Double Acquisition Neural Network for Matching Marine Debris Patches Across PlanetScope and Sentinel-2 Imagery

Gabrielė Tijūnaitytė[a]

[a] *Wageningen University, Wageningen, The Netherlands*

Abstract

The gradual accumulation of plastic waste in marine environments threatens biodiversity and human welfare. Monitoring this waste can inform policy evaluation and support waste recovery efforts. While remote sensing studies have primarily focused on detecting marine debris (MD), a proxy for plastic waste, in isolated satellite scenes, tracking MD agglomerations (patches) over time remains unexplored. This study presents the Double Acquisition Neural Network (NN), designed to match MD patches across PlanetScope (PS) and Sentinel-2 (S2) imagery captured within a one-hour interval. A new dataset of 3,445 annotated PS-S2 MD patch pairs was compiled to support the model development. The study systematically examined different designs of similarity estimation frameworks, conducted a hyperparameter search, and further investigated retrieval performance gains from restricting the candidate search scope, aided by varying prior knowledge on local drift. The optimal model, featuring two platform-specific ResNet-18 encoders initialised with pre-trained weights and trained with contrastive loss, achieved 37% top-1 and 57% top-3 retrieval accuracies under a local (bound to a study event) candidate selection regime. Incorporating drift knowledge into the candidate selection process significantly improved performance: the top-1 accuracy reached 62% and the top-3 accuracy reached 88%. The mean retrieval position of the true match was reduced from 12.6 to 1.9. These findings demonstrate that the proposed Double Acquisition NN can reliably match MD patches within two retrieval attempts, aided by drift knowledge. This work lays the foundation for automating MD patch tracking and enabling time-aware monitoring of marine plastic pollution.

Keywords: Marine Debris, Multi-Modal Features, Feature Similarity, Contrastive Learning

1. Introduction

1.1. Marine Plastics Pollution

Plastic production has increased disproportionately since its invention, reaching over 410 million metric tons (Mt) in 2023 alone (Plastics Europe, 2024). However, only a small portion of the generated plastic is recycled, leaving millions of metric tons as waste. Lebreton and Andrady (2019) estimated that in 2015, 60 to 99 Mt of plastic waste was improperly disposed of, some of which eventually found its way into the ocean. In 2010, an estimated 4.8 to 12.7 Mt of plastic waste entered the ocean (Jambeck et al., 2015), primarily via rivers, which are estimated to transport 1.15 to 2.41 Mt of plastic to the ocean each year (Lebreton et al., 2017).

Due to the long lifespan of plastic materials (Chamas et al., 2020), once plastic waste enters the ocean, it persists and accumulates, contributing to long-term pollution (Barnes et al., 2009; Andrady, 2015). This persistent pollution is a growing concern due to its potential direct and indirect adverse effects on marine biodiversity, ecosystems, and human well-being (Gall and Thompson, 2015; Carbery et al., 2018; Waring et al., 2018). Entanglement, ingestion, and laceration from plastic can lead to inflammation, starvation, and death in marine organisms (Gall and Thompson, 2015; Carbery et al., 2018). Moreover, plastic can enter food chains, posing broader risks across trophic levels (Waring et al., 2018). Furthermore, toxic additives used initially to improve plastic properties may be released into the water through degradation processes such as UV photo-oxidation, chemical and bacterial weathering, and abrasion from water and wind (Carbery et al., 2018; Waring et al., 2018).

The scale and persistence of marine plastic pollution call for urgent strategies to mitigate its harmful impact. Although reducing plastic production, consumption, and waste generation is the most direct solution to the plastic waste crisis, even substantial reductions will not eliminate the large quantities of plastic already discarded in the ocean (Barnes et al., 2009).

Monitoring is considered a key strategy for managing existing marine plastic pollution, as it advances our understanding of the global plastic inventory, its spatial distribution, and dispersal patterns (Maximenko et al., 2019; Karakuş, 2023). This knowledge is essential for supporting targeted waste recovery operations, identifying pollution sources, and enabling effective law enforcement (Kikaki et al., 2020; Karakuş, 2023). Moreover, long-term monitoring is critical for evaluating the effectiveness of mitigation policies and international agreements (Eriksen et al., 2023).

1.2. In Situ Monitoring and Plastic Dispersal Modelling

In situ monitoring through shipboard sampling or trawl studies provides direct and accurate measurements of plastic waste presence and concentration. However, these methods are expensive and require extensive human resources. Consequently, in situ efforts are spatially and temporally sparse and are often restricted to coastal areas (van Sebille et al., 2015; Kikaki et al., 2020; Salgado-Hernanz et al., 2021; Goddijn-Murphy et al., 2024).

To address the scarcity of direct measurements and obtain global estimates of plastic waste quantities, accumulation zones, and dispersal patterns, researchers employ the physical ocean drift models to simulate plastic movement (van Sebille et al., 2015; Lermusiaux et al., 2019; Bajon et al., 2023; Zhang et al., 2023). The resulting plastic dispersal models integrate oceanographic parameters such as currents and wind with available plastic concentration data to model the probable transport and fate of plastic.

Although these dispersal models are valuable tools, their accuracy is limited by multiple factors. Firstly, available in situ plastic measurement datasets often lack interoperability due to inconsistent methodologies (Goddijn-Murphy et al., 2024), hindering their integration and use for model calibration and validation. Secondly, this accuracy depends on the quality of the incorporated drift modelling approaches, which often rely on coarse-resolution oceanographic parameter inputs (Maximenko et al., 2019; Eriksen et al., 2023; Zhang et al., 2023) and may use simplifications within drift modelling frameworks (e.g., neglecting Stokes drift; Bosi et al. (2021)). Furthermore, drift model calibration often leverages tracked data from surface drifter buoys, which do not fully represent floating plastics due to differences in size, shape, density (Pereiro et al., 2018), and drag properties, particularly due to the presence of drogues (Béchaz, 2024). These limitations restrict the applicability of drift models for simulating plastic waste dispersal and trajectory forecasting, especially in complex coastal environments (Pereiro et al., 2018; Li et al., 2020).

Several studies have stressed the need for more observational data on plastic waste quantity and movement behaviours to improve the calibration and validation of plastic dispersal models and to enable more reliable monitoring and forecasting of plastic waste for various applications, yet collecting such data through in situ methods alone remains severely limited (van Sebille et al., 2015; Cózar et al., 2021; Zhang et al., 2023; Béchaz, 2024).

1.3. Remote Sensing Monitoring

An alternative data source to resource-intensive and sparse in situ observations is satellite-based Remote Sensing (RS), which has the potential to provide additional data for model calibration and validation. The high-resolution RS provides non-intrusive, repeatable and global measurements that offer extensive insights into spatial and temporal patterns of marine plastic waste (Biermann et al., 2020; Salgado-Hernanz et al., 2021; Topouzelis et al., 2021; Karakuş, 2023).

Although using RS for marine plastic waste monitoring is a relatively recent development (Salgado-Hernanz et al., 2021), several RS-based methods have been proposed for detecting marine plastic waste in satellite imagery. These include spectral indices (e.g., Themistocleous et al. (2020); Cózar et al. (2024)), classical machine learning algorithms (e.g., Basu et al. (2021); Duarte and Azevedo (2023)), emerging deep learning approaches (e.g., Mifdal et al. (2021); Rußwurm et al. (2023); Shen et al. (2024); Dalsasso et al. (2025)), and combinations of these techniques (e.g., Biermann et al. (2020); Lavender (2022); Sannigrahi et al. (2022); Kikaki et al. (2024)).

However, most studies face the challenge of insufficiently detailed RS data for direct plastic detection and therefore use Marine Debris (MD) as a proxy for plastic waste. Plastic has a weak and highly variable spectral signal ascribed to low concentrations, diverse plastic material types, degradation, and biofouling effects (Garaba et al., 2018; Cózar et al., 2021; Politikos et al., 2023). These factors make it challenging to distinguish plastic waste from other MD materials such as sea foam, seaweed, plankton, and other natural and non-natural debris (Biermann et al., 2020; Topouzelis et al., 2020, 2021; Mikeli et al., 2022; Karakuş, 2023). These materials, including plastic waste, often cluster into agglomerations—MD patches—that form more detectable targets for satellite sensors due to their larger size, which is sufficient for identification from space (Biermann et al., 2020). Moreover, plastic is often the most common non-natural component of MD patches, making these agglomerations particularly relevant for plastic waste detection (Biermann et al., 2020; Cózar et al., 2021).

Current RS approaches for plastic waste monitoring via MD patch detection rely on sensors aboard satellite missions such as Sentinel-1, Sentinel-2 (S2), Landsat, PlanetScope (PS), and WorldView-3 (Salgado-Hernanz et al., 2021; Karakuş, 2023; Politikos et al., 2023). Among these, S2 is the most commonly used due to its free availability, relatively high spatial resolution (up to 10 m) and abundant spectral information (in 13 spectral channels), which together enable the detection of small MD aggregations (Topouzelis et al., 2019; Biermann et al., 2020; Kikaki et al., 2020; Themistocleous et al., 2020; Topouzelis et al., 2020). In contrast, the PS platform offers scenes at a higher spatial resolution (3 m), but with fewer spectral channels (4, and 8 since 2020), making detecting plastic waste within MD patches more challenging. Moreover, as PS data is commercial, it has been used less frequently in MD detection studies (e.g. Kikaki et al. (2020); Shen et al. (2024); Dalsasso et al. (2025)).

