# SEORation: Curating SAR-EO Paired Data for Multi-Modal Remote Sensing Foundation Models

Seojun Kim*, Youngtack Oh*, Yejun Lee, Sohee Son†
SI Analytics
Daejeon, Republic of Korea
{seojun.kim, ytoh96, yejun.lee, shson}@si-analytics.ai

## Abstract

*SAR-EO paired data provide complementary supervision for multi-modal remote sensing foundation models. However, simply aggregating SAR-EO pairs from multiple sources can introduce semantic redundancy and weakly aligned cross-modal pairs. To address these issues, we propose **SEORation**, a two-stage pipeline for curating SAR-EO pairs. SEORation first performs remote-sensing-aware semantic deduplication using RemoteCLIP embeddings, and then prioritizes pairs with stronger scene-level compatibility through our proposed RSLIP score filtering. In this work, we release **OpenSEP** (Open SAR-EO Pairs), a 4.9M-pair multi-source SAR-EO data pool, and **OpenSEP-1.7M**, a curated subset selected from this pool according to retrieval performance on the OpenSEP validation split. We also provide empirical validation of SEORation through curation and retrieval experiments, demonstrating that the proposed pipeline improves SAR-EO pair selection for multi-modal pretraining. On QXS-SAROPT, external evaluation further shows that the RSLIP model trained on OpenSEP-1.7M improves the aggregate retrieval score from 445.10 to 510.06 compared with the raw candidate pool. These results highlight the importance of paired-data curation for reliable cross-modal alignment in SAR-EO multi-modal pretraining.*

## 1. Introduction

In remote sensing, EO (Electro-Optical) and SAR (Synthetic Aperture Radar) images provide complementary observations of the Earth. EO imagery captures rich visual and spectral characteristics of a scene, whereas SAR imagery provides active microwave observations that are relatively robust to illumination and weather conditions [10, 20]. Since the two modalities respond to different physical properties, their paired observations can provide complementary cues for scene understanding and representation learning. Recent multi-modal remote sensing models have therefore leveraged the complementarity of SAR-EO through cross-modal alignment, reconstruction, and temporal modeling [7, 11].

As remote sensing models scale toward foundation models, the role of data becomes increasingly central. Large-scale multi-modal pretraining depends not only on model architecture or training objectives, but also on the construction of the pretraining corpus itself [8]. Despite this importance, SAR-EO pretraining studies have largely treated the paired corpus as a fixed input to the model [16]. In practice, existing corpora are often simply constructed by aggregating available data sources to increase scale and coverage. However, such aggregation is not a neutral operation: each source reflects its own collection objective and acquisition setting, and naively merging them can over-represent certain sample distributions while introducing SAR-EO pairs with uneven cross-modal correspondence. In this work, we address these issues by analyzing SAR-EO paired data curation along two axes: sample-level semantic redundancy and pair-level consistency.

At the sample-level, redundancy arises from the way remote sensing corpora are constructed. Remote sensing datasets are often derived from large scenes, tiles, and patch archives [5, 23]. When such sources are cropped, sampled, or merged, spatially adjacent areas, overlapping coverage, and homogeneous land-cover patterns can produce many samples with highly similar semantics [9, 15]. This effect is particularly pronounced in broad ocean, cropland, grassland, and forest regions, where distinct patches may provide limited additional information in the representation space. In a merged SAR-EO corpus, such redundancy can over-represent certain sample distributions, increase training cost, and reduce the effective diversity of the pretraining data. We therefore aim to reduce excessive semantic redundancy while preserving useful diversity in SAR-EO paired data.

