SARMAE: Masked Autoencoder for SAR Representation Learning Supplementary Material

Danxu Liu[1,4*], Di Wang[2,4*], Hebaixu Wang[3,4*], Haoyang Chen[2,4*], Wentao Jiang[2]
Yilin Cheng[4,5], Haonan Guo[4,6], Wei Cui[1†], Jing Zhang[2,4†]
[1]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[2]School of Computer Science, Wuhan University, Wuhan, China
[3]School of Electronic Information, Wuhan University, Wuhan, China
[4]Zhongguancun Academy, Beijing, China
[5]School of Data Science, Fudan University, Shanghai, China
[6]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

1. Overview

This supplementary material provides comprehensive details for the proposed SARMAE framework and the constructed SAR-1M dataset. These details were omitted from the main paper due to space constraints. The supplementary material is organized as follows:

- Section 2. Detailed composition and statistics of SAR-1M dataset;
- Section 3. Implementation details of Speckle-Aware Representation Enhancement (SARE);
- Section 7. Fine-tuning configurations for downstream tasks;
- Section 8. Extended visualization results;
- Section 9. Datasheet for SAR-1M.

2. Details for SAR-1M.

SAR-1M aggregates 18 publicly available SAR datasets, encompassing diverse imaging conditions, geographic locations, and task scenarios. Images in RSAR are the same in SAR-Det100k. Tab. 1 presents the detailed breakdown of each source dataset, including the image quantity, task type, image size, target type and spatial resolution. Datasets with paired SAR&OPT images are indicated in the last column, comprising 1,042,156 pairs in total.

SAR-1M encompasses 57 distinct categories, covering a diverse range of scene types. These categories span maritime objects, aerial targets, ground vehicles, infrastructure elements, land cover types, and event-related scenes, with visualizations of different scenarios, either SAR images or SAR–optical paired images, shown in Fig. 1-2.

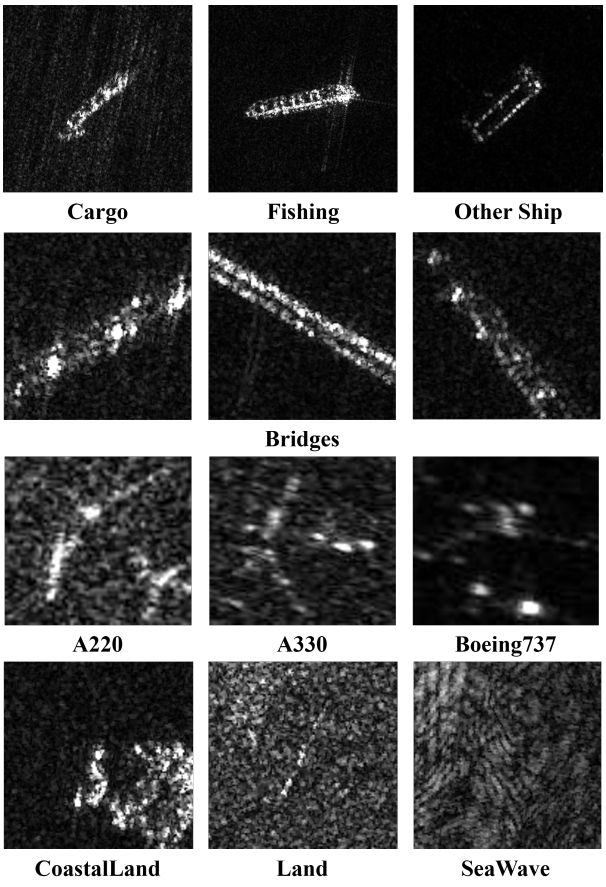

Figure 1. Various categories in SAR-1M.

SAR-1M employs hash-based deduplication to eliminate

[*]Equal contributions.
[†]Corresponding authors.

Table 1. Detailed composition of SAR-1M dataset.