While both S2 and PS platforms are effectively used to detect MD patches in individual scenes, each presents limitations for continuously tracking these patches across multi-temporal imagery. The five-day revisit time of S2 constrains its ability to re-detect MD patches on short

timescales. In contrast, despite its near-daily coverage, PS is limited by its spectral richness, reducing its standalone reliability. These limitations constrain the use of either platform alone for monitoring dynamic processes of the MD drift and dispersal over time.

Temporally coordinated overpasses of PS and S2 missions present a notable opportunity for time-aware MD monitoring, facilitated by frequent, near-simultaneous (typically within a one-hour interval) imagery acquisitions. These PS-S2 scene pairs, referred to as "double acquisitions", offer a unique opportunity to complement S2's spectral richness with PS's temporal frequency and spatial details. Kikaki et al. (2020) has noted the potential to leverage multiple satellite platforms for MD patch monitoring and tracking over time.

1.4. Tracking Through Double Acquisitions

Despite the potential of using multi-temporal satellite imagery for MD tracking, its implementation is largely underexplored. Although a few studies have attempted to manually infer MD patch trajectories and drift velocities from multi-temporal scenes from various platforms (Matthews et al., 2017; Kikaki et al., 2020; Weiß et al., 2022), to the best of the author's knowledge, no study has yet exploited the opportunity of using the PS-S2 double acquisitions to develop an automated MD patch tracking system. Furthermore, since neither the individual platforms nor their combination has previously been applied to track MD patches, no platform-specific or multi-platform dataset of MD patch annotations from double acquisitions currently exists.

This study draws inspiration from established object tracking pipelines commonly used in other domains to address the lack of time-aware monitoring and to leverage double acquisitions for automated MD patch tracking. For example, Ahn et al. (2023) proposed a traffic surveillance pipeline consisting of three modules: an object detection module (i.e., for cars), an object association module to match the same objects across video frames, and a trajectory estimation module.

Inspired by Ahn et al. (2023), an MD patch tracking pipeline could consist of: MD patch detectors for PS and S2 scenes, an MD object association module for matching detected MD patches across PS-S2 double acquisitions, and a trajectory estimation module (see Figure A.1 for the tracking pipeline composition illustration). Building on the extensive efforts toward addressing the MD detection problem, this study assumes the availability of reliable detectors for both PS and S2 imagery and focuses exclusively on the next step of the tracking pipeline: developing an object association module for MD patches detected in double acquisitions.

The association step addresses a fundamental retrieval problem: how to accurately retrieve the relevant candidate among a set of supplied candidates for each query. For this study, it can be reframed as a multi-platform retrieval task: for each MD patch detected in the PS scene (the query), the objective is to find (retrieve) its corresponding match in the S2 scene (the relevant candidate) from a set of detected MD patches (the candidate set).

In Ahn et al. (2023), object association is performed by ranking candidates using deep-learning-based similarity estimates between the query and each candidate. The retrieval process is further facilitated by constraining the candidate set to reduce the risk of false positives. This study adopts a similar two-factor strategy for the MD patch matching task by implementing candidate selection regimes to reduce the number of potential candidates and ranking candidates using similarity scores derived from the proposed Double Acquisition Neural Network (NN).

This study offers two key contributions. First, this study compiled the first dataset of tracked MD patches derived from double-acquisition imagery pairs of PS and S2, as outlined in Section 2. This dataset enabled the second contribution: developing a Double Acquisition Neural Network (NN) to match MD patches across PS and S2 double acquisition scenes (Section 3 outlines its architecture and training procedures). The latter contribution involved answering the following research questions (RQs):

RQ1 Which of the similarity estimation frameworks (outlined in Sections 3 and 4) performs the best for the MD patch matching task?

RQ2 What are the optimal training hyperparameters for the Double Acquisition NN in terms of batch size, augmentation strategies, base encoder depth, and pre-trained weight initialisation?

RQ3 How can the retrieval performance of the Double Acquisition NN be improved by restricting the candidate search space?

2. Dataset

2.1. Study Events

The MD patches in the curated dataset were annotated based on MD events previously reported in the literature and general media, which, upon inspection, were found to have PS-S2 double acquisitions (see Appendix B for specific scene IDs). The selected MD events include:

- Bay Islands, Honduras, 2017-10-09. Kikaki et al. (2020) noted a floating plastic debris event validated by in situ measurements on 2017-10-17.
- Venice, Italy, 2018-06-30 from Mifdal et al. (2021) study.
- Calabria, Italy, 2018-10-22. The MD patches formed following a severe flooding and outwash of trash into the ocean. Sannigrahi et al. (2022) found an agreement of spectral response curves for suspected plastic presence.

- Accra, Ghana 2018-10-31. According to Biermann et al. (2020), the MD patches captured in these scenes are primarily composed of macroalgae and spume, with some exhibiting the spectral signature of plastic. Rußwurm et al. (2023) has previously noted the existing PS-S2 double acquisition pair.
- Venice, Italy, 2018-10-31. MD patches visible just outside the Venetian Lagoon can be linked with a storm-induced flooding event in Venice at the end of October, 2018 (The Guardian, 2018).
- Lagos, Nigeria, 2019-01-01 from Mifdal et al. (2021).
- Durban, South Africa, 2019-04-24. It was reported that large amounts of plastic were washed out following a flood. The plastic presence was confirmed in situ (Biermann et al., 2020). Unfortunately, due to large cloud coverage, only a few double acquisition annotations of MD patches were possible.
- Thassos, Greece, 2021-04-30. The PS-S2 double acquisition pair was found based on Cózar et al. (2024) MD presence predictions.
- Marmara, Turkey, 2021-05-19 from Rußwurm et al. (2023). MD patches are relics of a large mucilage bloom, which, just like plastic, pose severe harm concerns (Yagci et al., 2022).

2.2. Data Acquisition and Pre-processing

Satellite scenes were retrieved for each reported MD event to enable the annotation of MD patches and provide data for the model training.

PlanetScope Scenes

PS scenes were manually downloaded from the Planet website by requesting analytical products filtered by MD event date and location. The downloaded scenes have a spatial resolution of 3 m, are projected to the local UTM coordinate system, and consist of 4 spectral channels: Red, Green, Blue (RGB) and Near-Infrared (NIR).

All acquired scenes were converted to top-of-atmosphere reflectance (TOARef) (see Appendix C for details). The pre-processed PS scenes were used to generate bounding boxes aggregated per study event.

Sentinel-2 Scenes

The corresponding double acquisition S2 scenes for each study event were downloaded from Google Earth Engine. The MD event date and bounding box coordinates were used as request parameters. The retrieved S2 scenes were Level-1 products, containing TOARef measurements across 13 spectral channels. Spectral channels with coarser spatial resolution were upscaled to match the highest available spatial resolution of 10 m. Additionally, the S2 scenes were re-projected to the local UTM coordinate system to ensure spatial alignment with the corresponding PS scenes.

Atmospheric Correction

There is no common agreement on the atmospheric correction (AC) application and its effectiveness, as many researchers highlight the risk of disrupting the plastic spectral signal due to AC (Garaba et al., 2018; Goddijn-Murphy et al., 2018; Biermann et al., 2020; Themistocleous et al., 2020; Topouzelis et al., 2020, 2021; Hu, 2022; Karakuş, 2023; Cózar et al., 2024). Additionally, ensuring that multi-platform ACs produce comparably corrected PS and S2 scenes is challenging (Karakuş, 2023). While AC can be deemed important for studies focusing on the spectral properties of MD, in this study, the short time interval between double acquisitions is expected to limit atmospheric condition variability. Therefore, both PS and S2 scenes are assumed to be affected by the atmosphere in a similar manner. Based on this, the decision not to apply ACs was made for both PS and S2 scenes.

2.3. Annotation Process

Although researchers have proposed several spectral indices to highlight MD patches and plastic waste in S2 scenes (Themistocleous et al., 2020; Karakuş, 2023), such as the Floating Debris Index (FDI; Biermann et al. (2020)), these indices often rely on spectral channels that are not available in PS imagery. Therefore, a simpler and consistent across-platform index was chosen to aid the MD patch annotation process. This study employed a Normalised Difference Index (NDI), which was calculated for each scene using the Blue and NIR spectral channels (Eq. 1). This index was chosen based on visual inspection in preliminary experiments.

$$NDI = \frac{\text{NIR} - \text{Blue}}{\text{NIR} + \text{Blue}} \tag{1}$$

Pairs of PS-S2 scenes and their corresponding NDI rasters were loaded into QGIS and visualised side by side. MD patches were annotated as pairs using the line tool, where each line started at the centre of the MD patch in the PS scene and ended at the same MD patch captured in the S2 scene (Figure 1). The spatial displacement between MD patch locations is attributed to drift occurring over the acquisition time gap, which in this study spanned from 2 to 46 minutes. Annotations were created for each MD patch in the PS scene when a corresponding match was identifiable in the S2 scene. In many cases, the emerging drift patterns, often forming parallel or gradually changing drift displacement trajectories, aided the annotation process (see Figure 1).

Line segments were post-processed to their endpoints, representing MD patch coordinates in the PS and S2 scenes. The original line segments served to associate matched MD patches into match pairs.