*Equal contribution.
†Corresponding author. Email: shson@si-analytics.ai

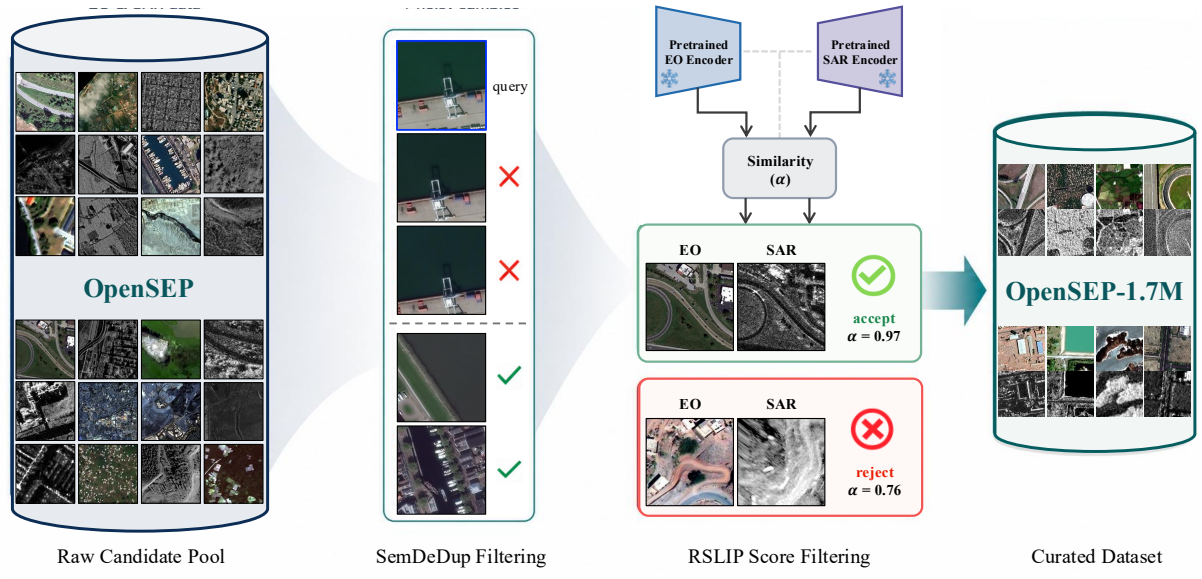


Figure 1. Overview of **SEORation**. Starting from **OpenSEP**, a raw SAR-EO candidate pool, SEORation first reduces sample-level semantic redundancy through SemDeDup filtering. It then applies RSLIP score filtering to prioritize SAR-EO pairs with stronger pair-level compatibility. The retained pairs form **OpenSEP-1.7M**, the final curated subset for SAR-EO multi-modal pretraining.

At the pair-level, geospatial co-location alone does not ensure reliable correspondence. A SAR-EO pair can serve as useful cross-modal supervision only when the two observations convey compatible scene content. However, modality-induced differences, such as radiometric discrepancies, geometric distortions, and residual registration errors, as well as acquisition-induced factors, such as acquisition-time gaps, surface changes, cloud contamination, and preprocessing mismatches, can undermine this condition [32]. When such imperfect correspondences are treated as positive pairs, they may provide misleading supervision, analogous to noisy correspondence in cross-modal matching [13]. We therefore examine pair-level consistency as a criterion for reliable SAR-EO supervision.

Based on this sample-level and pair-level curation view, we propose **SEORation**, a two-stage pipeline for curating SAR-EO paired data. The first stage applies semantic deduplication to reduce excessive redundancy while preserving SAR-EO pair structure. The second stage performs RSLIP score filtering to estimate semantic consistency between the two modalities. We further release **OpenSEP** (Open SAR-EO Pairs), a 4.9M-pair open SAR-EO data pool for multi-modal pretraining, and derive **OpenSEP-1.7M** as its SEORation-curated subset. Our experiments demonstrate that the proposed curation strategy is effective not only for in-domain, but also for out-of-domain generalization, using strength-varied curated subsets and evaluated under consistent bidirectional SAR-EO retrieval protocols.

## 2. Related Work

### 2.1. Multi-Modal Remote Sensing Foundation Models

Remote sensing foundation models have increasingly moved from single-modality pretraining toward multi-modal learning with SAR, EO, and temporal imagery. CROMA [7] learns radar-optical representations by combining cross-modal contrastive learning with masked reconstruction on spatially and temporally aligned SAR-optical samples. SkySense [11] and SkySense++ [27] study multi-modal remote sensing foundation models that integrate optical, SAR, and temporal imagery for broad Earth observation tasks. MaRS [31] further explores very-high-resolution SAR-optical foundation modeling with cross-modality and cross-granularity representation learning, while X-JEPA [3] studies predictive self-supervised alignment for cross-modal remote sensing retrieval. These studies demonstrate the importance of SAR-EO multi-modal representation learning, while our work focuses on the paired-data curation problem that precedes such pretraining.

### 2.2. SAR-EO Data Resources

A number of public resources provide SAR-EO or related multi-modal remote sensing data across different sensors, resolutions, and acquisition conditions. Sentinel-1/Sentinel-2 resources such as SSL4EO-S12 [26] and BigEarthNet-MM [23] support large-scale medium-resolution representation learning. DynamicEarthNet [24] provides daily Planet Fusion imagery and monthly Sentinel-

1/Sentinel-2 auxiliary imagery over the same areas of interest, enabling data fusion and multi-modal settings. High-resolution and heterogeneous SAR-EO resources include GUSO [30] and OSDataset2.0 [28] for SAR-optical registration and matching, SpaceNet6 [21] and the IEEE GRSS Data Fusion Contests [14, 18] for multi-sensor remote sensing benchmarks, and M4-SAR [25] and BRIGHT [2] for optical-SAR detection and disaster-response scenarios. OpenSEP integrates these public resources into a unified raw candidate pool and constructs patchified SAR-EO candidate pairs through a common preprocessing pipeline.

