhang, Y. Zhuang, H. Chen, L. Chen, G. Wang, P. Gao, and H. Dong, "Object-centric masked image modeling based self-supervised pretraining for remote sensing object detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, pp. 1–13, 01 2023.

[13] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 208–223, 2024.

[14] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, *et al.*, "Efficientsam: Leveraged masked image pretraining for efficient segment anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16111–16121, 2024.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[16] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[17] X. Cao, Y. Zhang, S. Lang, and Y. Gong, "Swin-transformer-based yolov5 for small-object detection in remote sensing images," *Sensors*, vol. 23, no. 7, p. 3634, 2023.

[18] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[19] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[20] Z. Han, Z. Zhang, S. Zhang, G. Zhang, and S. Mei, "Aerial visible-to-infrared image translation: dataset, evaluation, and baseline," *Journal of Remote Sensing*, vol. 3, p. 0096, 2023.

[21] J. Redmon *et al.*, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[22] A. Bochkovskiy, C.-Y. Wang, *et al.*, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[23] G. J. et al, "ultralytics/yolov5: v5.0," *online*, 2021.