Dataset	Year	Tasks	Imgs.	Img. Size (px)	Targets (Cls.)	Res. (m)	Modality
MSTAR[3]	1995	Cls.	14,577	128∼193	10	0.3	SAR
OpenSARShip[7]	2017	Cls.	26,679	9∼445	14	2.3∼17.4	SAR
SEN1-2[11]	2018	Det./Seg.	282,384	256	/	10	SAR&OPT
SAR-Ship[14]	2019	Det.	39,729	256	1	3∼25	SAR
FUSAR-ship[6]	2020	Cls.	5,243	512	10	/	SAR
HRSID[16]	2020	Det./Seg.	89,664	256	1	0.5∼3	SAR
SSDD[19]	2021	Det.	1,160	214∼668	1	1∼15	SAR
SADD[18]	2022	Det.	2,966	224	1	0.5∼3	SAR
MSAR[2]	2022	Det.	28,499	256∼2048	4	1	SAR
SAR-AIRcraft[15]	2023	Det.	18,818	512	7	1	SAR
SIVED[9]	2023	Det.	1,044	512	1	0.1∼0.3	SAR
OGSOD[22]	2023	Det./Seg.	22,366	256	3	3	SAR
SAR-Det100k[8]	2024	Det.	94,493	512	6	0.5∼25	SAR
RSAR[20]	2025	Det.	/	512	6	0.5∼25	SAR
M4_SAR[13]	2025	Det.	448,696	256	6	10,60	SAR&OPT
Bright[1]	2025	Seg.	149,872	256	3	0.3∼1	SAR&OPT
OpenEarthMap[17]	2025	Seg.	80,544	256	8	0.15∼0.5	SAR&OPT
AIR-PolSAR-Seg[23]	2025	Det./Seg.	6,168	512	6	8	SAR
SAR-1M	2025	Cls./Det./Seg.	1,312,902	9∼2048	57	0.1∼60	SAR&OPT

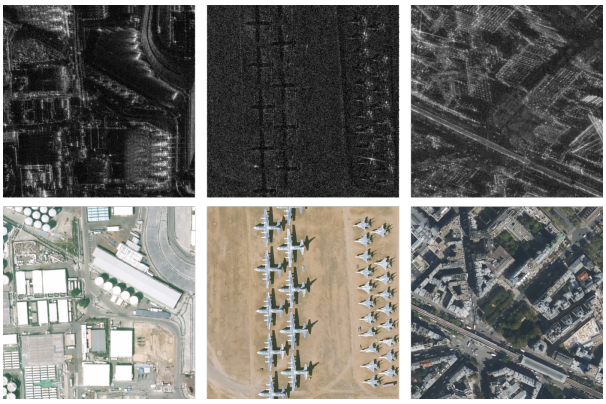

Figure 2. Diverse scenes with SAR-OPT pairs in SAR-1M.

any potential overlap with the test sets of the downstream datasets used in the main paper. Even when trained on only 30% of SAR-1M (Tab. 2), SARMAE-S still outperforms SUMMIT, demonstrating both the stronger capacity of our model and the benefits brought by increased data scale.

3. Implementation Details of SARE

To enable the model to learn noise-aware representations while preserving its ability to reconstruct clean SAR imagery, we implement SARE through a carefully designed noise injection strategy during pretraining. Unlike conventional denoising approaches that process all samples uniformly, we adopt a probabilistic augmentation scheme in which each training iteration carries a 50% chance of applying synthetic noise corruption to the input. This dynamic sampling mechanism allows the encoder to simultaneously learn to reconstruct clean SAR content and to handle diverse noise characteristics.

When an image is selected for augmentation, one of four physically motivated noise models is randomly chosen and applied.

The first category is additive Gaussian noise, which simulates random interference and is defined as:

$$x'(i,j) = x(i,j) + \mathcal{N}(0,\sigma^2), \tag{1}$$

where σ is randomly sampled from the range $[0.0, 0.5]$ to represent different noise levels.

The second category introduces multiplicative Rayleigh noise, which models the amplitude statistics of single-look SAR data:

$$x'(i,j) = x(i,j) \cdot \mathcal{R}(\sigma), \tag{2}$$

where $\mathcal{R}(\sigma)$ denotes a Rayleigh-distributed random variable with scale parameter σ sampled from $[0.0, 0.5]$.

The third noise type is Gamma-distributed multiplicative noise, representing the multi-look SAR intensity formation described in Eq.1 of the main text:

$$x'(i,j) \sim \text{Gamma}(L_{\text{syn}}, x(i,j)/L_{\text{syn}}), \tag{3}$$

where the synthetic look number L_{syn} is randomly selected from $1, 2, 3, 4$ to yield varying noise intensity levels.