### 2.3. Data Curation for Foundation Model Training

Data curation has become an important factor in large-scale foundation model training. DataComp [8] shows that data selection strategies can substantially affect image-text contrastive pretraining under fixed training and evaluation protocols. MetaCLIP [29] analyzes CLIP [19] pretraining from a data-centric perspective and proposes metadata-balanced curation, while Data Filtering Networks [6] studies learned filtering for selecting useful subsets from large uncurated pools. SemDeDup [1] is especially relevant to our work because it identifies semantic duplicates in embedding space rather than exact duplicates in input space, showing that redundant samples can be removed while preserving performance and improving training efficiency. SEORation extends this curation perspective to SAR-EO paired data in remote sensing. In addition to sample-level redundancy, SAR-EO corpora require pair-level validation because geographically corresponding samples can still provide ambiguous cross-modal supervision. This is related to noisy correspondence in cross-modal matching, where mismatched pairs can degrade representation learning [13].

## 3. Method

### 3.1. Overview

We construct **OpenSEP** by first normalizing public remote sensing datasets into a unified SAR-EO candidate pool. We then apply **SEORation**, a two-stage curation pipeline consisting of semantic deduplication and RSLIP score filtering. The first stage reduces sample-level semantic redundancy while preserving the SAR-EO pair structure. The second stage uses a GUSO-trained RSLIP scorer to estimate scene-level compatibility between SAR and EO images. Fig. 1 illustrates the overall pipeline.

To study the effect of curation choices, we vary the semantic embedding space used for semantic deduplication and the pruning strength used for RSLIP score filtering. Each subset is used to train the same RSLIP architecture under a controlled protocol, and the resulting models are evaluated with SAR-to-EO and EO-to-SAR retrieval. This separates the use of RSLIP as a curation scorer from its use as a SAR-EO foundation model trained on raw or curated paired corpora. We refer to the final SEORation-curated subset as **OpenSEP-1.7M**, and use **OpenSEP** to denote the released SAR-EO paired data resource.

Table 1. Composition of the OpenSEP training pool and source-provided held-out pools. The held-out column reports publicly available paired validation or test samples used as OpenSEP validation pools, and "–" indicates that no official paired held-out split is provided.

| Dataset | # Train | # Held-out |
|---|---|---|
| SpaceNet6 [21] | 43,109 | – |
| BigEarthNet-MM [23] | 600,326 | – |
| DynamicEarthNet [24] | 606,432 | 15,330 |
| DFC2023 [18] | 30,484 | 2,146 |
| SSL4EO-S12 [26] | 1,044,316 | – |
| M4-SAR [25] | 312,912 | 33,946 |
| OSDataset2.0 [28] | 15,212 | – |
| DFC2025 [14] | 69,328 | – |
| BRIGHT [2] | 51,824 | 156 |
| GUSO [30] | 2,160,572 | 48,994 |
| **Total** | **4,934,515** | **100,572** |

### 3.2. OpenSEP Candidate Pool Construction

To construct OpenSEP, we aggregate public remote sensing datasets that provide paired or pairable SAR and EO observations. As summarized in Table 1, the raw candidate pool aggregates ten public resources and contains 4.9M patchified SAR-EO candidate pairs. These aggregated sources cover diverse sensors, spatial resolutions, acquisition conditions, scene types, and geographic regions. Since these datasets were originally collected for different tasks and sources, we treat the aggregated data as a raw candidate pool rather than a directly curated pretraining corpus. We convert each source into a unified patch-level SAR-EO format while retaining available metadata such as dataset origin, sensor type, spatial resolution, and acquisition information. After patchification and train-split construction, the raw pool with **4.9M SAR-EO candidate pairs** is then passed to SEORation.

