SAR OBJECT DETECTION WITH SELF-SUPERVISED PRETRAINING AND CURRICULUM-AWARE SAMPLING

Yasin Almalioglu, Andrzej Kucik, Geoffrey French, Dafni Antotsiou,
Alexander Adam, and Cedric Archambeau
Helsing, 27 Mortimer St, London, W1T 3BL, UK
{first.last}@helsing.ai

ABSTRACT

Object detection in satellite-borne Synthetic Aperture Radar (SAR) imagery holds immense potential in tasks such as urban monitoring and disaster response. However, the inherent complexities of SAR data and the scarcity of annotations present significant challenges in the advancement of object detection in this domain. Notably, the detection of small objects in satellite-borne SAR images poses a particularly intricate problem, because of the technology's relatively low spatial resolution and inherent noise. Furthermore, the lack of large labelled SAR datasets hinders the development of supervised deep learning-based object detection models. In this paper, we introduce TRANSAR, a novel self-supervised end-to-end vision transformer-based SAR object detection model that incorporates masked image pre-training on an unlabeled SAR image dataset that spans more than $25,700$ km^2 ground area. Unlike traditional object detection formulation, our approach capitalises on auxiliary binary semantic segmentation, designed to segregate objects of interest during the post-tuning, especially the smaller ones, from the background. In addition, to address the innate class imbalance due to the disproportion of the object to the image size, we introduce an adaptive sampling scheduler that dynamically adjusts the target class distribution during training based on curriculum learning and model feedback. This approach allows us to outperform conventional supervised architecture such as DeepLabv3 or UNet, and state-of-the-art self-supervised learning-based arhitectures such as DPT, SegFormer or UperNet, as shown by extensive evaluations on benchmark SAR datasets.

1 INTRODUCTION

Recent advances in self-supervised learning (SSL) have significantly improved computer vision (He et al., 2022; Bao et al., 2022) and remote sensing (Toker et al., 2022; Shermeyer et al., 2020; Tuia et al., 2024), enabling the extraction of high-level representations from unlabelled data. SSL frameworks typically involve pretraining using contrastive learning (Chen et al., 2020) or masked image modelling (MIM) (He et al., 2022), followed by fine-tuning on tasks such as object detection and segmentation (Scheibenreif et al., 2022; Zorzi et al., 2022).

Synthetic Aperture Radar (SAR) is widely used in environmental monitoring (Bountos et al., 2022), disaster management (Shermeyer et al., 2020), and military surveillance (Moreira et al., 2013). Its unique ability to operate in all weather conditions and independent of external illumination makes it invaluable for remote sensing. However, SAR imagery presents challenges such as speckle noise, geometric distortions, and low resolution, particularly for small objects like vehicles, which are often represented by only a few pixels (Zhu et al., 2021). These challenges are compounded by the severe class imbalance in SAR imagery and the scarcity of annotated datasets due to the high cost and expertise required for labeling.

While SSL techniques have been adapted for remote sensing tasks like land cover segmentation (Scheibenreif et al., 2022) and change detection (Mall et al., 2023a), SAR object detection remains underexplored. Vision transformers (ViTs) (Dosovitskiy et al., 2021) have proven effective for SSL with segmentation (Jain et al., 2023) and MIM (Bao et al., 2022), but SAR-specific challenges like class imbalance are not fully addressed. Traditional sampling approaches in SAR object detection, such as offline hard-negative mining (Hughes et al., 2018) and oversampling, often lead to overfitting or fail to generalize well due to the severe class imbalance. While adaptive learning strategies like curriculum learning (Huang et al., 2020) and the small-loss criterion (Mit, 2024; Jiang

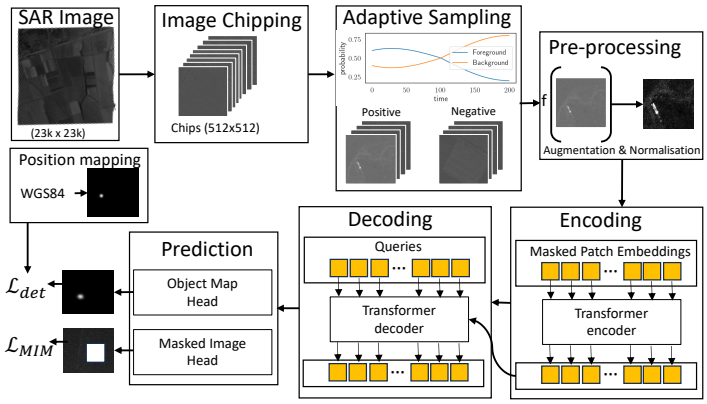

Figure 1: The proposed SSL SAR object detection pipeline with adaptive sampling. Adaptive sampling balances foreground and background in each batch, guided by prediction performance. The vision transformer processes image patches, embedding them with positional encoding. Lightweight prediction heads handle object map predictions during pretraining and reconstruction during fine-tuning.

et al., 2018) have demonstrated effectiveness in handling imbalanced datasets, their application in SAR remains limited. Recent work in remote sensing foundation and vision-langugage models, such as GeoPixel (Shabbir et al., 2025), GRAFT (Mall et al., 2023b) and GeoChat (Kuckreja et al., 2024), have highlighted the potential of large-scale pretraining for geospatial tasks (Danish et al., 2024). However, these models primarily focus on optical imagery, their expansion to SAR remains an open challenge, reinforcing the need for SAR-specific self-supervised approaches.