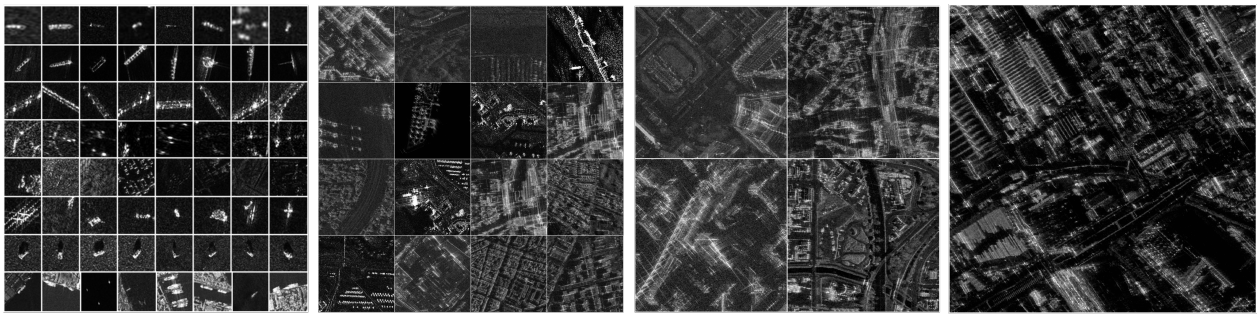

Figure 3. Images of different resolutions in SAR-1M.

Table 2. Comparison of performance across different datasets. All models adopt ViT-B as the backbone.

Model	Dataset	Dataset Scale	FUSAR_30% [6]	SAR-ACD [21]
SUMMIT [4]	MuSID	560k	71.91	84.25
SARMAE-S	SAR-1M(30%)	300k	88.27	93.76
SARMAE	SAR-1M	1000k	**92.92**	**95.06**

The fourth category is additive uniform noise, which simulates sensor-induced perturbations. It is formulated as:

$$x'(i,j) = x(i,j) + \mathcal{U}(-\alpha, \alpha), \tag{4}$$

where $\mathcal{U}(-\alpha, \alpha)$ denotes a uniform random variable drawn from the interval $[-\alpha, \alpha]$, and α is randomly sampled from $[0.0, 0.5]$ to control the perturbation magnitude.

Each chosen noise model further samples its related hyperparameters from the corresponding ranges, ensuring diverse corruption patterns both across iterations and within each training batch.

During training, if augmentation is applied, the corrupted image x' is masked and encoded following the MAE protocol, while the reconstruction target remains the original clean patch x. This formulation compels the encoder to map noisy, incomplete inputs back to clean and complete scene content. For iterations where augmentation is skipped, standard MAE reconstruction is performed without synthetic noise. This probabilistic dual-mode training paradigm enables SARMAE to learn noise-robust and semantically rich representations without compromising reconstruction quality or training stability.

Pretraining with pseudo-clean targets generated by SAR denoisers (LEE, BM3D, FROST) yields suboptimal classification accuracy (Tab. 3), as such targets often distort semantic structures. SARE learns to adaptively model complex SAR noise and capture task-relevant representations. As shown in Tab. 4, SARE outperforms both standard MAE and Gaussian denoising baselines. These results provide strong evidence that physics-driven noise modeling in SARE offers more robust SAR representations than standard augmentation strategies.

Table 3. Classification accuracy by pretraining on varied denoiser.

Denoiser	None	LEE	BM3D	FROST
MSTAR [3]	90.66	77.81	82.16	76.30

4. Implementation Details of SARC

We exclude optical images with cloud cover $\geq 20\%$ to ensure data quality. To avoid over-regularization from the optical modality, SARC is assigned a low loss weight ($\lambda = 0.1$), and additional unpaired SAR samples (0.3M) are incorporated to preserve feature autonomy in the SAR branch. As shown in Tab. 4, replacing SARC with contrastive loss leads to performance degradation, likely because speckle noise in SAR imagery makes negative pairs highly unreliable and interferes with stable feature alignment. In contrast, cosine similarity focuses on aligning positive pairs without aggressive repulsion, which is more suitable for bridging these distinct modalities. Moreover, Tab. 4 shows that SARE and SARC complement each other, yielding further performance gains when combined.

5. Rationale for DINOv3

DINOv3 [12] offers superior optical representations, compared to ImageNet pretrained MAE-ViT (see Tab. 5, both models adopt the Base version). We froze DINOv3 encoder to prevent it from adapting to SAR domain, which provides purely semantic guidance in the optical modality.

Table 4. Ablation study on noise addition and loss functions.