### 3.3. Semantic Deduplication

The first stage of SEORation reduces sample-level semantic redundancy in the SAR-EO candidate pool. We perform deduplication at the pair level using EO-side embeddings as the semantic anchor. When an EO sample is identified as redundant, its associated SAR-EO pair is removed together. For each EO image $I_i$, rather than treating redundancy as pixel-level near-duplication, we extract a RemoteCLIP embedding [17] to capture remote-sensing semantics and apply $\ell_2$ normalization:

$$\mathbf{z}_i = \frac{f(I_i)}{\|f(I_i)\|_2}.$$

Following SemDeDup [1], we cluster the normalized embeddings and perform pruning independently within each cluster. For a sample at position $j$ in a cluster, we compute its maximum cosine similarity to the preceding samples in the same cluster:

$$s_j = \max_{1 \le i < j} \mathbf{z}_i^\top \mathbf{z}_j,$$

with $s_1 = 0$. Given a pruning parameter $\varepsilon$, the corresponding SAR-EO pair is removed if

$$s_j > 1 - \varepsilon.$$

We construct the deduplicated candidate pool with $\varepsilon = 0.07$, corresponding to a cosine similarity threshold of $0.93$. The retained pairs are then passed to the RSLIP score filtering stage.

### 3.4. RSLIP Score Filtering

Motivated by the pair-level inconsistency discussed above, for the second stage of SEORation, we apply RSLIP score filtering to estimate the compatibility of each SAR-EO pair. For this, we propose **RSLIP**, a cross-modal contrastive learning model that aligns SAR and EO images in a shared semantic embedding space. Inspired by CLIP [19], RSLIP adopts a dual image-encoder architecture consisting of an EO encoder and a SAR encoder. Given a mini-batch of paired samples $\mathcal{B} = \{(x_i^{eo}, x_i^{sar})\}_{i=1}^N$, RSLIP produces normalized embeddings $z_i^{eo}$ and $z_i^{sar}$ and is trained with a symmetric contrastive objective:

$$\mathcal{L}_{eo \to sar} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(z_i^{eo}, z_i^{sar})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(z_i^{eo}, z_j^{sar})/\tau)},$$

$$\mathcal{L} = \frac{1}{2} \left( \mathcal{L}_{eo \to sar} + \mathcal{L}_{sar \to eo} \right),$$

where $\tau$ denotes the temperature parameter.

For RSLIP score filtering, we first train a **GUSO-trained RSLIP scorer**. GUSO [30] is a large-scale SAR-EO registration dataset originally introduced for SAR-optical image registration and provides high-quality cross-modal correspondences across diverse cities and scenes. This makes it suitable for learning an initial SAR-EO compatibility scorer. Since OpenSEP intentionally spans diverse sensors, spatial resolutions, and acquisition conditions, we train the scorer with random cropping, blur, downsampling, and geometric jitter to reduce reliance on high-resolution texture or strict registration cues. The trained scorer is then applied to OpenSEP candidate pairs to produce the RSLIP scores used for filtering.

For each deduplicated candidate pair $(x^{eo}, x^{sar})$, we compute the RSLIP score as the cosine similarity between the normalized EO and SAR embeddings:

$$s(x^{eo}, x^{sar}) = z^{eo\top} z^{sar}.$$

A higher score indicates stronger compatibility in the learned embedding space. We use this score as a filtering signal rather than an absolute quality label. For each semantically deduplicated pool, we construct multiple curated subsets by varying the RSLIP score filtering strength. Specifically, we retain the top-$p\%$, with $p \in \{10, 30, 50, 70\}$ candidate pairs according to the RSLIP score. These subsets are used to train the same RSLIP architecture under a controlled protocol to analyze how pair-level filtering strength affects SAR-EO pretraining.

## 4. Experiments

We evaluate the effect of SEORation on SAR-EO paired-data curation through cross-modal retrieval experiments. We first describe the common training setup and retrieval protocol, and then use the OpenSEP validation pool to sequentially select the curation strengths for semantic deduplication and RSLIP score filtering. We analyze the RSLIP model trained on the resulting curated dataset, OpenSEP-1.7M. Finally, we evaluate SAR-to-EO and EO-to-SAR retrieval on QXS-SAROPT, an external SAR-EO dataset that is not used in OpenSEP construction, to examine whether the selected curation configuration transfers beyond the source datasets.

Table 2. Training configuration used for all RSLIP pretraining experiments.

| **Setting** | **Value** |
|---|---|
| Model | ViT-B/16 |
| Input size | $224 \times 224$ |
| Optimizer | AdamP |
| Learning rate | $5 \times 10^{-4}$ |
| Weight decay | $1 \times 10^{-4}$ |
| Batch size | 128 |
| Training epochs | 25 |
| Backbone freezing | First 2 epochs |
| LR schedule | Constant for 15 epochs, then $0.1\times$ decay |
| Gradient clipping | 2.0 |

### 4.1. Experimental Setup

All pretraining experiments are conducted under a consistent setup of [4], while varying only the training data. We compare models trained on the raw OpenSEP candidate pool, semantically deduplicated pools, and curated subsets obtained after RSLIP score filtering. This controlled setup allows us to analyze how each curation stage and curation strength affect SAR-EO representation learning.