This paper introduces TRANSAR, a vision transformer model for SAR object detection based on SSL. TRANSAR addresses class imbalance with a novel adaptive sampling scheduler and incorporates an SSL-MIM phase tailored for SAR representation learning. Additionally, we introduce an auxiliary semantic segmentation-based method process to enhance small-object detection during post-training. Our experiments demonstrate that TRANSAR outperforms architectures like DeepLabv3 (Chen et al., 2017), DPT (Ranftl et al., 2021), and SegFormer (Xie et al., 2021), establishing its effectiveness for SAR object detection.

2 METHOD

TRANSAR is a transformer-based SAR object detection model that combines SSL-MIM pretraining; supervised auxiliary binary semantic segmentation to segregate small objects; and an adaptive sampling scheduler to address class imbalance.

2.1 ARCHITECTURE AND TRAINING

TRANSAR architecture builds on ViTs with Swin transformers' shifted window approach(Liu et al., 2021) for scale invariance and efficiency. The encoder backbone includes four blocks, each combining a patch merging layer and Swin transformer blocks with multi-head self-attention and residual post-normalization (Liu et al., 2022a). The architecture scales by adjusting dimensionality, pooling layers, and patch embeddings. For SSL-MIM pretraining, we use a CNN-based pixel-shuffling lightweight reconstruction head (Liu et al., 2021) with block-wise masking (Bao et al., 2022), enabling the model to learn SAR intensity patterns without relying on pixel interpolation.

During fine-tuning, the MIM head is replaced with a detection head comprising convolution and pixel-shuffling layers (Liu et al., 2022a; Shi et al., 2016). We represent annotations as Gaussian blobs centered on object coordinates, designed to segregate small, point-like objects, as shown in Fig. 1. The target tensor is convolved with a 2D Gaussian kernel to encode object positions. The loss function combines binary cross-entropy (BCE) and Dice loss (Milletari et al., 2016; Sudre et al., 2017) to handle class imbalance effectively:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \alpha \text{BCE}(\mathbf{y}, \hat{\mathbf{y}}) + \beta \text{Dice}(\mathbf{y}, \hat{\mathbf{y}}), \tag{1}$$

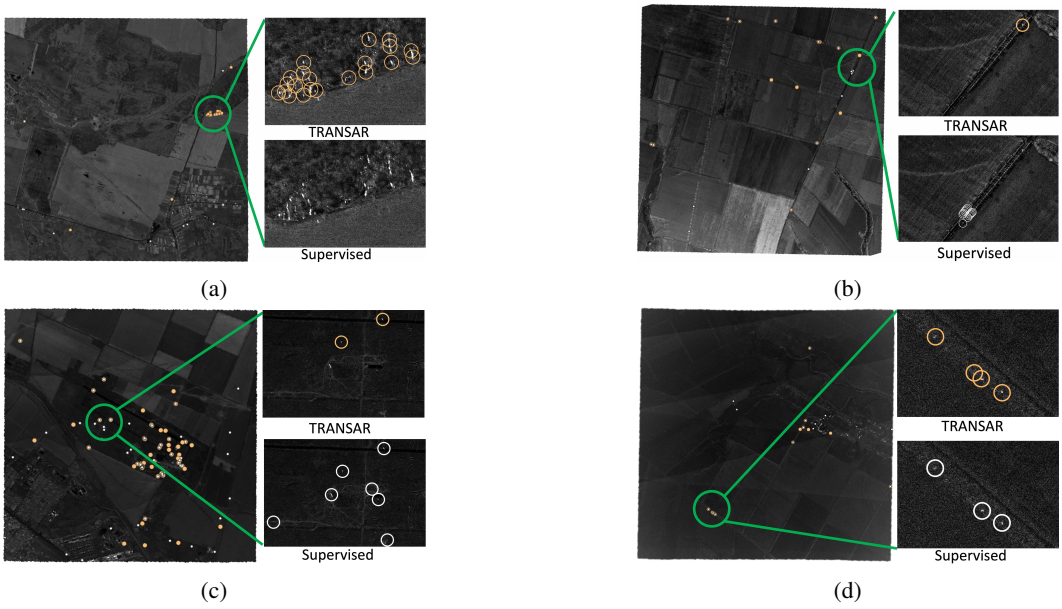

Figure 2: Example qualitative SAR object detection results. **a.** Fine-grained detection of small objects. The supervised model fails at distinguishing the concentrated target objects **b.** Robust to false reflective objects. The supervised model generates false positives. **c.** Precise detection. The supervised model has mixed false and true predictions. **d.** Similar performance in the rural areas where reflective objects are distinct from the background.

where α and β weight the loss components. To address imbalance further, the loss is adaptively weighted, and unpopulated pixels caused by imperfect projection of the georeferenced data onto the rigid tensor grid are excluded from calculations.