Model	Noise Addition	Loss Function	FUSAR_40 [6]	SSDD [19]	AIR-PolSAR-Seg [23]
MAE [5]	-	-	82.22	64.20	64.36
MAE + G	Gaussian	-	85.16	62.60	63.19
SARE	Ours	-	86.80	<u>64.40</u>	<u>65.15</u>
SARE + C	Ours	Contrastive Loss	<u>86.95</u>	62.40	63.41
SARMAE	Ours	SARC	**89.30**	**68.10**	**66.53**

Table 5. Comparison of different teachers for the Optical Branch. ViT-B is adopted as the backbone.

Optical Branch Init.	FUSAR_40 [6]	SAR-ACD [21]
MAE-ViT [5]	84.29	89.87
DINOv3 [12] (Ours)	89.30	95.06

6. Model Scalability Analysis

In terms of efficiency, scaling from ViT-B to ViT-L introduces a moderate training overhead (+33%) on 8×A800 GPUs (Tab. 6), indicating that SARMAE scales efficiently with model size. While ViT-L tends to overfit on smaller datasets (FUSAR-SHIP, MSTAR) due to its higher capacity, it consistently outperforms ViT-B on large-scale benchmarks (SARDet-100K). These results suggest that SARMAE follows standard scaling behavior, where larger models benefit from sufficient training data.

Table 6. Efficiency Analysis.

Model	Params (M)	FLOPs (G)	Training Cost (h)
ViT-B (Baseline)	86	17.6	∼45
ViT-L	307	61.6	∼60

7. Fine-tuning Configurations for Downstream Tasks

All experiments are conducted using the pretrained ViT-B and ViT-L backbones initialized with SARMAE weights. And all experiments are conducted on 8 NVIDIA A800 GPUs (40GB). We use PyTorch Distributed Data Parallel (DDP) for multi-GPU training. Gradient clipping with a maximum norm of 1.0 is applied across all tasks. For ViT-L models, we apply checkpoints to maintain the effective batch size when GPU memory is limited.

7.1. Target Classification

For target classification tasks, we evaluate SARMAE on three datasets: FUSAR-SHIP [6], MSTAR [3], and SAR-ACD [21]. The pretrained ViT encoder is adapted for classification by appending a global average pooling layer followed by a linear classification head. The number of output dimension in the linear layer corresponds to the number of classes in each dataset. The training configurations are detailed in Tab. 7. For the 40-shot experiments on FUSAR-SHIP and MSTAR, we randomly sample 40 images per class for training while using the full test set for evaluation. For the 30% labeled setting on FUSAR-SHIP, MSTAR and SAR-ACD, we randomly select 30% of the training data while keeping the test set unchanged. Each experiment is repeated 3 times with different random seeds, and we report the average accuracy.

7.2. Horizontal&Oriented Object Detection

For horizontal bounding box detection, we integrate the pretrained SARMAE backbone into the Faster R-CNN [10] framework with a Feature Pyramid Network (FPN) neck. We evaluate on two datasets: SSDD and SARDet-100k. And for oriented bounding box detection on the RSAR dataset, we adopt Oriented R-CNN as the detection framework, which extends Faster R-CNN with rotated Region of Interest (RoI) features and oriented bounding box regression. The training configurations are detailed in Tab. 7. To preserve the pretrained representations, we freeze all layers of the ViT backbone except the final layer during fine-tuning. This approach maintains the general SAR features learned during pretraining while allowing task-specific adaptation through the detection head.

7.3. Semantic Segmentation

For pixel-level semantic segmentation on the AIR-PolSAR-Seg dataset, we utilize UperNet as the segmentation framework. For the multi-class segmentation task, we report mean Intersection over Union (mIoU) across all categories. For the single-class water extraction task, we report the IoU for the water class. The training settings have been shown in Tab. 7.

8. Extended visualization results

Fig. 4 presents detection results on SSDD, SARDet-100k, and RSAR datasets. The visualizations demonstrate the model's capability to accurately localize ships in diverse scenarios, including multi-scale detection, dense

Table 7. Training configurations for different tasks.

Config	Classification	Detection	Segmentation
optimizer	AdamW	AdamW	AdamW
base learning rate	1.0×10^{-3}	1.0×10^{-4}	6.0×10^{-5}
weight decay	0.05	0.05	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$	$\beta_1, \beta_2 = 0.9, 0.95$	$\beta_1, \beta_2 = 0.9, 0.99$
batch size	25	16	4
learning rate schedule	Cosine	Step	Polynomial
warmup iterations	2000	1000	1500
warmup type	Constant	Linear	Linear
warmup learning rate	1.0×10^{-5}	0.33333	6.0×10^{-8}

harbor scenes, and oriented bounding box prediction for arbitrarily-oriented vessels.