We use an epoch-matched training protocol for all models. Each subset is trained for the same number of epochs, so smaller curated subsets require fewer optimization steps than larger pools. This setting evaluates whether curated data can provide more effective training signals under the

Table 3. Validation retrieval results for semantic deduplication and RSLIP score filtering. Semantic deduplication is evaluated on OpenSEP validation pools from BRIGHT, DFC2023, DynamicEarthNet, GUSO, and M4-SAR. For RSLIP score filtering, filtered subsets are constructed from the RemoteCLIP $\varepsilon = 0.07$ deduplicated subset; the Top 100% baseline is omitted since it is identical to the RemoteCLIP $\varepsilon = 0.07$ column. The final score is reported as dataset-size-weighted R@sum, where each dataset-level R@sum is weighted by its number of validation pairs. Underlined values indicate the best result within semantic deduplication, while bold values indicate the best result within RSLIP score filtering.

| **Dataset** | **Semantic Deduplication** | | | **RSLIP Score Filtering** | | | |
|---|---|---|---|---|---|---|---|
| | **Raw** | **DINOv3** ($\varepsilon$=0.07) | **RemoteCLIP** ($\varepsilon$=0.07) | **Top 10%** | **Top 30%** | **Top 50%** | **Top 70%** |
| BRIGHT | 576.92 | <u>582.05</u> | 567.95 | 581.41 | **592.31** | **592.31** | 588.46 |
| DFC2023 | 513.98 | 517.89 | <u>519.71</u> | 484.30 | 523.11 | 523.16 | **524.23** |
| DynamicEarthNet | 7.38 | <u>8.70</u> | 7.99 | 3.54 | 6.57 | 6.41 | **7.16** |
| GUSO | 466.72 | 483.01 | <u>489.62</u> | 412.83 | 526.02 | **532.82** | 528.21 |
| M4-SAR | 406.58 | 403.46 | <u>412.47</u> | 347.07 | 442.44 | **454.08** | 427.00 |
| **Weighted R@sum** | 377.58 | 384.76 | <u>390.93</u> | 330.03 | 418.67 | **425.89** | 414.63 |

same epoch budget. The common training configuration is summarized in Table 2.

Curation strengths are selected using the OpenSEP validation pools listed in Table 1, which are constructed from the held-out splits available in the source datasets. For each validation pool, we evaluate both EO-to-SAR and SAR-to-EO retrieval and report Recall@1, Recall@5, and Recall@10. We define R@sum as the sum of Recall@1, Recall@5, and Recall@10 over both retrieval directions. Since the validation pools differ in size across sources, we aggregate validation performance using dataset-size-weighted R@sum, where each dataset-level R@sum is weighted by the number of validation pairs in that dataset.

For external evaluation, we use QXS-SAROPT [12], a SAR-EO paired dataset that is not included in OpenSEP construction. Given a query image from one modality, the model retrieves its paired counterpart from the other modality. We report Recall@1, Recall@5, Recall@10, and R@sum for both EO-to-SAR and SAR-to-EO retrieval.

### 4.2. Effect of Embedding Space for Semantic Deduplication

This section analyzes the embedding space used for semantic deduplication, the first stage of SEORation. Since semantic deduplication removes samples based on similarity in an embedding space, the choice of encoder directly affects which SAR-EO pairs are considered redundant. Remote sensing imagery differs from natural imagery in viewing geometry, scene layout, and land-cover patterns, suggesting that a remote-sensing-aware representation may provide a more suitable redundancy criterion.

To examine this effect, we fix the pruning strength to $\varepsilon = 0.07$ and compare RemoteCLIP [17] with DINOv3 [22]. RemoteCLIP provides remote-sensing-aware embeddings, whereas DINOv3 provides general-purpose visual embeddings. For each embedding space, we apply semantic deduplication to the raw OpenSEP candidate pool, train the same RSLIP architecture on the resulting pool, and evaluate SAR-to-EO and EO-to-SAR retrieval on the OpenSEP validation pools.

Table 3 reports the validation retrieval results. Both semantic deduplication settings improve over the raw candidate pool, which achieves a weighted R@sum of 377.58. RemoteCLIP-based deduplication achieves the highest weighted R@sum of 390.93, outperforming DINOv3-based deduplication at 384.76. This suggests that remote-sensing-aware embeddings provide a more effective space for identifying redundant SAR-EO samples. Based on this result, we use the RemoteCLIP $\varepsilon = 0.07$ deduplicated pool for the subsequent RSLIP score filtering stage.