2.2 Adaptive Sampling Scheduler

We describe the adaptive sampling scheduler (AS) to handle class imbalance. The AS is based on curriculum learning (Bengio et al., 2009) and hard negative sampling (Shrivastava et al., 2016) that demonstrates a strategy of gradual learning from easy to hard samples significantly improves the generalisation and imbalanced learning (Guo et al., 2018). The scheduler dynamically tailors the target distribution during the training process, transitioning it from an imbalanced to a balanced state within a batch.

Let C_i be the cardinality of the set of samples belonging to class i. We introduce the vector $\mathbf{d}^{\text{train}}$ to represent the training class distribution. For each attribute, the i^{th} element of the class distribution $\mathbf{d}^{\text{train}}$ is defined as the ratio of $C_{max} - C_i$ to the size of majority class, C_{max} to over-sample the minority classes in the early stages. Explicitly, we define $\mathbf{d}^{\text{train}} = (d_i^{\text{train}})_{i=1}^{K}$ by setting $d_i = 1 - C_i/C_{max}$, where K represents the number of classes. The sampling scheduler dictates the target data distribution $\mathbf{d}^{\text{target}} = (d_i^{\text{target}})_{i=1}^{K}$ of the attributes in each batch as a function of epoch t. Initially, the target distribution of one attribute in a batch is set to imbalanced $\mathbf{d}^{\text{train}}$. During the training process, it gradually transfers to a balanced distribution with the subsequent function at epoch t:

$$d_i^{\text{target}}(t) = \begin{cases} d_i^{\text{train}} & \text{if } t = 0 \\ \left(d_i^{\text{train}}\right)^{\alpha g(t) + (1-\alpha) h(t)} & t > 0 \end{cases}, \tag{2}$$

where $g(t) \in [0, 1]$ is a sampling scheduler function such as linear, cosine and exponential following the prior work of Wang et al. (2019), $h(t) \in [0, 1]$ is a sampling regulariser function based on model performance such as the F1 score, α is the sampler weight. Function $g(t)$ is akin the class-balancing techniques recommended for handling long-tailed class distributions (Wei et al., 2021; Hoyer et al., 2022). We introduce the function $h(t)$ to assist the scheduler in monitoring model performance and dynamically prioritizing classes that may lead to false detections. It is worth noting that $h(t)$ can exhibit fluctuations across epochs without a regulariser, $g(t)$. Thus, we put a larger weight

Method	Model	Pretraining	AP25	AP50	AP75	mAP	Precision	Recall	F1
SSL	ViT-Uper(Liu et al., 2022b; Jain et al., 2023)	RGB	36.77	40.09	41.26	38.87	40.14	42.26	41.17
		SAR-MIM	44.86	51.97	54.26	47.73	52.06	48.27	50.09
	ViT-MAE(He et al., 2022)	RGB	34.78	38.29	39.81	36.27	38.75	41.16	39.92
		SAR-MIM	46.12	53.79	58.44	50.14	55.51	47.94	51.45
	DPT(Ranftl et al., 2021)	RGB	36.43	38.37	40.29	37.61	39.25	42.18	40.66
		SAR-MIM	51.10	55.03	58.84	52.89	55.29	62.15	58.52
	SegFormer(Xie et al., 2021)	RGB	34.81	36.39	40.87	35.79	38.59	42.76	40.57
		SAR-MIM	45.28	49.74	51.23	47.10	50.6	60.38	55.06
	TRANSAR-medium	SAR-MIM	50.56	53.41	60.70	51.17	56.72	82.06	67.08
	TRANSAR-large	RGB	54.57	62.11	64.81	59.84	62.86	63.57	63.21
		SAR-MIM	60.10	68.90	85.45	**66.77**	77.86	80.53	**79.17**
Supervised	DeepLabv3(Chen et al., 2017)	RGB	25.7	28.25	28.89	27.80	28.90	59.20	38.84
		No	23.59	26.38	26.69	25.71	26.68	49.84	34.76
	UNet-SENet(Shermeyer et al., 2020)	No	29.37	34.09	37.55	32.16	35.18	38.49	36.76
		SAR	43.63	45.19	47.41	**44.52**	46.10	56.10	**50.61**

Table 1: Comparative object detection performance of TRANSAR with the state-of-the-art supervised and self-supervised architectures.

on $g(t)$ to ensure a gradual change in the distribution (Appendix Fig. 3 illustrates the evolution of the foreground and background sample ratio for $K = 2$ and $\alpha = 0.8$). According to target distribution $\mathbf{d}^{\text{target}}(t)$, the majority class samples are dynamically selected and the minority class samples are re-weighted in different epochs to update the current class distribution. Due to the dynamic class weights calculated by the scheduler, we adapt the class weights in the loss function described in Sec. 2.1 proportionate to $\mathbf{d}^{\text{train}}$. For each epoch t, the per pixel weighted loss is defined as $\mathcal{L}_t^{\text{AS}}(\mathbf{y}, \hat{\mathbf{y}}) = \mathbf{w}(t) \cdot L(\mathbf{y}, \hat{\mathbf{y}})$, where L is the per pixel, unweighted objective function given by Eq. (1). The loss weight, $\mathbf{w}(t)$, at epoch t is given by $\mathbf{w}(t) = \max\left\{\frac{\mathbf{d}^{\text{train}}}{\mathbf{d}^{\text{target}}(t)}, 1\right\}$, where $\max\{\cdot\}$ is the element-wise maximum.