2.4. Annotation Dataset

The final dataset contains 3,445 match pair annotations of MD patches re-detected across PS-S2 double acquisitions (Table 1). These pairs correspond to 3,445 MD

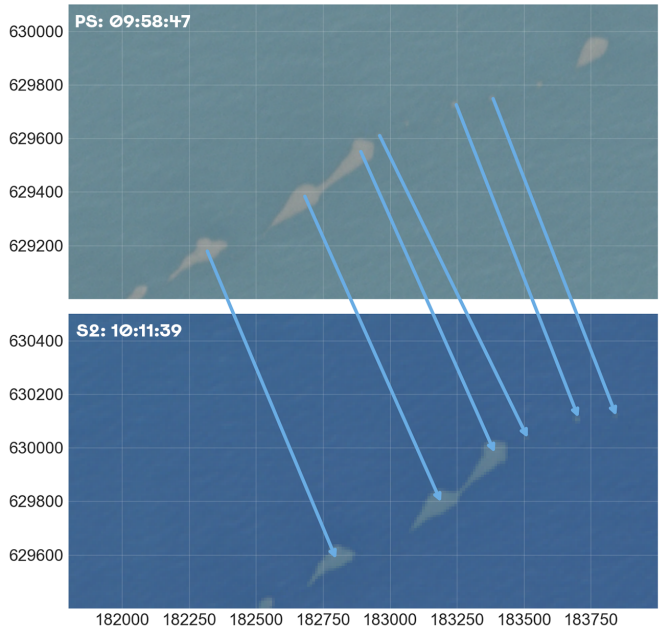

Figure 1: Example of Marine Debris (MD) patch pair annotations for the MD event in Accra, Ghana, on 2018-10-31. During the annotation process, PlanetScope (PS) and Sentinel-2 (S2) scenes were visualised side-by-side, and arrows were placed to point from an MD patch centre in the PS scene to its corresponding location in the S2 scene. Note that in this figure, the arrow annotations do not objectively reflect the drift displacement, as both images have overlapping extents.

patch locations annotated in PS scenes and 3,080 in S2 scenes. The full dataset is published alongside this study, and more technical details about the dataset structure and composition can be found in Appendix B.

The discrepancy between the number of annotations from each platform is due to the different number of scenes available per study event. Due to the longer revisit time of the S2 platform, only one S2 scene is available for each MD event. In contrast, the PS platform captured the same event in multiple, sometimes overlapping scenes. This overlap allowed the same MD patch to be annotated multiple times in the PS scenes but only once in the S2 scene. Consequently, there were multiple double acquisition pairs for some MD patches. For example, the same MD patch locations PS_1, PS_2 and $S2_1$ produced two match pairs: $(PS_1, S2_1)$ and $(PS_2, S2_1)$.

2.5. Dataset Implementation

Annotations from the compiled dataset were used to generate multispectral tensor tiles of MD patches, which served as inputs for model training and evaluation.

Tensor Tile Extraction

MD patch annotations were used as centre anchors to crop tensor tiles from pre-processed PS and S2 scenes (see Section 2.2). PS tiles were cropped to 256×256 pixels, and S2 tiles to 64×64 pixels. These dimensions were selected to ensure approximately similar spatial coverage, while also retaining computational efficiency benefits of input sizes that are powers of two. This tile creation strategy was designed to centre the MD patch within each tile while also including both the spectral and spatial context from the surrounding area.

Table 1: The number of annotations in the training, validation, and test sets of MD patches that were re-detected across the PS-S2 double acquisitions, categorised by the corresponding MD events.

	MD event	Matches & Patches in PS	Patches in S2
Train	Bay Islands, Honduras	115	108
	Venice (2018-06), Italy	163	163
	Calabria, Italy	182	173
	Accra, Ghana	67	67
	Venice (2018-10), Italy	154	152
	Lagos, Nigeria	15	15
	Durban, South Africa	13	13
	Marmara, Turkey	1415	1096
	Thassos, Greece	42	42
	Total	**2166**	**1829**
Validate	Accra, Ghana	365	350
	Marmara, Turkey	240	240
	Total	**605**	**590**
Test	Accra, Ghana	278	268
	Marmara, Turkey	396	393
	Total	**674**	**661**

Both PS and S2 TOARef scenes are originally scaled by 10^4 to store reflectance values as integer data type instead of float data type, reducing memory usage. However, such scaling is suboptimal for neural network optimisation, as large input values can lead the activations, such as sigmoid and tanh, into their saturation regimes, producing diminished gradients and consequently slowing convergence (Ioffe and Szegedy, 2015). To address this, all tile values were rescaled by dividing by 10^4. Since reflectance is a proportional unit between reflected and incident radiation, it has an inherent range of $[0, 1]$.

Tiles were grouped into PS-S2 pairs and assigned match labels. A pair was labelled positive if the same MD patch was centred in both tiles, and negative if the tiles contained different MD patches.

Data Split

To reflect variability across MD events while preserving spatial independence, the Marmara and Accra MD events were spatially divided into training, validation, and test sets (see Figure D.1 for an illustration of the spatial extents). All remaining MD events were used exclusively for training. This spatial division ensured that model performance is evaluated on spatially distinct areas of the Accra and Marmara MD events, promoting a more robust assessment of generalisation. The resulting data split included

2,166 MD patch pairs for training, 605 for validation, and 674 for testing (Table 1).

3. Methodology

This section presents the architecture, optimisation and training strategies of the **proposed Double Acquisition NN**, identified as the most effective model, based on experimental evaluation (detailed later), for estimating similarity values of MD patches and ranking candidates.

3.1. Model Architecture

Since matching MD patches are represented by PS and S2 image tiles, originating from different satellite platforms, with varying spatial resolutions and spectral characteristics, estimating pairwise similarity between them constitutes a heterogeneous (multi-platform and multi-temporal) image matching problem.

Traditional feature-based image matching methods in computer vision, such as the Scale-Invariant Feature Transform (SIFT) introduced by Lowe (2004), compare images based on the computed similarity between extracted features.

Deep learning similarity estimation frameworks have emerged from these traditional methods. These approaches focus on feature-level similarity estimation, leveraging neural networks to extract high-level representations (features) from inputs and project them into a latent space. In this space, semantically similar inputs are mapped closely, while non-matching inputs are mapped farther apart. Similarity is then quantified by comparing the embedded features using a distance metric such as cosine similarity (Chicco, 2020).

Matching features in heterogeneous RS imagery requires bridging the information gap introduced by platform differences. These differences can be treated as differences between data modalities (e.g., images and text). This outlook allows for drawing inspiration from multimodal matching approaches.

An example of such a matching approach is the Contrastive Language-Image Pre-training (CLIP) framework from Radford et al. (2021), which is designed to as a pretext task to match image-caption pairs for self-supervised contrastive learning. The CLIP framework addresses a multi-modal matching task by employing two individual encoders to extract features from images and text individually. These features are then projected into a common, modality-invariant embedding space and compared using cosine similarity, quantifying the angular distance θ between vectors independently of their magnitude (Eq. 2).

$$\text{Cosine Similarity} = \cos\theta = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \\ = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the projected embedding vectors of two inputs, $\|\boldsymbol{A}\|$ and $\|\boldsymbol{B}\|$ are their corresponding Euclidean norms, n is their dimension.

Implemented Architecture

This framework inspired the architecture of the proposed Double Acquisition NN. However, instead of individual image and text encoders, this network employs two distinct image encoders—one for PS tiles and one for S2 tiles, referred to as the PS and S2 branches, respectively. Each branch independently processes its respective input tile to extract features.

In image-based tasks, ResNet architectures of varying depths (He et al., 2016) are a common choice for image encoding due to their strong feature extraction capabilities. In the context of image matching, they have been applied in domains such as forensic science (Du et al., 2017; Tang et al., 2019), robotics (Qiu et al., 2018), and pedestrian re-identification (Zheng et al., 2017). The original CLIP framework also employs a ResNet architecture for the image encoder, which is modified with antialiased rect-2 blur and attention pooling layers (Radford et al., 2021). To maintain architectural simplicity and the possibility to leverage publicly available pre-trained weights (discussed later in Section 5.3.4), this study opts for a standard ResNet-18 architecture for both the PS and S2 branches.

As ResNet is originally designed for RGB images, architectural modifications are necessary to accommodate the multispectral nature of RS data. Specifically, the first convolutional layer is modified to intake 4-channel inputs in the PS branch and 13-channel inputs in the S2 branch. Furthermore, the final fully connected layer is replaced with an identity mapping in both branches. The resulting 512-dimensional feature vectors are then linearly projected to 128 dimensions. These projection layers are specific to each branch and do not share weights. Figure 2 illustrates an overview of the architecture.

3.2. Model Optimisation

The proposed Double Acquisition NN is optimised using the contrastive learning strategy from the inspirational CLIP framework, which originally was adapted from the InfoNCE loss proposed by Oord et al. (2018). While the CLIP framework is originally designed to jointly pre-train image and text encoders for direct use in zero-shot classification inference, in this study, its optimisation strategy is adopted as the primary **supervised contrastive similarity learning** objective, rather than as a pre-training step.