Fig. 5 illustrates semantic segmentation results on AIR-PolSAR-Seg dataset. The model achieves precise pixel-level classification for multiple terrain categories and accurate water body extraction, demonstrating strong performance on fine-grained segmentation tasks.

9. Datasheets

9.1. Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

1. *"For what purpose was the dataset created?"*
   A: SAR-1M was created to address the lack of large-scale, diverse SAR datasets for self-supervised representation learning. Existing SAR datasets are limited in scale (100k-500k) and diversity, hindering the development of foundation models for SAR imagery.
2. *"Who created the dataset (e.g., which team, research group) and on behalf of which entity?"*
   A: The dataset was curated by us as part of research on SAR representation learning. It aggregates 18 publicly available SAR datasets.
3. *"Who funded the creation of the dataset?"*
   A: The dataset creation was funded by the affiliations of the authors involved in this work.

9.2. Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

1. *"What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?"*
   A: The dataset primarily comprises SAR imagery captured by satellites. All datasets utilized in SAR-1M are publicly accessible and nonprofit.
2. *"How many instances are there in total (of each type, if appropriate)?"*
   A: SAR-1M contains 1.3 million SAR image instances captured by satellites and 1 million paired SAR-OPT, 2.3 million in total.
3. *"Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?"*
   A: Yes, our dataset contains all possible instances that have been collected so far.
4. *"Is there a label or target associated with each instance?"*
   A: No, our dataset is intended for self-supervised learning. Therefore, each instance is an individual SAR/OPT image and does not contain annotations.
5. *"Is any information missing from individual instances?"*
   A: No.
6. *"Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?"*
   A: Yes, the relationship between individual instances is explicit.
7. *"Are there recommended data splits (e.g., training, development/validation, testing)?"*
   A: Yes, the entire dataset is intended for self-supervised methods, and we recommend using the whole dataset for self-supervised learning research.
8. *"Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?"*