### 4.3. Selecting the RSLIP Score Filtering Strength

After selecting the RemoteCLIP $\varepsilon = 0.07$ pool from the semantic deduplication stage, we select the filtering strength for the second stage of SEORation. For each SAR-EO pair, a GUSO-trained RSLIP scorer computes a compatibility score. We then construct curated subsets by retaining the top-$p\%$ pairs according to the RSLIP score, where $p \in \{10, 30, 50, 70\}$.

Table 3 reports OpenSEP validation retrieval results for each RSLIP score-filtered subset. The Top 50% subset achieves the highest weighted R@sum, while the Top 30% and Top 70% subsets also outperform the unfiltered Top 100% base pool. In contrast, the Top 10% subset performs substantially worse, indicating that overly aggressive filtering can remove useful diversity. These results suggest that RSLIP score filtering improves SAR-EO alignment when it balances pair-level reliability and data diversity.

Based on this validation result, we select RemoteCLIP $\varepsilon = 0.07$ semantic deduplication followed by RSLIP Top 50% score filtering as the final curation configuration. We use OpenSEP to denote the released SAR-EO

Table 4. Source composition of **OpenSEP-1.7M**. The number of train patches denotes the patchified SAR-EO candidate pairs retained after RemoteCLIP $\varepsilon = 0.07$ semantic deduplication and RSLIP Top 50% score filtering. The keep rate denotes the proportion of retained pairs relative to the raw candidate pool of each source dataset.

| Dataset | # Train patches | Keep Rate |
|---|---|---|
| SpaceNet6 [21] | 32,300 | 74.93% |
| BigEarthNet-MM [23] | 9,557 | 1.59% |
| DynamicEarthNet [24] | 1,250 | 0.21% |
| DFC2023 [18] | 12,219 | 40.08% |
| SSL4EO-S12 [26] | 313,629 | 30.03% |
| M4-SAR [25] | 139,833 | 44.69% |
| OSDataset2.0 [28] | 13,037 | 85.70% |
| DFC2025 [14] | 12,471 | 17.99% |
| BRIGHT [2] | 16,731 | 32.28% |
| GUSO [30] | 1,153,061 | 53.37% |
| **Total** | **1,704,088** | **34.53%** |

paired data resource, and **OpenSEP-1.7M** to denote this final SEORation-curated subset used in our experiments. Table 4 reports the source composition of OpenSEP-1.7M, which retains 1.70M pairs, corresponding to 34.53% of the raw candidate pool. The source-wise keep rates vary because SEORation filters candidates by semantic redundancy and RSLIP scene-level compatibility rather than enforcing a fixed per-source ratio.

### 4.4. External Evaluation on QXS-SAROPT

Finally, we evaluate retrieval performance on QXS-SAROPT [12], an external SAR-EO paired dataset that is not included in OpenSEP construction. This evaluation examines whether the curation configuration selected on the OpenSEP validation pools remains effective beyond the source datasets used to build OpenSEP.

Table 5 reports SAR-to-EO and EO-to-SAR retrieval results on QXS-SAROPT. The raw candidate pool serves as the uncurated baseline. The Top 30% subset achieves the best R@sum on QXS-SAROPT. However, the validation-selected OpenSEP-1.7M configuration, corresponding to Top 50%, is second-best in R@sum and SAR-to-EO metrics, remains competitive in EO-to-SAR retrieval, and substantially outperforms both the raw candidate pool and the unfiltered Top 100% subset. This indicates that the curation strength selected on OpenSEP validation transfers well to an external SAR-EO dataset. Although the external optimum shifts from Top 50% to Top 30%, moderate RSLIP score filtering consistently outperforms both the raw candidate pool and the unfiltered Top 100% subset.

### 4.5. Qualitative Analysis of RSLIP Scores

We also qualitatively examine how RSLIP scores reflect compatibility between SAR and EO images. Fig. 2 shows examples of SAR-EO pairs from the OpenSEP candidate pool sorted by their RSLIP scores. Low-score pairs often exhibit weak structural or semantic correspondence across modalities. For example, major terrain or object layouts may differ, or structures visible in one modality may not have clear counterparts in the other. In contrast, high-score pairs tend to show more consistent scene-level structures, such as roads, buildings, water bodies, and terrain patterns, across SAR and EO observations.