This supervised contrastive objective leverages a batch of N positive pairs to construct all possible mismatched (negative) pairs within the batch (see Figure E.1 for an illustration of this pair set composition). The model is encouraged to correctly identify positive pairs among all possible pairing combinations. This is achieved by implementing the CLIP loss (Eq. 3), a symmetric cross-entropy

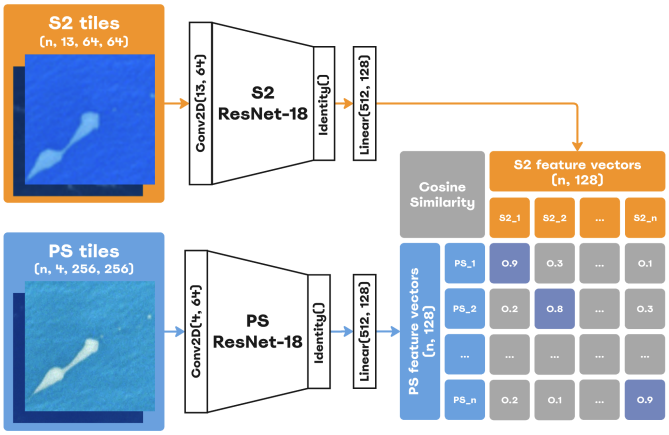

Figure 2: Double Acquisition Neural Network (NN) architecture inspired by Contrastive Language-Image Pre-training (CLIP) framework introduced by Radford et al. (2021). The two-branch model utilises ResNet-18 backbones with modified first and last layers to extract features from PS and S2 tiles containing MD patches. Feature embeddings are projected to 128 dimensions. The similarity matrix (right) is necessary for CLIP loss calculation and contains cosine similarities between embedded S2 and PS tile features. True matches are highlighted in purple; n is a batch size; Conv2D - a convolutional layer with input and output channel dimension parameters.

loss applied over cosine similarity scores between extracted and then projected PS-S2 feature pairs. The total loss averages the losses obtained from matching PS-to-S2 tiles (Eq. 4) and matching S2-to-PS tiles (Eq. 5), thereby ensuring balanced learning for both encoders. During training, similarity values are scaled by a temperature parameter, which is jointly optimised with the model.

$$\mathcal{L}_{\mathbf{CLIP}} = \frac{1}{2N}\left[\sum_{i=1}^{N}\mathcal{L}_{\mathrm{PS}}(\mathbf{PS}_i, \mathbf{S2}_{i,\ldots,N}) + \sum_{i=1}^{N}\mathcal{L}_{\mathrm{S2}}(\mathbf{S2}_i, \mathbf{PS}_{i,\ldots,N})\right] \tag{3}$$

$$\mathcal{L}_{\mathrm{PS}}(\mathbf{PS}_i, \mathbf{S2}_{i,\ldots,N}) = -\log \frac{\exp\left(\frac{\cos(\mathbf{PS}_i, \mathbf{S2}_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\cos(\mathbf{PS}_i, \mathbf{S2}_j)}{\tau}\right)} \tag{4}$$

$$\mathcal{L}_{\mathrm{S2}}(\mathbf{S2}_i, \mathbf{PS}_{i,\ldots,N}) = -\log \frac{\exp\left(\frac{\cos(\mathbf{S2}_i, \mathbf{PS}_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\cos(\mathbf{S2}_i, \mathbf{PS}_j)}{\tau}\right)} \tag{5}$$

where **PS** and **S2** are the projected feature vectors of PS and S2 tiles, respectively. $\cos(\cdot,\cdot)$ denotes cosine similarity, calculated for index i (positive pairs) and index j—all candidate pairs in the batch, including negatives. τ is a learnable temperature parameter.

3.3. Model Training

The model was trained for 800 epochs, with a batch size of 128. The temperature parameter was initialised at $\tau = 0.07$. An initial learning rate of 0.0005 was used with a cosine learning rate scheduler and the Adam optimiser. No data augmentations were applied to the input tensors. The branch-specific ResNet-18 encoders were initialised with general pre-trained weights from Wang et al. (2023), for more details refer to the (b) weight initialisation outlined in Section 5.3.4. Final model weights were selected based on the highest top-5 validation accuracy (described in Section 5.1), which was monitored for each epoch.

4. Related Methods

This section describes other similarity estimation frameworks tested for the MD patch matching task, in addition to the proposed Double Acquisition NN outlined in Section 3.

4.1. Siamese Neural Network

Using features extracted by neural networks for similarity estimation was introduced by Bromley et al. (1993), who proposed the Siamese Neural Network (SiamNN) for signature forgery detection. The objective of this network is to extract features using a time-delay neural network encoder and compare pairs of these extracted features by measuring their cosine similarity. The architecture employs a single shared-weight backbone, forming a two-stream network. SiamNN is optimised to maximise similarity for matching pairs and minimise it for non-matching pairs. Notably, the original SiamNN is a verification model, designed to solve a one-shot classification problem of verifying a given signature against a reference.

Since its introduction, the SiamNN architecture has been applied in numerous fields. Specialised variants are used in handwriting (Du et al., 2017), finger vein (Tang et al., 2019), fingerprint, and face-verification tasks (Chicco, 2020). Its application in remote sensing is also established; for example, He et al. (2018) used SiamNN to match satellite imagery with complex background variations such as land use changes and differing atmospheric conditions.

Implemented Architectures

The **SiamNN** implementation for MD matching involved architectural modifications. Due to differences in spectral and spatial resolutions and varying tile sizes between PS and S2 inputs, the use of a shared-weight backbone architecture could not be directly implemented. To address the mismatch in input properties, three architectural configurations were tested.

First, the **SiamNN-Single** configuration included simple branch-specific pre-encoders for PS and S2 inputs, followed by a shared-weight ResNet-18 backbone (Figure 3a). Second, the **SiamNN-3-layer** configuration replaced the simple pre-encoders with deeper 3-layer Convolutional Neural Networks (CNNs), while keeping the

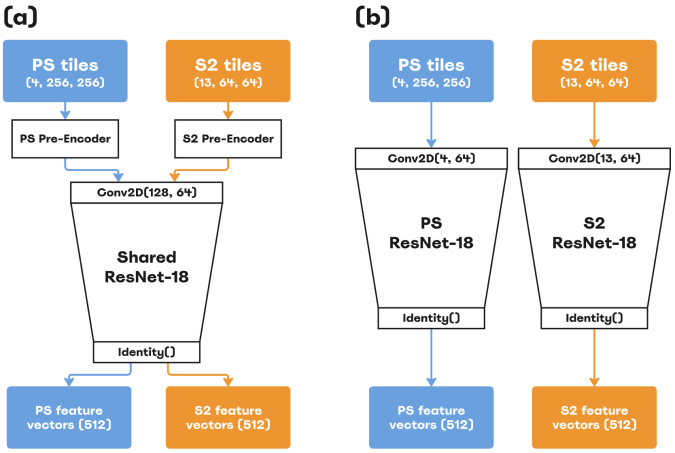

Figure 3: Siamese Neural Network architectures for the PS-S2 MD patch matching problem. (a) Architecture with branch-specific pre-encoders for PS and S2 input tiles to standardise input dimensionality and enable a shared-weight ResNet-18 backbone encoder; (b) Architecture with fully independent branch-specific ResNet-18 encoders. In both cases, the model outputs 512-dimensional feature vectors.

shared-weight backbone unchanged (Figure 3a). Lastly, the **SiamNN-Individual** configuration was composed of entirely independent encoders (Figure 3b), identical to those used in the proposed Double Acquisition NN (Section 3.1). None of the configurations included projection layers, as this component was not part of the original SiamNN design. Detailed architectural descriptions are provided in Appendix F.

Optimisation

All SiamNN models were optimised using the Cosine Embedding loss (Eq. 6). For effective training, the loss requires negative pairs as input. Therefore, to produce a balanced dataset, half of the positive pairs in each input batch were rearranged into negative pairs, resulting in a balanced training set with 50% positive and 50% negative pairs.

$$\mathcal{L}_{\text{Cosine Embedding}}(\boldsymbol{PS}, \boldsymbol{S2}) = \\ = \begin{cases} 1 - \cos(\boldsymbol{PS}, \boldsymbol{S2}), & \text{if } y = 1 \\ \max(0, \cos(\boldsymbol{PS}, \boldsymbol{S2})), & \text{if } y = -1 \end{cases} \quad (6)$$

where $\cos(\boldsymbol{PS}, \boldsymbol{S2})$ denotes the cosine similarity between PS and S2 embedded vectors, $y = 1$ is the pair label for positive pairs and $y = -1$ for negative pairs.

4.2. Deep Relational Similarity Learning

The proposed method in this study is based on the CLIP framework, which assumes that features extracted from different modalities (e.g., image-text or PS-S2) can be aligned in a shared latent space. This modality-invariant space then enables similarity quantification using standard distance metrics. However, Wang et al. (2021) argued that relying on a modality-invariant latent space can be impractical, as data from different modalities may contain different structures and information amounts. As an alternative, researchers introduced the Deep Relational Similarity Learning (DRSL) framework, which learns the optimal similarity function between modalities rather than relying on extracted feature alignment in a common embedding space. Fusing extracted multi-modal features and passing them through a learnable Relation Network allows estimation of the pairwise similarity.

Implemented Architecture

To adapt the original **DRSL** for the MD patch matching task, the DRSL architecture was configured with two individual ResNet-18 encoders for the PS and S2 branches. The first convolution layers were modified to intake modality-specific input channel dimensions, as described in Section 3.1. The final fully connected layers in both branches were replaced by branch-specific fully connected neural networks (FNNs), each with two hidden layers. The feature vectors produced by these FNNs were concatenated and passed through a relational similarity NN consisting of two hidden layers to output a relational pairwise similarity score. Appendix G contains an illustration and details of the implemented DRSL architecture.

Optimisation

Following the original DRSL framework, model training utilised a batch to generate all possible pairwise combinations of PS-S2, forming a full similarity matrix per batch (see illustration in Figure G.1). The network was optimised using a Mean Squared Error (MSE) loss to learn relational similarities that approximate a prior matrix, consisting of ones for positive and zeros for negative pairs.

4.3. Supervised SimCLR

Pair matching is a frequent pretext task in contrastive self-supervised learning (SSL), where models are pre-trained to associate representations of semantically similar inputs and learn robust feature extractions without the need for labelled data (Wang et al., 2022).