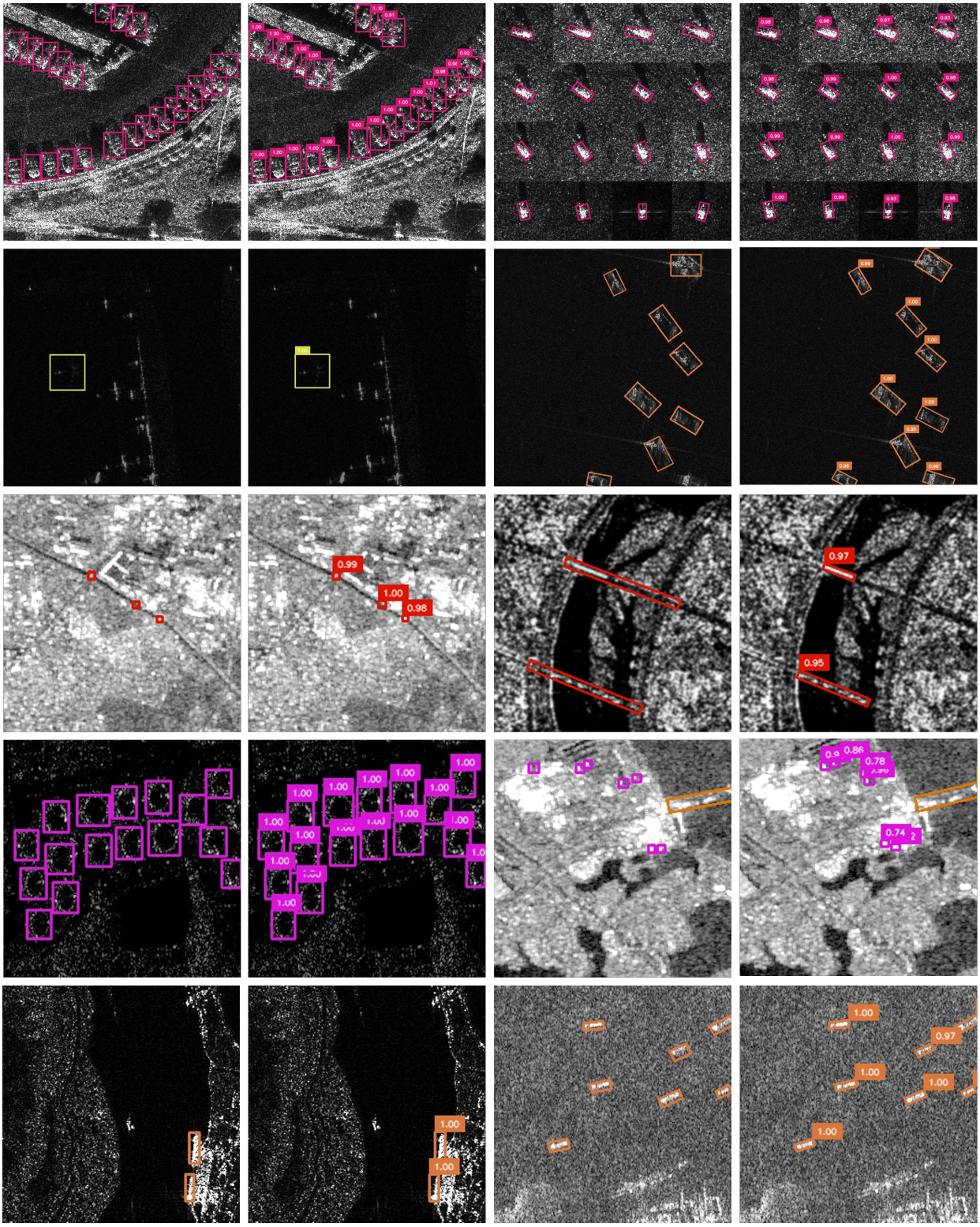

Figure 4. Object detection visualization results.