3 Results and Discussion

We demonstrate that the proposed approach outperforms state-of-the-art supervised and self-supervised methods on benchmark SAR object detection datasets. Additionally, we present a detailed ablation study on the SSL architecture, adaptive sampling, and SAR preprocessing, along with qualitative insights. The evaluation is conducted on the X-band, single-polarized (HH) spotlight SAR imagery datasets with approximately 0.3m resolution. The first is a proprietary vehicle dataset comprising 134 images annotated by SAR analysts. The second dataset includes 1028 unlabelled, geo-coded, and terrain-corrected satellite images from Capella Space (Cap), covering about $25700\ \text{km}^2$. These unlabelled images are used exclusively for SSL-MIM pretraining.

We compare TRANSAR with state-of-the-art SSL architectures, such as UperNet (Liu et al., 2022b; Jain et al., 2023), MAE (He et al., 2022), DPT (Ranftl et al., 2021), and SegFormer (Xie et al., 2021), as well as traditional supervised methods like DeepLabv3 (Chen et al., 2017) and UNet-SENet (Shermeyer et al., 2020). Tab. 1 presents detailed comparative detection scores, including average precision (AP) calculated at 20 intervals in $\{0.05, 0.10, ..., 0.95\}$, as suggested in (Lin et al., 2014). We report precision, recall, and F1 scores at the threshold yielding the maximum F1 score. For fair comparison, large-size models of competing SSL architectures are used, with details in the supplementary material.

The results show that TRANSAR-large outperforms other approaches with consistent mAP scores across varying precision thresholds and achieves a well-balanced F1 score between precision and recall. TRANSAR-medium, despite its smaller size, achieves comparable performance to larger SSL models, while TRANSAR-tiny underperforms due to its inability to distinguish foreground objects from the background.

Fig. 2 provides qualitative comparisons between the best-performing supervised UNet-SENet model and TRANSAR. TRANSAR excels in densely distributed objects (Fig. 2a), addressing non-uniform object distribution challenges. It is also more robust to false positives caused by radar-reflective objects like pylons (Fig. 2b) and achieves higher precision in detecting true positives while reducing false positives (Fig. 2c). In simpler scenarios (Fig. 2d), both models perform similarly, especially in rural areas with distinct SAR intensity differences from the background. However, urban areas with dense reflections remain challenging for all SAR object detection models, as shown in supplementary figures. We provide a further discussion of TRANSAR in Appendix G.

REFERENCES

Capella Space Synthetic Aperture Radar (SAR) Open Dataset - Registry of Open Data on AWS. https://registry.opendata.aws/capella_opendata/.

Mitigating data imbalance and noise: A divergence-based approach with enhanced sample selection. *Neurocomputing*, 605:128269, November 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.128269.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers, September 2022.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, June 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380.

Nikolaos Ioannis Bountos, Ioannis Papoutsis, Dimitrios Michail, Andreas Karavias, Panagiotis Elias, and Isaak Parcharidis. Hephaestus: A Large Scale Multitask Dataset Towards InSAR Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1453–1462, 2022.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation, December 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks, November 2024.

Armin Doerry. SAR Image Scaling Dynamic Range Radiometric Calibration and Display. Technical Report SAND–2019-2371, 1761879, 673186, March 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9924–9935, 2022.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2020.

Lloyd Haydn Hughes, Michael Schmitt, and Xiao Xiang Zhu. Mining Hard Negative Samples for SAR-Optical Image Matching Using Generative Adversarial Networks. *Remote Sensing*, 10(10): 1552, October 2018. ISSN 2072-4292. doi: 10.3390/rs10101552.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer To Rule Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2023.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2304–2313. PMLR, July 2018.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, New Orleans, LA, USA, June 2022b. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01167.

Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-Aware Sampling and Contrastive Learning for Satellite Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5261–5270, 2023a.

Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment, December 2023b.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, October 2016. doi: 10.1109/3DV.2016.79.

Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-Supervised Vision Transformers for Land-Cover Segmentation and Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1422–1431, 2022.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S. Khan, and Salman Khan. GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing, January 2025.

Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan Lewis. SpaceNet 6: Multi-Sensor All Weather Mapping Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 196–197, 2020.

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training Region-Based Object Detectors With Online Hard Example Mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, 2016.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu (eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pp. 240–248, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67558-9. doi: 10.1007/978-3-319-67558-9_28.

Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixé. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21158–21167, 2022.

Devis Tuia, Konrad Schindler, Begüm Demir, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N. van Rijn, Holger H. Hoos, Fabio Del Frate, Mihai Datcu, Volker Markl, Bertrand Le Saux, Rochelle Schneider, and Gustau Camps-Valls. Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–25, 2024. ISSN 2168-6831. doi: 10.1109/MGRS.2024.3425961.

Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic Curriculum Learning for Imbalanced Data Classification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5016–5025, October 2019. doi: 10.1109/ICCV.2019.00512.

Sandhi Wangiyana, Piotr Samczyński, and Artur Gromek. Data Augmentation for Building Footprint Segmentation in SAR Images: An Empirical Study. *Remote Sensing*, 14(9):2012, April 2022. ISSN 2072-4292. doi: 10.3390/rs14092012.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10857–10866, 2021.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021.

Xiao Xiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler. Deep learning meets SAR: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):143–172, 2021.

Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. PolyWorld: Polygonal Building Extraction With Graph Neural Networks in Satellite Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1848–1857, 2022.

SUPPLEMENTARY MATERIAL

In the following pages, we provide supplementary information related to the methodology, experimental setup and results.

A TRANSAR AT DIFFERENT SCALES

We describe the architecture details of TRANSAR in Sec. 2. We provide additional details of model parameters at different scales in Tab. 2. The provided configurations pertain to the TRANSAR backbone models, a deep learning architecture used for SAR object detection tasks. These settings define critical parameters for the model's architecture and training process. Notably, the "hidden size" determines the dimensionality of the encoder layers and the pooler layer, while "image size" specifies the resolution of each image. The "patch size" and "window size" parameters determine the size of patches and windows, respectively. "num channels" signifies the number of input channels, and "embed dim" defines the dimensionality of patch embedding. The "depths" list characterises the depth of each layer in the Transformer encoder, and "num heads" represents the number of attention heads in each encoder layer. "qkv bias" indicates whether biases are added to queries, keys, and values. These configurations collectively shape the behaviour and capabilities of the TRANSAR models for a given task.

Configuration	**Tiny**	**Medium**	**Large**
#params (m)	27	86	195
image size	512	512	512
hidden size	768	1024	1536
patch size	4	4	4
window size	7	7	12
embed dim	96	128	192
depths	[2, 2, 6, 2]	[2, 2, 18, 2]	[2, 2, 18, 2]
num heads	[3, 6, 12, 24]	[4, 8, 16, 32]	[6, 12, 24, 48]
qkv bias	true	true	true

Table 2: Details of the TRANSAR models at tiny, medium and large scales.

A.1 ABLATION STUDIES

We run ablation studies to better understand the role of TRANSAR components. Appendix A.1 summarises the results on different adaptive sampling schedulers, normalisation functions, and various pretraining mask sizes. For the experiments in each ablation category, we keep the best setting in the other categories. Although the adaptive sampling schedulers achieve overall best-performing results compared to the baseline SSL and supervised methods, the cosine scheduler achieves the best F1 scores with a clear margin. When we disable the scheduler, we observe a significant drop in the performance, highlighting the major contribution of the adaptive sampling. With respect to the normalisation, we observe less variation in the performance of various normalisation functions compared to the schedulers. The logarithmic normalisation achieves the best F1 score among the linear and arctan functions, which is aligned with the existing SAR radiometry visualisations Cap. Finally, we investigate the effect of the block-masking size used during the pretraining MIM stage. We observed that the larger mask values ignore the background intensity variations and generate poor reconstructed regions. On the other hand, when we use smaller mask sizes such as 4, the model creates blurry reconstructions. Thus, the model pretrained with the mask size of 8 gives the best detection scores in the fine-tuning stage.

A.2 PRETRAINING.

We analyse the model performance on different domains with pretraining on RGB and SAR images to prove the effectiveness of the proposed approach regardless of the pretraining domain in Tab. 1.

Ablation	Settings	Precision	Recall	F1
AS Scheduler	None	56.08	68.29	61.59
	linear	68.87	70.6	69.72
	exponential	70.26	72.51	71.37
	cosine	77.86	80.53	**79.17**
Normalisation	linear	73.56	73.12	73.34
	arctan	75.91	74.28	75.09
	log	77.86	80.53	**79.17**
Mask size	32	69.98	71.13	70.55
	16	74.59	76.18	75.38
	8	77.86	80.53	**79.17**

Table 3: Ablation study on TRANSAR design choices. Each ablation category uses the best setting from the other categories.

The MIM pretraining improves the overall performance of the SSL models, showing the effective use of the unlabelled data via the block-masking strategy. Specifically, the TRANSAR-large model pretrained with SAR MIM achieves the best performance in terms of both mAP and F1 scores, outperforming the competing approaches with significant margins. Supervised UNet-SENet pretrained on SAR outperforms the SSL models with a clear margin in terms of both mAP and F1 scores. However, it performs poorly when compared in the same SAR pretraining setting, indicating the importance of effective SAR domain-related pretraining.