SimCLR (Simple Framework for Contrastive Learning of Visual Representations; Chen et al. (2020)) is a contrastive learning framework that employs a Siamese-like architecture to identify positive pairs of two augmented views of the same input images among all other possible pairings in the batch, including negative pairs. This architecture consists of a shared-weight backbone and a non-linear projection head. It is optimised using the Normalised Temperature-scaled Cross Entropy (NT-Xent) loss, a form of the InfoNCE loss (Oord et al., 2018).

A key difference between SimCLR and CLIP (which inspired the proposed Double Acquisition NN) lies in the encoder design. SimCLR uses a single modality-specific encoder to process augmented views of the same image.

In contrast, CLIP employs two separate modality-specific encoders to align image-text pairs.

This study implemented a **supervised** adaptation of SimCLR with the primary objective of matching multi-platform PS and S2 tiles with MD patches, rather than pre-training an image encoder. In this adaptation, the two views correspond to two tiles with the same MD patch captured by different platforms (PS and S2), replacing SimCLR's self-supervised augmentation-based positive pairs with ground truth multi-platform MD patch matches.

Implemented Architectures

To enable SimCLR shared-weight encoder design for the multi-platform tile matching task, the architecture was modified to accommodate platform-specific inputs. The tested architectures mirrored those used in the SiamNN implementation. The first **SimCLR-3-layer** configuration used 3-layer CNN branch-specific pre-encoders, followed by a shared-weight ResNet-18 backbone (Figure 3a). The second **SimCLR-Individual** configuration employed two branch-specific ResNet-18 encoders (Figure 3a). In both variants, encoders were followed by a shared-weight projection head with one hidden layer, outputting 128-dimensional embeddings, consistent with Chen et al. (2020).

Optimisation Variations

When directly applying the SimCLR optimisation strategy to the MD patch matching task, the definition of negative pairs expands to include not only multi-platform negatives (i.e., $(\boldsymbol{PS}_i, \boldsymbol{S2}_j)$ for $j \neq i$), but also same-platform pairs such as $(\boldsymbol{PS}_i, \boldsymbol{PS}_j)$ and $(\boldsymbol{S2}_i, \boldsymbol{S2}_j)$ for $j \neq i$. To reflect this broader definition, the corresponding NT-Xent loss is referred to as **NT-Xent-Full** (Eq. 8).

$$\mathbf{MD} = \{\mathbf{PS}_i\}_{i=1}^{N} \cup \{\mathbf{S2}_i\}_{i=1}^{N} \tag{7}$$

$$\mathcal{L}_{NT-Xent-Full} = -\log \frac{\exp\left(\frac{cos(\mathbf{MD}_i, \mathbf{MD}_p)}{\tau}\right)}{\exp\left(\frac{cos(\mathbf{MD}_i, \mathbf{MD}_p)}{\tau}\right) + \sum_{n \in \mathcal{N}} \exp\left(\frac{cos(\mathbf{MD}_i, \mathbf{MD}_n)}{\tau}\right)} \tag{8}$$

where N is batch size, τ is the temperature parameter, $\cos(\mathbf{MD}_i, \mathbf{MD}_p)$ denotes the cosine similarity between the embedded anchor tile $\mathbf{MD}_i$ and its positive match $\mathbf{MD}_p$ from the other platform (a PS-S2 pair), $\cos(\mathbf{MD}_i, \mathbf{MD}_n)$ denotes similarity between the same embedded anchor tile and each negative sample, including those originating from the same platform. The negative set is defined as $\mathcal{N} = \mathbf{PS}_1, \mathbf{S2}_1, \ldots, \mathbf{PS}_n, \mathbf{S2}_n$, where $n \notin \mathcal{P}$.

However, since this study focuses on matching PS to S2 MD patches (not PS-PS or S2-S2), modified optimisation strategies were introduced to discard same-platform negatives. In addition, these adaptations also accounted for multiple positive matches, as some MD patches were captured in multiple PS scenes. This study tested two loss variants: **NT-Xent-In** (Eq. 9), suitable for datasets with distinct positives-negatives, and **NT-Xent-Out** loss (Eq. 10), which works better with noisy and inaccurate data (Hoffmann et al., 2022).

$$\mathcal{L}_{NT-Xent-In} = -\sum_{p \in \mathcal{P}} \log \frac{\exp\left(\frac{cos(\mathbf{PS}_i, \mathbf{S2}_p)}{\tau}\right)}{\exp\left(\frac{cos(\mathbf{PS}_i, \mathbf{S2}_p)}{\tau}\right) + \sum_{n \in \mathcal{N}} \exp\left(\frac{cos(\mathbf{PS}_i, \mathbf{S2}_n)}{\tau}\right)} \tag{9}$$

$$\mathcal{L}_{NT-Xent-Out} = -\log \frac{\sum_{p \in \mathcal{P}} \exp\left(\frac{cos(\mathbf{PS}_i, \mathbf{S2}_p)}{\tau}\right)}{\sum_{p \in \mathcal{P}} \exp\left(\frac{cos(\mathbf{PS}_i, \mathbf{S2}_p)}{\tau}\right) + \sum_{n \in \mathcal{N}} \exp\left(\frac{cos(\mathbf{PS}_i, \mathbf{S2}_n)}{\tau}\right)} \tag{10}$$

In both equations, $\cos(\mathbf{PS}_i, \mathbf{S2}_p)$ denotes the similarity between embeddings of the anchor PS tile $\mathbf{PS}_i$ and its positive S2 match tile $\mathbf{S2}_p$, $\cos(\mathbf{PS}_i, \mathbf{S2}_n)$ refers to similarity between embeddings of the same anchor tile and all negative samples originating from the S2. The negative set is defined as $\mathcal{N} = \mathbf{S2}_1, \ldots, \mathbf{S2}_n$, where $n \notin \mathcal{P}$.

In all NT-Xent loss variants, cosine similarity is scaled by a temperature parameter τ, which is fixed during training. Preliminary experiments showed that a commonly used temperature value of 0.07 performed better than the original value of 0.5 used by Chen et al. (2020), for the MD patch matching task.

5. Experimental Design

This section outlines the experimental designs for evaluating different similarity estimation frameworks and implementation approaches for matching PS-S2 tiles with MD patches. Each set of experiments corresponds to one of the three RQs defined in the introduction and supports the development of the proposed Double Acquisition NN, which demonstrated the best overall performance.

The proposed model and all experiments were implemented using the PyTorch and PyTorch Lightning frameworks. Each experimental model was trained under comparable conditions, with identical dataset splits and initialisation seeds. Unless specified otherwise, training parameters (where applicable), weight initialisation, and final weight selection strategies were identical to those used for training the proposed Double Acquisition NN (see Section 3.3).

5.1. Evaluation

To assess and compare experimental model designs, this study used two groups of metrics: classification and retrieval. While both metric groups provide insights into model behaviour details, retrieval metrics were the primary focus given the retrieval formulation of the MD patch matching task.

5.1.1. Classification Evaluation

Classification metrics were used to evaluate the model's ability to correctly assign positive or negative pair labels based on a similarity threshold of 0.5.

Dedicated **balanced** validation and test datasets were constructed to evaluate classification performance, each containing 50% positive and 50% negative pairs. Positive pairs were sampled directly from the full evaluation dataset. In contrast, negative pairs were generated by reassigning a non-matching S2 tile to each of the remaining PS anchor tiles (see dataset composition in Figure E.2).

This balanced dataset was also used to compute **classification accuracy** and average similarity scores: the **mean average similarity** (calculated over the entire balanced dataset), and the **average positive and average negative similarities** (calculated over positive and negative pairs separately in this balanced dataset). These metrics were applied selectively to provide a complementary measure of the model's capacity to differentiate between positive and negative pair inputs in key experiments.

5.1.2. Retrieval Evaluation

The retrieval metrics are generally used to quantify a model's ability to correctly rank the relevant candidate for a given query among a set of candidates based on their ranking scores. This study used two retrieval metrics: mean position and top-k accuracy.

The **mean position** describes the average position of the correct match within the ranked list of candidates based on similarity scores (Eq. 11).

$$\text{Mean Position} = \frac{1}{N}\sum_{i=1}^{N}\text{position}(\mathbf{S2}_i^{\text{p}}) \tag{11}$$

where N is the number of query PS tiles, and $\text{position}(\mathbf{S2}_i^{p})$ is the rank of the true matching S2 tile (with the same physical MD patch) in a set of S2 tile candidates for the i-th PS query.

Top-k accuracy measures the proportion of queries for which the true match is positioned within the top k most similar predictions (Eq. 12). This study monitored $k \in \{1, 3, 5, 10, 50\}$.

$$\text{Top-}k\text{ Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & \text{if } \mathbf{S2}_i^{p} \in (\mathbf{S2}_i)_j \\ 0 & \text{if } \mathbf{S2}_i^{p} \notin (\mathbf{S2}_i)_j\end{cases} \tag{12}$$

where N is the number of query PS tiles, $\mathbf{S2}_i$ denotes the ranked set of candidates for the i-th PS query, p is the relevant candidate (i.e., true match in S2) and $(\mathbf{S2}_i)_j$ where $j \in \{1, ..., k\}$ denotes the top-k candidate subset.

Since the MD patch matching task is formulated as a retrieval problem, the optimal Double Acquisition NN configuration in each experiment was selected based on the top-1 accuracy on the test dataset, using the full S2 set as the candidates, reflecting the global retrieval regime, which is detailed later in Section 5.4.