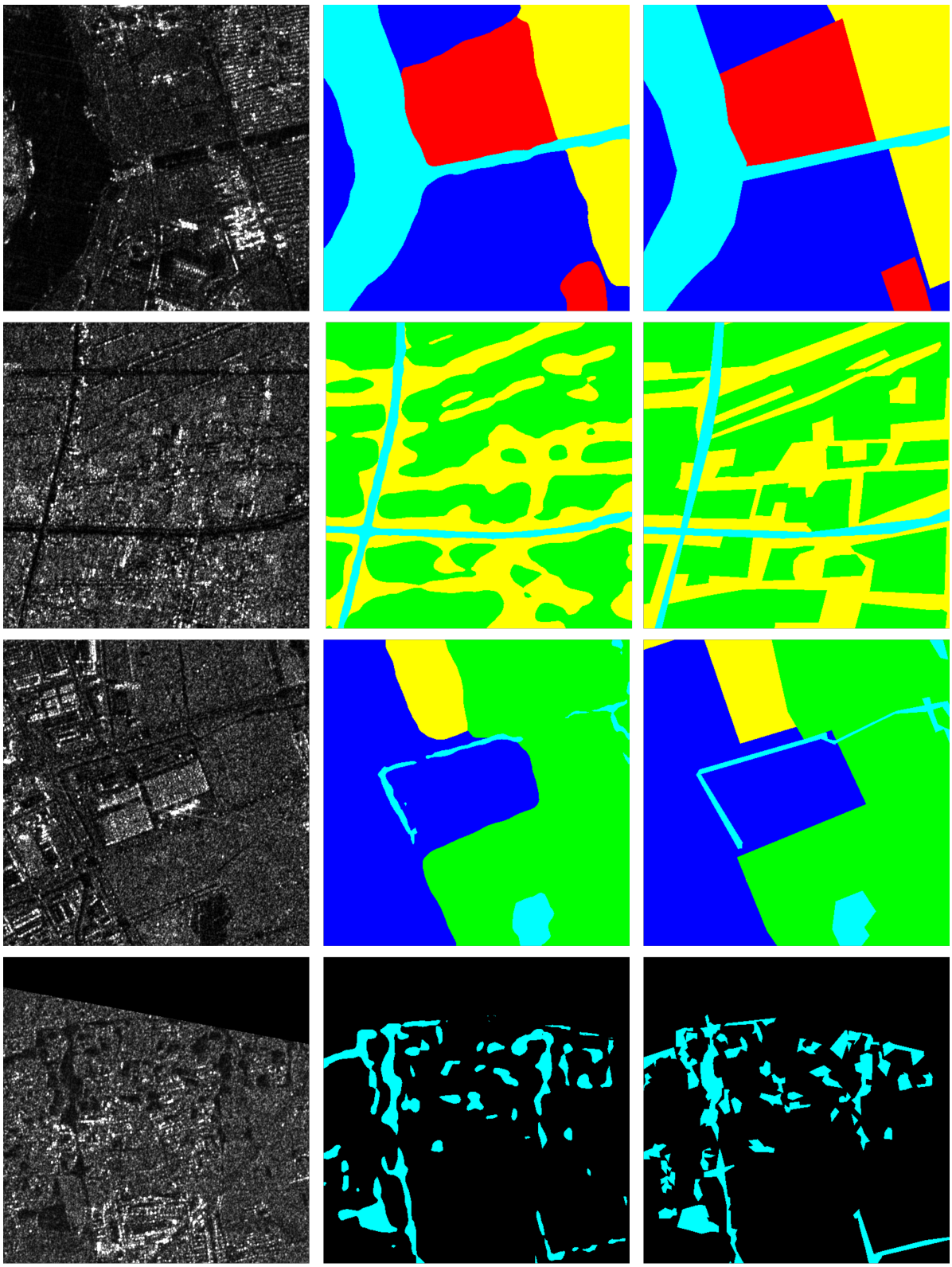

Figure 5. Semantic segmentation visualization results. Blue: Industrial. Green: Natural. Red: Land Use. Cyan: Water. White: Other. Yellow: Housing.

A: Yes, our dataset relies on many publicly available SAR datasets, which we have detailed in the main text.

9. *"Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?"*
   A: No, all data are clearly licensed.
10. *"Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?"*
    A: No.

9.3. Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

1. *"How was the data associated with each instance acquired?"*
   A: Please refer to the details listed in the main text Sec. 2.
2. *"What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?"*
   A: Please refer to the details listed in the main text Sec. 2.
3. *"If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?"*
   A: Please refer to the details listed in the main text Sec. 2.

9.4. Preprocessing, Cleaning, and Labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. *"Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?"*
   A: Yes, we preprocessed and cleaned data in our dataset.
2. *"Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?"*
   A: Yes, raw data is accessible.
3. *"Is the software that was used to preprocess/clean/label the data available?"*
   A: Yes, the necessary software used to preprocess and clean the data is publicly available.

9.5. Uses

The questions in this section are intended to encourage dataset creators to reflect on tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers make informed decisions, thereby avoiding potential risks or harms.

1. *"Has the dataset been used for any tasks already?"*
   A: No.
2. *"Is there a repository that links to any or all papers or systems that use the dataset?"*
   A: Not yet, but we will provide such links in our GitHub repository soon in the future.
3. *"What (other) tasks could the dataset be used for?"*
   A: The dataset could be used for training the SAR foundation models with the self-supervised learning method.
4. *"Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?"*
   A: N/A.
5. *"Are there tasks for which the dataset should not be used?"*
   A: N/A.

9.6. Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. *"Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?"*
   A: No.
2. *"How will the dataset be distributed (e.g., tarball on website, API, GitHub)?"*
   A: Very likely to be distributed by website, API, and GitHub repository.
3. *"When will the dataset be distributed?"*
   A: The datasets are publicly accessible, our SAR-1M will be publicly available soon.
4. *"Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?"*
   A: Yes, the dataset is under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
5. *"Have any third parties imposed IP-based or other restrictions on the data associated with the instances?"*
   A: No.
6. *"Do any export controls or other regulatory restrictions*

apply to the dataset or to individual instances?"
A: No.

9.7. Maintenance

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

1. *"Who will be supporting/hosting/maintaining the dataset?"*
   A: The authors of this work serve to support, host, and maintain the datasets.
2. *"How can the owner/curator/manager of the dataset be contacted (e.g., email address)?"*
   A: The curators can be contacted via the email addresses listed on our webpage.
3. *"Is there an erratum?"*
   A: There is no explicit erratum; updates and known errors will be specified in future versions.
4. *"Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?"*
   A: Yes, for the current version. Future updates (if any) will be posted on the dataset website.
5. *"Will older versions of the dataset continue to be supported/hosted/maintained?"*
   A: Yes. This is the first version of the release; future updates will be posted and older versions will be replaced.
6. *"If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?"*
   A: Yes, we provide detailed instructions for future extensions.