A.3 Inference.

Our inference pipeline differs from the conventional bounding box paradigm in favour of a segmentation-based approach to object detection. The model outputs probability heatmaps, $\hat{\mathbf{y}}$, which represent the likelihood of object presence at each pixel location. To distil this dense probabilistic information into object locations, we apply thresholding at a confidence level $c = 0.5$, thereby retaining only the most salient features of the heatmap. Subsequently, we employ a peak detection algorithm to decode these thresholded heatmaps into Cartesian coordinates, identifying local maxima that serve as putative object centres.

In lieu of the Intersection over Union (IoU) metric used for for Non-Maximum Suppression (NMS), our framework introduces a distance-based criterion. We compute the pairwise distances between the detected peaks, and if any two peaks are within a predefined non-maximum suppression distance d_{NMS}, we suppress the peak with the inferior confidence score. This approach ensures that each object is represented by a single, distinct peak. The accuracy of our object localization is quantified using a hit distance metric, d_{hit}, which offers a more direct measure of spatial accuracy than IoU. A prediction is deemed a true positive if it lies within the d_{hit}-distance of a ground truth target.

B Implementation Details.

We use the same number of epochs, data processing, dataset splits, hyperparameter set, early stopping criteria and backbone size for a fair comparison of the models. We follow the existing work for the pretraining Capella dataset and the annotated dataset, we use 80% of the SAR images for the training, 10% for the validation, and 10% for the test. We optimise the pretraining and fine-tuning stages with the Adam optimiser and a warm-up cosine learning rate scheduler. We train the models for 100 epochs for pretraining and 25 epochs for fine-tuning with 4000 iterations per epoch; and batch size 16 for pretraining and 8 for fine-tuning on two NVIDIA v100 GPUs, where each iteration samples chips from the next image in the dataset. The supervised experiments follow the same settings as the fine-tuning experiments. We use the widely used object detection metrics such as F1, average precision (AP) and mean AP (mAP) at different thresholds Lin et al. (2014). We set $s_{norm} = 16$ for the SAR image normalisation constant, and the Gaussian kernel $\sigma = 10$, and loss weights $\alpha = 0.05$, $\beta = 1$ for all the experiments.

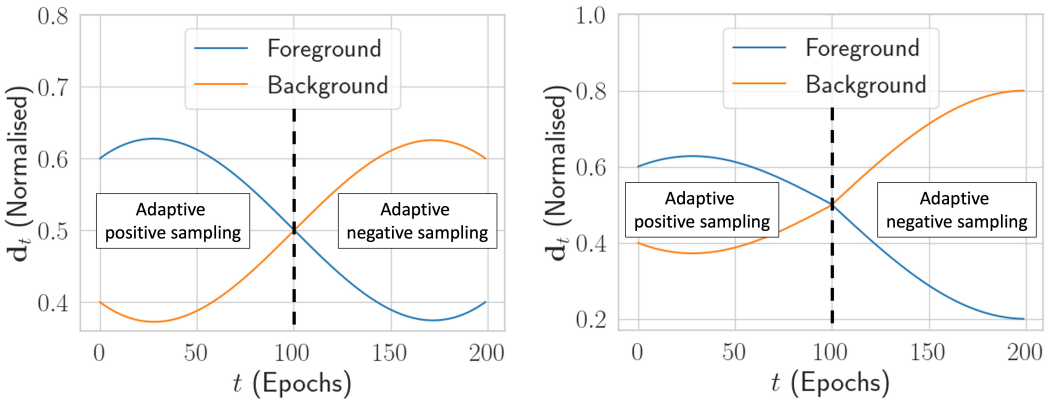

Figure 3: Example normalised sampling distribution of foreground (positive) and background (negative) samples in different precision performances. The sampler frequently draws foreground samples in the early epochs and switches to background samples to improve the precision. **a.** Constantly improving precision. **b.** Precision stalls after epoch 100 and the distribution shifts towards heavy negative sampling.

C Detailed parameters of the self-supervised baselines

We use the model parameters suggested by the authors for both the supervised and the self-supervised models. However, DPT Ranftl et al. (2021) and SegFormer Xie et al. (2021) models offer various model scales with slightly different architectures. ViT-Uper (Liu et al., 2022b) implementation used in the experiments is based on the implementation by Jain et al. (2023). We provide the detailed architectural configurations of the baseline methods used in this study for a clearer comparison.

DPT Ranftl et al. (2021)

- **#params (m):** 305, Number of trainable parameters of the model.
- **hidden size:** 1024, Dimensionality of the encoder layers and the pooler layer.
- **num hidden layers:** 24, Number of hidden layers in the Transformer encoder.
- **num attention heads:** 16, Number of attention heads for each attention layer in the Transformer encoder.
- **intermediate size:** 4096, Dimensionality of the "intermediate" (i.e., feed-forward) layer in the Transformer encoder.
- **patch size:** 16, The size (resolution) of each patch.
- **neck hidden sizes:** [256, 512, 1024, 1024], The hidden sizes for projecting the feature maps of the backbone.