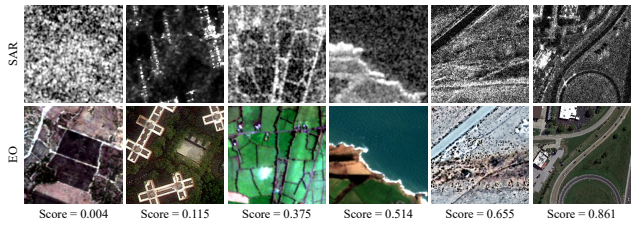


Figure 2. Qualitative examples of SAR-EO pairs sorted by the RSLIP score. The score increases from left to right. Low-score pairs show weak or ambiguous cross-modal correspondence, whereas high-score pairs exhibit clearer consistency between SAR and EO observations.

Table 5. External retrieval results on QXS-SAROPT. Top-$p$% subsets are constructed from the RemoteCLIP $\varepsilon = 0.07$ deduplicated pool using RSLIP scores. Top 100% denotes the deduplicated base pool before RSLIP score filtering, and **OpenSEP-1.7M** denotes the validation-selected Top 50% subset. R@sum is computed as the sum of R@1, R@5, and R@10 over both retrieval directions. Best results are shown in bold and second-best results are underlined.

| Subset | EO→SAR | | | SAR→EO | | | R@sum |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| Raw candidate pool | 67.68 | 82.23 | 86.42 | 56.35 | 73.55 | 78.89 | 445.10 |
| Top 100% | 72.48 | 84.77 | 88.16 | 52.78 | 70.91 | 76.86 | 445.95 |
| Top 70% | 78.98 | <u>87.97</u> | **90.44** | 75.14 | 85.51 | 88.57 | 506.59 |
| **OpenSEP-1.7M** (Top 50%) | <u>79.01</u> | 87.88 | 90.28 | <u>77.00</u> | <u>86.61</u> | <u>89.28</u> | <u>510.06</u> |
| Top 30% | **79.66** | **88.06** | <u>90.36</u> | **78.67** | **86.94** | **89.38** | **513.07** |
| Top 10% | 54.98 | 74.05 | 79.99 | 48.98 | 68.69 | 75.64 | 402.34 |

## 5. Conclusion

In this paper, we studied the importance of paired-data curation for SAR-EO multi-modal remote sensing pretraining and proposed SEORation, a two-stage curation pipeline for this purpose. SEORation reduces sample-level semantic redundancy through semantic deduplication using RemoteCLIP embeddings and filters SAR-EO pairs with low compatibility through our proposed RSLIP score filtering. We further aggregated and curated multiple public SAR-EO and related remote sensing resources to construct OpenSEP-1.7M, a curated SAR-EO paired data resource constructed from public sources. In our experiments, we used the OpenSEP validation pool to sequentially select the curation subsets for semantic deduplication and RSLIP score filtering. On QXS-SAROPT, an external evaluation dataset,

the RSLIP model trained on OpenSEP-1.7M improves the aggregate SAR-to-EO and EO-to-SAR retrieval score from 445.10 to 510.06 compared with the raw candidate pool. These results show that constructing a large SAR-EO paired corpus is not sufficient by itself, and that curation considering both semantic redundancy and pair-level compatibility is important for learning reliable cross-modal alignment.

**Limitations and Future Work.** Several limitations remain. Our semantic deduplication analysis is limited to a fixed pruning threshold and a small set of embedding spaces, and a more comprehensive study across deduplication strengths and downstream tasks is left for future work. Although we mitigate this through augmentation and external evaluation on QXS-SAROPT, future work could explore more diverse scorer training sources or scorer ensembles. We also plan to evaluate OpenSEP and SEORation beyond retrieval, including dense prediction, fusion, change detection, and disaster response.

## Acknowledgements

This work was supported by the National IT Industry Promotion Agency (NIPA) grant funded by the Korea government (MSIT) through the Advanced GPU Utilization Support Program (No. 02-26-01-0197).