References

[1] Hongruixuan Chen, Jian Song, Olivier Dietrich, Clifford Broni-Bediako, Weihao Xuan, Junjue Wang, Xinlei Shao, Yimin Wei, Junshi Xia, Cuiling Lan, et al. Bright: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *Earth System Science Data Discussions*, 2025:1–51, 2025. 2

[2] Jie Chen, Zhixiang Huang, Runfan Xia, Bocai Wu, Lei Sheng, Long Sun, and Baidong Yao. Large-scale multi-class sar image target detection dataset-1.0. *Journal of Radars*, 14: 1488, 2022. 2

[3] Joseph R Diemunsch and John Wissinger. Moving and stationary target acquisition and recognition (mstar) model-based automatic target recognition: Search technology for a robust atr. In *Algorithms for Synthetic Aperture Radar Imagery V*, pages 481–492. SPIE, 1998. 2, 3, 4

[4] Yuntao Du, Yushi Chen, Lingbo Huang, Yahu Yang, Pedram Ghamisi, and Qian Du. Summit: A sar foundation model with multiple auxiliary tasks enhanced intrinsic characteristics. *International Journal of Applied Earth Observation and Geoinformation*, 141:104624, 2025. 3

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 4

[6] Xiyue Hou, Wei Ao, Qian Song, Jian Lai, Haipeng Wang, and Feng Xu. Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, 63(4):140303, 2020. 2, 3, 4

[7] Boying Li, Bin Liu, Lanqing Huang, Weiwei Guo, Zenghui Zhang, and Wenxian Yu. Opensarship 2.0: A large-volume dataset for deeper interpretation of ship targets in sentinel-1 imagery. In *SAR in Big Data Era: Models, Methods and Applications*, pages 1–5. IEEE, 2017. 2

[8] Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *NeurIPS*, 37:128430–128461, 2024. 2

[9] Xin Lin, Bo Zhang, Fan Wu, Chao Wang, Yali Yang, and Huiqin Chen. Sived: A sar image dataset for vehicle detection based on rotatable bounding box. *Remote Sensing*, 15 (11), 2023. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 4

[11] Michael Schmitt, Lloyd Hughes, and Xiao Xiang Zhu. The sen1-2 dataset for deep learning in sar-optical data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 141–146, 2018. 2

[12] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3, 4

[13] Chao Wang, Wei Lu, Xiang Li, Jian Yang, and Lei Luo. M4-sar: A multi-resolution, multi-polarization, multi-scene, multi-source dataset and benchmark for optical-sar fusion object detection. *arXiv preprint arXiv:2505.10931*, 2025. 2

[14] Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. A sar dataset of ship detection for deep learning under complex backgrounds. *Remote Sensing*, 11(7):765, 2019. 2

[15] Zhirui Wang, Yuzhuo Kang, Xuan Zeng, et al. Sar-aircraft-1.0: High-resolution sar aircraft detection and recognition dataset. *Journal of Radars*, 12(4):906–922, 2023. 2

[16] Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8: 120234–120254, 2020. 2

[17] Junshi Xia, Hongruixuan Chen, Clifford Broni-Bediako, Yimin Wei, Jian Song, and Naoto Yokoya. Openearthmap-sar: A benchmark synthetic aperture radar dataset for global high-resolution land cover mapping. *arXiv preprint arXiv:2501.10891*, 2025. 2

[18] Peng Zhang, Hao Xu, Tian Tian, Peng Gao, Linfeng Li, Tianming Zhao, Nan Zhang, and Jinwen Tian. Sefepnet: Scale

expansion and feature enhancement pyramid network for sar aircraft detection with small sample dataset. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3365–3375, 2022. 2

[19] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021. 2, 4

[20] Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, and Xiang Li. Rsar: Restricted state angle resolver and rotated sar benchmark. *CVPR*, 2025. 2

[21] Chenxi Zhao, Daochang Wang, Siqian Zhang, and Gangyao Kuang. Global discriminative information search and focus for sar target recognition. *IEEE Sensors Journal*, 25(9): 15735–15749, 2025. 3, 4

[22] Zhicheng Zhao, Changfu Zhou, Yu Zhang, Chenglong Li, Xiaoliang Ma, and Jin Tang. Text-guided coarse-to-fine fusion network for robust remote sensing visual question answering. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:1–17, 2025. 2

[23] WANG Zhirui, ZHAO Liangjin, WANG Yuelei, ZENG Xuan, KANG Jian, YANG Jian, and SUN Xian. Air-polsar-seg-2.0: Polarimetric sar ground terrain classification dataset for large-scale complex scenes. *Journal of Radars*, 14(2): 353–365, 2025. 2, 4