SegFormer Xie et al. (2021) We use the largest model provided by the authors.

- **#params (m):** 82, Number of trainable parameters of the model.
- **num encoder blocks:** 4, The number of encoder blocks (i.e., stages in the Mix Transformer encoder).
- **depths:** [3, 6, 40, 3], The number of layers in each encoder block.
- **sr ratios:** [8, 4, 2, 1], Sequence reduction ratios in each encoder block.
- **hidden sizes:** [64, 128, 320, 512], Dimension of each of the encoder blocks.
- **patch sizes:** [7, 3, 3, 3], Patch size before each encoder block.
- **num attention heads:** [1, 2, 5, 8], Number of attention heads for each attention layer in each block of the Transformer encoder.

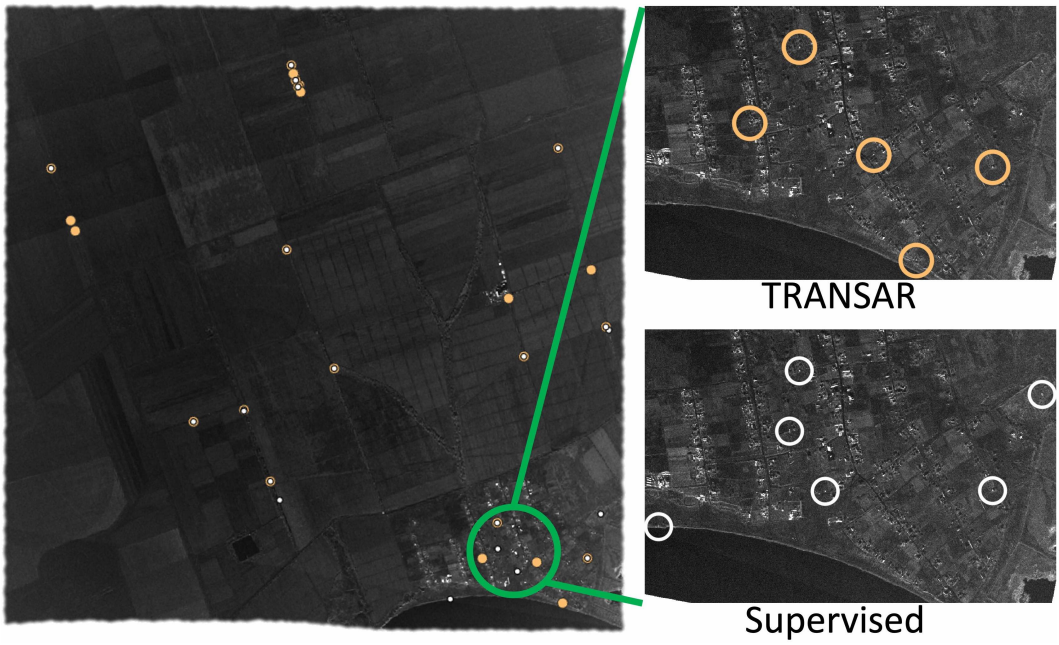

Figure 4: Example detection results in urban areas. Highly reflective objects pose a significant challenge for the models. Both TRANSAR and supervised approaches generate false predictions as shown in the chips.

D Dataset Details

We use 1028 unlabelled X-band spotlight SAR images at the spatial resolution of 0.35 m for both range and azimuth, provided by Capella Cap. Each image covers a nominal scene size of 5×5 km. The sensor load utilises nine looks to capture the scene, providing an azimuth resolution of 0.5 m. The ground range resolution varies between $0.4 - 0.7$ m, while the pixel spacing is at 0.35 m. The look angle range for this product spans from 25° to 50°. Our 134 X-band SAR images have the same specifications as the unlabelled data. The annotations are in WGS84 point coordinates.

D.1 Data Processing

Training samples are composed of batches of $L_W \times L_W$ chips randomly cropped out of georeferenced and terrain-corrected images. We apply geometric (random flips, affine transformations) and radiometric (brightness, contrast, gamma) augmentation. Although some of these augmentations result in physically implausible SAR images Wangiyana et al. (2022), we found that they still improve the overall performance, aligned with the observations in earlier work Shermeyer et al. (2020). SAR data are usually discretised into 16-bit images, and the pixel values approximately follow the Rayleigh distribution. Inspired by the normalisation techniques used in the geographic information systems to make SAR images visually comprehensible Doerry (2019), we propose a logarithmic normalisation as part of the TRANSAR transformation pipeline, i.e.,

$$\hat{\mathbf{x}} = \log_2(\mathbf{x})/s_{norm}, \tag{3}$$

where s_{norm} is a normalisation scale constant. The effectiveness of the normalisation pipeline is shown in Appendix A.1. Finally, the normalised input data is normalised as follows:

$$\hat{\mathbf{x}}_{norm} = (\hat{\mathbf{x}} - \mu_c)/\sigma_g, \tag{4}$$

where the mean, μ_c, is calculated per chip and the standard deviation, σ_g, is the global standard deviation of the training dataset.