5.2. Framework Variation Experiments (RQ1)

This study investigated the effectiveness of various established similarity estimation frameworks for the MD patch matching. The following frameworks were tested: the proposed Double Acquisition NN, SiamNN-Single, SiamNN-3-layer, SiamNN-Individual, DRSL, SimCLR-3-layer and SimCLR-Individual. All SimCLR-based architectures were optimised using the NT-Xent-Out loss, while the other loss variants (NT-Xent-In and NT-Xent-Full) were tested on the SimCLR-Individual architecture.

All models were trained for 200 allocated epochs using a batch size of 128 positive pairs. Subsequent experiments involved the best-performing configuration identified from this comparison.

5.3. Hyperparameter Optimisation (RQ2)

This section describes hyperparameter experiments conducted on the proposed Double Acquisition NN (as described in Section 3).

5.3.1. Batch Size Experiments

This study investigated the effects of increasing batch size on model performance. Experiments were conducted using batch sizes of 16, 32, 64, 128, and 256. Initially, all configurations were trained for 200 allocated epochs. However, due to the unexpected behaviour of smaller batch sizes performing slightly better, additional experiments were carried out to extend the training duration of larger batches such that each configuration performed approximately the same number of model weight updates (i.e., steps). The training for a model with a batch of 32 remained for 200 allocated epochs, for a batch size of 64, it was increased to 400 allocated epochs, 128—800 allocated epochs, and 256—1600 allocated epochs.

The optimal batch size and training duration were selected from extended-duration training experiments.

5.3.2. Augmentation Strategy Experiments

Data augmentations were tested to enhance the model's generalisation abilities and virtually increase the size of the training set. Augmentations were independently applied to PS and S2 training tiles, allowing each tile in a pair to undergo a different (randomised) augmentation strategy.

The benefits of augmentations on model performance were evaluated across multiple augmentation strategies: no augmentations, mild, medium, and harsh augmentations, and the spectral-channel-shuffling strategy. Detailed descriptions of these augmentation strategies are provided in Appendix H, and Figure H.1 shows examples of each strategy. All configurations were trained using a batch size of 128 for 800 allocated epochs.

5.3.3. Base Encoder Depth Experiments

ResNet-18 was selected as the primary base architecture for image encoding in all experiments due to its relatively small size and simplicity, making it more suitable for extensive experimentation. Once the proposed architecture configuration was identified, a deeper and more complex ResNet-50 base architecture was tested within the Double Acquisition NN framework. Both branch-specific ResNet-50 encoders incorporated the same modifications applied to the ResNet-18 encoder: adjustments to the first convolutional layer and replacement of the last fully connected layers (see Section 3.1).

These experiments were trained using a batch size of 128 for 800 allocated epochs and employed slightly different weight initialisation strategies. MoCo weights were the only ones available for ResNet-18 in the SSL4EO project (Self-Supervised Learning for Earth Observation; Wang et al. (2023)), thus they were used by default. However, SSL4EO provides additional weight options for the ResNet-50 architecture; therefore, to further investigate the effects of base encoder architecture on model performance and to account for possible biases introduced by different weight initialisations, the ResNet-50 encoders were initialised separately with MoCo and DINO weights from the SSL4EO study.

5.3.4. Pre-Training Variation Experiments

This study tested three initialisation configurations for the proposed Double Acquisition NN to investigate the benefits of using pre-trained encoders. These configurations were subsequently fine-tuned (the entire network) on the MD patch matching task for 200 allocated epochs, with the remaining training parameters kept consistent as described in Section 3.3. The models were initialised with:

(a) random weights;

(b) general pre-trained weights;

(c) weights obtained through self-supervised SimCLR refinement of the general pre-trained weights.

(a) Random Initialisation Weights

In this initialisation configuration, the entire network (except the temperature parameter) had randomly initialised weights, and it served as a baseline to assess the impact of pre-trained weights.

(b) General Pre-trained Weights

This initialisation configuration was used in all experiments described above. Here, the ResNet encoders in each model branch (or the shared encoder, where applicable) were initialised with publicly available weights from the SSL4EO project (Wang et al., 2023). These weights were obtained through pre-training conducted by the SSL4EO team on S2 Level-1 data from all 13 spectral channels using the MoCo SSL framework. Pre-training data were sampled around 10,000 of the most populated cities, making these weights *general* for broad use of S2 applications in the Earth observation domain.

For encoders requiring a different number of input channels (e.g., PS encoder, or shared-weight encoders in the SiamNN and supervised SimCLR frameworks), the first convolutional layer, along with other introduced or modified layers (e.g., pre-encoders, projection layers, FNNs, Relational Similarity NN), was randomly initialised.

To the best of the author's knowledge, there are no publicly available, generally pre-trained ResNet architectures for PS data. Therefore, an ablation experiment was conducted on the Double Acquisition NN to investigate the applicability of S2-based general pre-trained weights for the PS modality. This experiment used MoCo pre-trained weights for the S2-branch encoder, while the PS-branch and the remaining architecture parts were randomly initialised.

(c) Self-Supervised Refined Weights

Initialisation configuration (c) was built upon configuration (b) by refining the SSL4EO weights through additional self-supervised pre-training using the SimCLR framework, applied independently to each branch-specific encoder.

For the PS branch, the encoder architecture was configured as in the proposed Double Acquisition NN (see Section 3.1), except that the final projection layer was replaced with a non-linear projection head producing 128-dimensional feature vectors, as proposed by Chen et al. (2020). Pre-training was performed using the NT-Xent-Full loss (Eq. 8) on batches of positive pairs, composed of two augmented views derived from the same platform, specifically, PS tiles, resulting in PS-PS pairs. Augmented views were generated using the harsh augmentation strategy (see Appendix H for more details). This ResNet-18 base encoder was initialised with the SSL4EO weights.

Self-supervised SimCLR pre-training was configured with the following hyperparameters: temperature $\tau = 0.07$, batch size of 512, and 200 training epochs. The same pre-training procedure was replicated independently for the S2 branch encoder, using S2 tiles.

The final encoder weights used in downstream fine-tuning for both branch-specific encoders were selected based on the top-3 validation accuracies monitored during self-supervised SimCLR pre-training. The non-linear projection heads used during pre-training were replaced with randomly initialised linear projection layers for the transfer to the MD patch matching task.

As the model initialised with refined weights performed worse than those using the original SSL4EO weights, additional experiments were conducted to investigate whether this performance drop can be associated with catastrophic forgetting. This phenomenon, described by McCloskey and Cohen (1989), occurs when a neural network forgets previously learned useful representations as its weights adjust to a new training objective. To investigate whether

self-supervised SimCLR pre-training induced such forgetting, the Double Acquisition NN was fine-tuned on the MD patch matching task after 1, 3, 5, 10, 15, 20, and 25 epochs of self-supervised SimCLR pre-training for each branch encoder.

5.4. Candidate Selection Regimes (RQ3)

This set of experiments was no longer focused on architectural and training-related experiments. Instead, it examined how to implement the proposed Double Acquisition NN most efficiently by leveraging prior knowledge to enhance performance further.

Since the MD patch matching problem is formulated as a retrieval task, applying prior-knowledge-based constraints to limit the candidate set is expected to reduce task complexity and improve retrieval performance. This study explored three candidate selection regimes: global, local and drift-bound, each applying progressively stronger constraints to narrow the candidate search space. In these experiments, only retrieval metrics were evaluated.

Global Regime

In the global regime, each PS MD patch tile in the evaluation dataset was compared against all available S2 tiles with MD patches from the S2 platform (in the same dataset). MD patches originating from one MD event were compared against those from all events in the dataset, simulating a completely unconstrained search scenario.

For the global regime, the global tok-k accuracy was computed individually for each query and then averaged across all queries.

Local Regime

The local regime restricts the candidate set to only those tiles with MD patches from the same MD event as the query MD patch. This regime simulated a scenario where a single MD event is investigated at a time, without any prior knowledge of the expected maximum drift speed.

For the local regime, the local top-k accuracies were first averaged per test MD event, and then across all events.

Drift-bound Regime

The drift-bound regime is based on the intuition that MD patches can only drift within a limited spatial radius. Due to uncertainties in drift model accuracy for estimating precise expected MD patch locations, this study adopts a simplified approach of drift-informed restrictions. Specifically, it applies a constant maximum drift speed per MD event, thereby avoiding the need for complex drift modelling.

The expected search area for candidate selection is determined by combining the maximum drift speed with the time elapsed between acquisitions. Since MD annotations include temporal information, an expected drift radius can be estimated for each MD patch query and used to filter the candidate set accordingly.

Maximum drift speed values were determined through an optimal drift speed search conducted for each test MD event. This study tested drift speed values ranging from 2 cm/s to 2 m/s. Initial values were drawn from the literature: Gerin et al. (2013) reported an average drift speed of 2 cm/s ($\pm$20 cm/s standard deviation) for drifters in the Marmara Sea, and Kikaki et al. (2020) reported 2-14 cm/s for MD off the coast of Honduras. However, these speed values proved insufficient to retrieve any candidate matches.

It is important to note that this drift speed search was not conducted to tune the search radius to the test data and introduce bias, but rather to simulate an applied scenario where the expected maximum drift speed is known, for example, from oceanographic forecasts. A full applied case study of drift-informed candidate selection is beyond the scope of this study.

In the drift-bound regime, top-k accuracies were averaged per MD event and across all events. Candidate set sizes varied for each query and were sometimes smaller than the k retrieval scope. The corresponding query was excluded from the mean position calculation if the true match was outside the candidate set. For top-k accuracy, queries were assigned a value of 1 if the true match was within the candidate set, and 0 otherwise.