## References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 3, 4

[2] H. Chen, J. Song, O. Dietrich, C. Broni-Bediako, W. Xuan, J. Wang, X. Shao, Y. Wei, J. Xia, C. Lan, K. Schindler, and N. Yokoya. BRIGHT: a globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *Earth System Science Data*, 17(11):6217–6253, 2025. 3, 6

[3] Shabnam Choudhury, Yash Salunkhe, Vaibhav Rajan, Subhasis Chaudhuri, and Biplab Banerjee. X-JEPA: A novel joint learning cross-modal predictive alignment framework for remote sensing image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4355–4364, 2026. 2

[4] Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024. 4

[5] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 1264–1268. IEEE, 2025. 1

[6] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T. Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, 2024. 3

[7] Anthony Fuller, Koreen Millard, and James R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems*, 2023. 1, 2

[8] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, 2023. 1, 3

[9] Daniel A. Griffith and Yongwan Chun. Spatial autocorrelation and uncertainty associated with remotely-sensed data. *Remote Sensing*, 8(7):535, 2016. 1

[10] Chengyan Guo, Zhiyuan Zhang, Kexin Huang, Lan Luo, Ziqing Yang, Shuyun Shi, and Junpeng Shi. Deep learning methods for SAR and optical image fusion: A review. *Remote Sensing*, 18(8):1196, 2026. 1

[11] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[12] Meiyu Huang, Yao Xu, Lixin Qian, Weili Shi, Yaqin Zhang, Wei Bao, Nan Wang, Xuejiao Liu, and Xueshuang Xiang. The QXS-SAROPT dataset for deep learning in SAR-optical data fusion. *arXiv preprint arXiv:2103.08259*, 2021. 5, 6

[13] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems*, 2021. 2, 3

[14] IEEE GRSS Image Analysis and Data Fusion Technical Committee. 2025 IEEE GRSS data fusion contest: All-weather land cover and building damage mapping. https://www.grss-ieee.org/community/technical-committees/2025-ieee-grss-data-fusion-contest/, 2025. Accessed: 2026-05-02. 3, 6

[15] N. Karasiak, J.-F. Dejoux, C. Monteil, and D. Sheeren. Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 111:2715–2740, 2022. 1

[16] Danxu Liu, Di Wang, Hebaixu Wang, Haoyang Chen, Wentao Jiang, Yilin Cheng, Haonan Guo, Wei Cui, and Jing Zhang. Sarmae: Masked autoencoder for sar representation learning. *arXiv preprint arXiv:2512.16635*, 2025. 1

[17] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou.

RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3, 5

[18] Claudio Persello, Ronny Hänsch, Gemine Vivone, Kaiqiang Chen, Zhiyuan Yan, Deke Tang, Hai Huang, Michael Schmitt, and Xian Sun. 2023 IEEE GRSS data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction. *IEEE Geoscience and Remote Sensing Magazine*, 11(1):94–97, 2023. 3, 6

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4

[20] Michael Schmitt, Florence Tupin, and Xiao Xiang Zhu. Fusion of SAR and optical remote sensing data: Challenges and recent trends. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5458–5461, 2017. 1

[21] Jacob Shermeyer, Adam Van Etten, Daniel Hogan, Ryan Lewis, Max Ehrlich, Saksham Gupta, Tao Chen, Logan Montgomery, Robert Mook, Laila Bashmal, et al. Spacenet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3, 6

[22] 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. 5

[23] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3): 174–180, 2021. 1, 2, 3, 6

[24] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andres Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Caglar Senaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixe. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21135, 2022. 2, 3, 6

[25] 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. 3, 6

[26] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 2, 3, 6

[27] Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, Qi Zhu, Jian Wang, Ming Yang, Jingdong Chen, Yongjun Zhang, and Yansheng Li. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, 7:1235–1249, 2025. 2

[28] Yuming Xiang, Jinyang Chen, Zhonghua Hong, Niangang Jiao, Feng Wang, Hongjian You, and Xiaohua Tong. OS-Dataset2.0: SAR-optical image matching dataset and evaluation benchmark. *Journal of Radars*, 2025. In press. 3, 6

[29] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations*, 2024. 3

[30] Heng Yan, Ailong Ma, Hong Shu, Yuting Wan, Liangpei Zhang, and Yanfei Zhong. Ultra-high-resolution SAR and optical image registration: From global benchmark dataset to frequency-guided registration method. *ISPRS Journal of Photogrammetry and Remote Sensing*, 235:190–210, 2026. 3, 4, 6

[31] Ruoyu Yang, Yinhe Liu, Heng Yan, Yiheng Zhou, Yihan Fu, Han Luo, and Yanfei Zhong. MaRS: A multi-modality very-high-resolution remote sensing foundation model with cross-granularity meta-modality learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11685–11693, 2026. 2

[32] Wenfei Zhang, Ruipeng Zhao, Yongxiang Yao, Yi Wan, Peihao Wu, Jiayuan Li, Yansheng Li, and Yongjun Zhang. Multi-resolution sar and optical remote sensing image registration methods: A review, datasets, and future perspectives. *arXiv preprint arXiv:2502.01002*, 2025. 2