During the evaluation, the chips are sampled from the SAR images as a regular, overlapping grid, and only the centre cropped $L_H^E \times L_W^E$ output pixels are used for calculating the evaluation metrics. This is to provide sufficient context and avoid lower-confidence predictions close to the boundary of the receptive field.

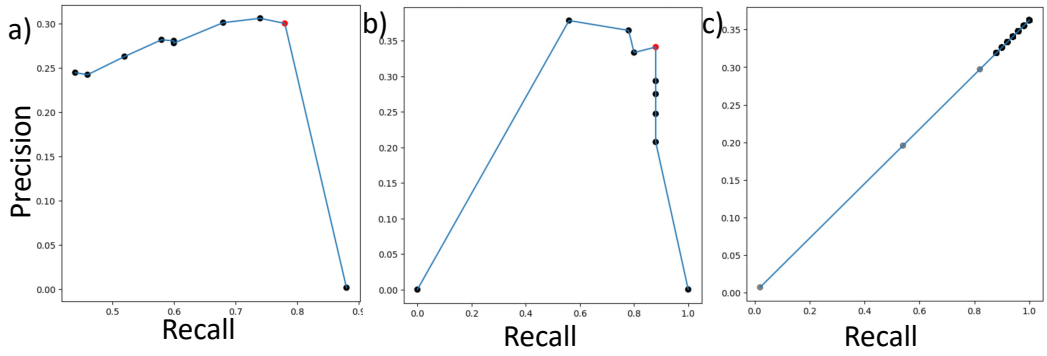

Figure 5: Sensitivity analysis on auxiliary segmentation task in terms of precision recall curves. a) NMS distance. b) Confidence threshold. c) Hit distance.

E Challenges in Urban Environments

In urban areas, object detection faces unique challenges, particularly when dealing with highly reflective objects. These objects, characterised by their reflective surfaces and diverse shapes, often pose a significant challenge for detection models. In our experiments, as illustrated in Fig. 4 in the supplementary material, we observe instances where both TRANSAR and traditional supervised approaches generate false predictions. These false predictions are particularly evident in the visual chips (all the detections in the chips are false-positives), highlighting the complexity of accurately identifying and delineating highly reflective objects within urban environments. This challenge underscores the importance of robust and adaptive detection techniques to effectively address the intricacies of object detection in urban settings.

F Sensitivity Analysis

We present sensitivity analysis for the auxiliary segmentation task on precision-recall curves applied to the labelled data, as illustrated in Fig. 5 for the baseline DeepLabv3 model. Our objective was to analyse the precision and recall of this fixed model on a predefined test set, while varying three critical hyperparameters:

1. **NMS Distance**: We investigated the model's performance over a range of NMS distances from 0 to 500, with 10 data points. Our default value for this parameter is 23.
2. **Confidence Threshold**: We assessed the model's behavior by varying the confidence threshold from 0 to 1 in increments of 0.1. The default value for this threshold is set at 0.5.
3. **Hit Distance**: To explore the impact of hit distance on the model's output, we examined a range from 1 to 10,000, with 2,000 data points. The default hit distance value is 45.

Adjusting the first two parameters necessitated re-running inference, which is why there are only ten points on the corresponding curves. However, the hit distance could be modified without the need for re-inference by filtering the output. The red data point on these curves signifies the threshold values we selected for the deployed model, or very close approximations. In the case of hit distance, there are numerous data points, and the red dot is located at approximately P = 0.318 and R = 0.88 on the far right.

G Discussions

SSL is promising for SAR object detection. Detailed comparative evaluations of the supervised and SSL models show that the SSL approaches suggest superior performance for SAR object detection. Nevertheless, detecting objects in urban areas continues to pose a significant challenge for

both SSL and supervised models. We foresee that the SSL approaches will gain more prominence as more SAR imagery becomes available for research. We hope to inspire wider object detection research in related challenges such as high data imbalance and tiny object detection, especially in the SAR domain.

Benchmarks and datasets are needed. We observe that more well-defined benchmarks and annotated multi-variate datasets are needed in this domain to have a more profound analysis and solutions to the problem. Although there are several datasets emerging for various other tasks such as segmentation and temporal analysis, the limited number and content of the existing datasets constitute a bottleneck for the advancements in the domain.

Ethical Considerations. The TRANSAR models in satellite-borne SAR imagery enhance object detection capabilities. We believe the results reported in this paper and advances made are of general interest to the researchers studying computer vision in remote sensing. Example applications include disaster management, environmental monitoring, and urban monitoring. However, this technology could be repurposed, e.g., for military applications. Potential misuse are not inherent to the TRANSAR models we introduce in this paper, but a risk that exists with SAR technology in general. We believe that it is of general interest to the community and the public to raise awareness of what is possible today. We recommend mitigating the risks associated to potential misuse by the implementation of strict access control and clear usage policies that ensure that the technology is used ethically and responsibly.