6. Results

6.1. Performance of Framework Variations (RQ1)

The test classification and retrieval results for all evaluated frameworks are summarised in Table 2. The proposed Double Acquisition NN achieved the highest retrieval performance across all metrics, including a global top-1 accuracy of 26.7%, a mean position of 20.8. It also reached the second-best classification accuracy of 87.1%, confirming the model's robustness for classification and retrieval tasks.

In contrast, DRSL and all SiamNN variants demonstrated substantially lower retrieval performance, with global top-1 accuracies below 3% and mean positions exceeding 100. DRSL and SiamNN-Single also achieved classification accuracies (of 50%) close to random chance. Among the SiamNN variants, the SiamNN-Individual architecture performed best in the retrieval task, although it remained far behind the proposed model (global top-1 accuracy of 0.6 vs 26.7%). Despite their poor retrieval performance, SiamNN-3-layer and SiamNN-Individual reached some of the highest classification accuracies of 88.3 and 85.2%, respectively.

Supervised SimCLR frameworks achieved the second-best retrieval performance overall. The top-performing (based on global top-1 accuracy) SimCLR variant stayed behind the proposed model by 3.7%. Within the SimCLR variants, the architecture with individual encoders consistently outperformed the 3-layer shared architecture, as

Table 2: Test classification and retrieval performance of all evaluated framework variants. Classification metrics are evaluated on the balanced dataset, and retrieval metrics under the global candidate selection regime (with 661 MD patches). All frameworks were trained for 200 allocated epochs (ep). Bold values indicate the best-performing scores per metric. Frameworks are grouped by type, with variants listed together. DRSL-Deep Relational Similarity Learning; SiamNN - Siamese NN; SimCLR-a supervised Simple Framework for Contrastive Learning implementation; NT-Xent-Normalised temperature-scaled cross entropy loss variants

Framework	Classification Accuracy (%)	Global Retrieval					
		Mean Position	Top-k Accuracy (%)				
			1	3	5	10	50
Double Acquisition NN (200ep)	87.1	**20.8**	**26.7**	**44.7**	**56.5**	**66.2**	**89.0**
DRSL	50.0	102.8	2.5	6.5	8.9	15.1	43.2
SiamNN-Single	50.0	266.2	0.3	0.4	0.4	1.2	2.4
SiamNN-3-layer	**88.3**	126.7	0.6	2.5	3.7	7.3	31.3
SiamNN-Individual	85.2	133.1	0.6	2.2	4.3	8.3	26.3
SimCLR-3-layer (NT-Xent-Out)	73.0	77.5	13.4	25.4	31.6	41.5	66.2
SimCLR-Individual (NT-Xent-Out)	86.9	29.4	23.0	42.0	48.7	61.9	84.1
SimCLR-Individual (NT-Xent-In)	83.8	27.4	21.8	38.1	47.3	61.9	86.1
SimCLR-Individual (NT-Xent-Full)	87.1	31.6	20.2	38.4	46.3	59.3	85.0

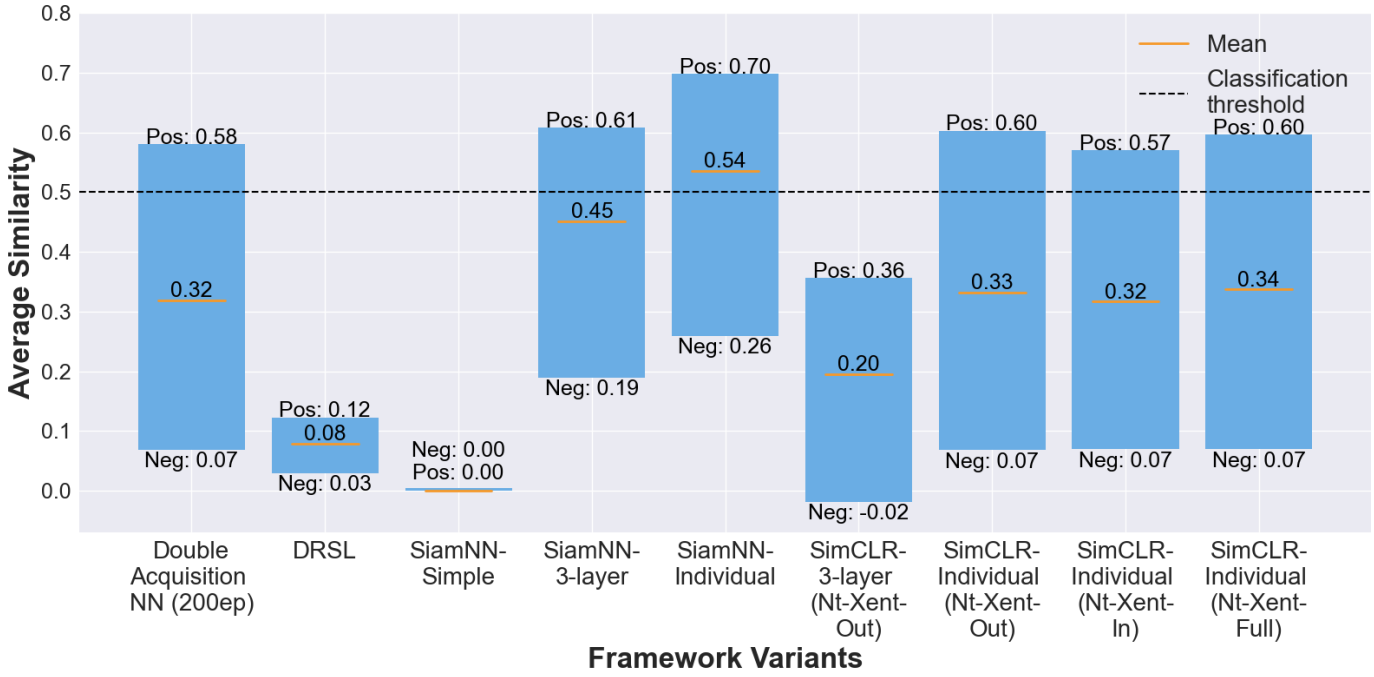

Figure 4: Average mean, positive (pos) and negative (neg) pair similarities for each tested framework variation. All frameworks were trained for 200 allocated epochs (ep). The classification threshold equals 0.5.

reflected by the drop in mean position from 77.5 to 29.4. However, none of the three NT-Xent loss variants used in SimCLR-Individual showed a clear advantage over one another across retrieval metrics.

Figure 4 presents the average mean, positive, and negative similarities across all frameworks. These patterns relate to the observed classification performances. Frameworks such as DRSL, SiamNN-Single, and SimCLR-3-layer, which had the lowest classification accuracies, also exhibited average positive similarities below the 0.5 classification threshold. The SiamNN-3-layer variant had the highest classification accuracy and the mean average similarity just below the classification threshold. In contrast, the highest-performing models in the retrieval task (the proposed network and SimCLR-Individual variants) demonstrated a wider gap between average positive and negative similarities. This effect was largely induced by substantially lower average negative similarity values, particularly compared to the SiamNN variants.

Table 3: Test classification and retrieval performance of the proposed Double Acquisition NN under different training batch sizes and extended training durations. Classification metrics are evaluated on the balanced dataset, and retrieval metrics under the global candidate selection regime (with 661 MD patches). Results are grouped into two training regimes: a baseline regime with 200 allocated epochs, and extended training regimes for larger batch sizes. Bold values indicate the best-performing scores per metric for each training regime.

Batch Size/ epochs	Classification Accuracy (%)	Global Retrieval					
		Mean Position	Top-k Accuracy (%)				
			1	3	5	10	50
16/200	**93.2**	19.0	26.1	46.9	56.5	69.6	**90.5**
32/200	90.8	**18.5**	**31.5**	**52.5**	**60.2**	**71.2**	89.9
64/200	90.5	21.6	28.5	47.9	57.3	68.1	88.9
128/200	87.1	20.8	26.7	44.7	56.5	66.2	89.0
256/200	73.1	38.3	15.3	29.4	37.5	49.9	78.6
64/400	**89.5**	22.1	29.7	49.6	59.6	69.1	89.3
128/800	89.0	**15.5**	**37.5**	**56.5**	**65.7**	**75.7**	**92.3**
256/1600	73.7	22.3	24.3	41.4	49.7	65.0	89.5

6.2. Optimised Hyperparameters (RQ2)

6.2.1. Batch Size

Table 3 summarises the test retrieval and classification performance of the proposed Double Acquisition NN across varying batch sizes and training durations. Results are grouped into two training regimes: a baseline regime with 200 allocated training epochs, and an extended training regime for larger batch sizes.

The highest retrieval performance was observed with a batch size of 128 under the extended training regime. Within the baseline regime, a batch size of 32 had the best overall performance, except for global top-50 retrieval accuracy and classification accuracy, which were slightly higher for the model trained with a batch size of 16.

Figure 5 shows the average mean, positive, and negative similarities for each batch size. Smaller batch sizes (16, 32, and 64) yielded higher classification accuracies (90.5-93.2%) and higher average positive similarity scores. In contrast, models trained with larger batch sizes (128 and 256) under the baseline regime showed reduced classification accuracies (87.1% and 73.1%, respectively) and lower average positive similarity scores (e.g., 0.45 for batch size of 256, just below the classification threshold). When extended training was applied, a batch size of 128 improved classification and retrieval performance compared to the corresponding 64 and 32 batch size models.

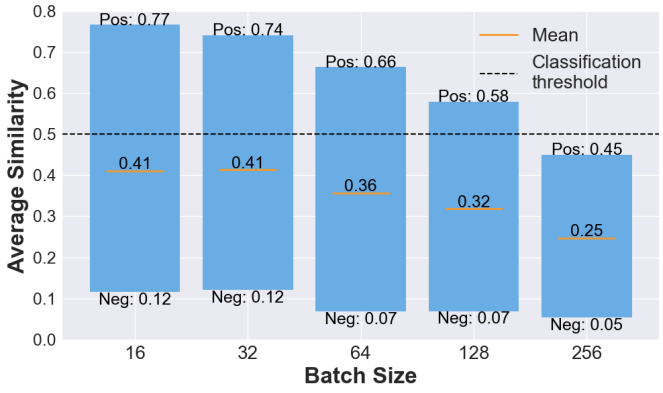

Figure 5: Average mean, positive and negative pair similarities for the proposed Double Acquisition NN trained with varied batch sizes. All models were trained for 200 allocated epochs. The classification threshold equals 0.5.

6.2.2. Augmentations Strategies

Table 4 presents the test retrieval and classification performance of the proposed Double Acquisition NN trained with different augmentation strategies: none, mild, medium, harsh, and spectral-channel-shuffling.

The model trained without augmentations achieved the highest global top-1, top-3, top-5, and top-10 retrieval accuracies. However, the model trained with the medium augmentation strategy slightly outperformed in mean position (15.5 vs 13.9) and global top-50 accuracy (92.3% vs 92.9%), indicating improved ranking performance beyond the top-10. The highest classification accuracy was achieved by the model trained with harsh augmentations. However, the same harsh strategy and the spectral-channel-shuffling strategy resulted in the lowest retrieval performance scores.

6.2.3. Base Encoder Depth

Table 5 reports the test retrieval and classification performance of the proposed Double Acquisition NN with different base encoder architectures. Results grouped by encoder type: ResNet-18 (initialised with MoCo pre-trained weights) and ResNet-50 (evaluated with both MoCo and DINO pre-trained weights).

The model equipped with ResNet-18 as the base encoder for each branch outperformed both ResNet-50 variants across all classification and retrieval metrics. It achieved a global top-1 accuracy of 37.5%, compared to 26.4% for the ResNet-50 with MoCo weights and 15.1%

Table 4: Test classification and retrieval performance of the proposed Double Acquisition NN trained with different data augmentation strategies. Classification metrics are evaluated on the balanced dataset, and retrieval metrics under the global candidate selection regime (with 661 MD patches). All experimental models were trained for 800 allocated epochs. Bold values indicate the best-performing scores per metric.

Augmentation Strategy	Classification Accuracy (%)	Global Retrieval					
		Mean Position	Top-k Accuracy (%)				
			1	3	5	10	50
None	89.0	15.5	**37.5**	**56.5**	**65.7**	**75.7**	92.3
Mild	90.9	15.0	33.1	53.7	62.9	72.7	92.3
Medium	93.0	**13.9**	34.4	55.8	63.4	74.5	**92.9**
Harsh	**94.2**	18.9	25.7	44.8	53.1	65.3	90.2
Spectral-channel-shuffling	85.2	25.5	31.8	47.8	56.4	68.2	88.1

Table 5: Test classification and retrieval performance of the proposed Double Acquisition NN when employing different base architectures for image encoding. In brackets, the indication of which general pre-trained weights from SSL4EO (Wang et al., 2023) were used for base-encoder initialisations. All experimental models were trained for 800 allocated epochs. Classification metrics are evaluated on the balanced dataset, and retrieval metrics under the global candidate selection regime (with 661 MD patches). Bold values indicate the best-performing scores per metric.

Base Architecture	Classification Accuracy (%)	Global Retrieval					
		Mean Position	Top-k Accuracy (%)				
			1	3	5	10	50
ResNet-18-MoCo	**89.0**	**15.5**	**37.5**	**56.5**	**65.7**	**75.7**	**92.3**
ResNet-50-MoC0	86.9	22.4	26.4	43.2	52.7	62.5	88.4
ResNet-50-DINO	86.6	40.4	15.1	29.4	38.4	50.3	80.7

for the ResNet-50 with DINO weights.

Between the two ResNet-50 configurations, the MoCo-initialised model outperformed the DINO-initialised model across all retrieval metrics, with improvements ranging from 7.7 to 18.0%. The difference in classification capacity was marginal (0.3%), with the MoCo variant being better.

6.2.4. Pre-Trained Weight Initialisation

Table 6 summarises the impact of different weight initialisation strategies obtained through different pre-training strategies for the encoders in the proposed Double Acquisition NN on test retrieval and classification performance.

The model initialised with SSL4EO pre-trained weights on both the PS and S2 encoders outperformed the randomly initialised variant across all retrieval metrics. Notably, global top-1 accuracy increased by 11.4%, and other global top-k metrics showed relative improvements ranging from 9.2 to 18.5%. Classification accuracy was slightly higher (87.1% vs 89.6%) for the randomly initialised variant.

When SSL4EO weights were initialised only in the S2 encoder, performance decreased across all metrics compared to random initialisation, except for a marginal improvement in global mean position (by 1.4) and global top-50 accuracy (by 0.8%).

Further refinement of SSl4EO weights through self-supervised SimCRL pre-training decreased performance, with global top-1 and top-3 accuracies below those obtained with random initialisation. Overall, this refined-weight configuration performed worse than the unrefined SSL4EO weight initialisation.

Figure 6 shows the downstream performance of refined-weight models across self-supervised SimCLR pre-training epochs. Retrieval performance began to decline from the first epoch and continued to decline until epoch 15, after which it began to recover slightly. However, the refined weights never regained the original performance level of the unrefined SSL4EO weights.

6.3. Impact of Candidate Selection Restrictions (RQ3)

Table 7 summarises the retrieval performance of the proposed Double Acquisition NN under varying levels of prior knowledge, defined by applying different candidate selection regimes (global, local, and drift-bound) to the testing procedures.

Under the global regime, the model ranked the true match, on average, as the 15.5^{th} most similar tile. It correctly retrieved the top-1 prediction in 37.5% of test cases. With three attempts, the model successfully retrieved the true match in 56.6% of cases (see Figure 7 for some top-3 retrieval examples), and with 50 attempts, in 92.3% of

Table 6: Test classification and retrieval performance of the proposed Double Acquisition NN fine-tuned on different weight initialisations. All experimental models were trained for 200 allocated epochs. Classification metrics are evaluated on the balanced dataset, and retrieval metrics under the global candidate selection regime (with 661 MD patches). Results are organised into three groups: random weight initialisation, SSL4EO pre-trained weights (applied either to both PS and S2 encoders or to the S2 encoder only), and SSL4EO weights further refined using a self-supervised SimCLR pre-training strategy. Bold values indicate the best-performing scores per metric.

Weight Initialisation	**Classification Accuracy (%)**	**Global Retrieval**					
		Mean Position	**Top-k Accuracy (%)**				
			1	**3**	**5**	**10**	**50**
Random	**89.6**	39.0	15.3	31.5	38.0	50.4	79.8
SSL4EO pre-trained weights in:							
PS and S2 branches	87.1	**20.8**	**26.7**	**44.7**	**56.5**	**66.2**	**89.0**
S2 branch	86.8	37.6	11.9	27.0	36.2	49.7	80.6
SSL4EO weights refined with self-supervised SimCLR	88.1	32.8	15.0	30.9	39.0	50.4	82.8

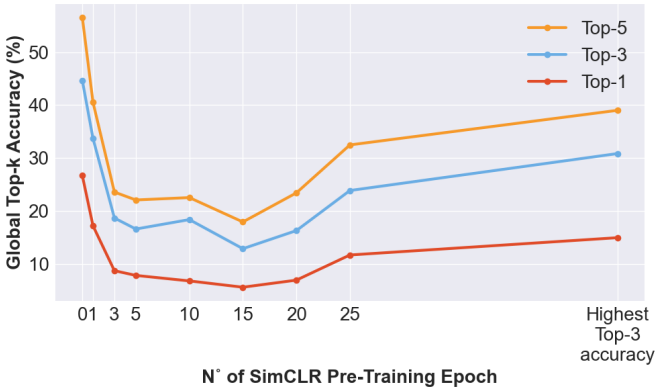

Figure 6: Global top-1, top-3, and top-5 test accuracies of Double Acquisition Neural Networks fine-tuned for the MD patch matching task. Models were initialised either with MoCo weights from the original SSL4EO (epoch 0) or with weights obtained from additional self-supervised SimCLR pre-training for 1, 3, 5, 10, 15, 20, and 25 epochs, as well as from an extra checkpoint selected based on the highest top-3 validation accuracy during self-supervised SimCLR pre-training.

cases, indicating that the model was able to narrow down the correct match to a smaller candidate set.

Under the local regime, the model retrieved the true match more efficiently: on average, within 12.6 attempts. Local top-5, top-10, and top-50 accuracy scores were only marginally higher (by 0.1-1.2%) than in the global regime, whereas the local top-1 accuracy was 0.5% lower than the global top-1 accuracy.

Figure 8 shows retrieval results of the optimal maximum drift speed search for each test MD event. The reported drift speed values in the literature (2-20 cm/s) failed to retrieve any candidate MD patches for most queries, resulting in low retrieval accuracies and simulating a maximum drift speed underestimation scenario. Gradually increasing the drift value improved retrieval performance up to a breakpoint. This breakpoint had the highest performance scores and was used to select the optimal maximum drift speeds. They were: 90 cm/s for the Accra MD event, and 40 cm/s for Marmara. Increasing the drift speed further and simulating the overestimation scenarios resulted in a gradual decline in performance.

Introducing these optimised maximum drift speed val-

Table 7: Test retrieval performance of the proposed Double Acquisition NN under global, local, and drift-bound candidate selection regimes. Results for local and drift-bound regimes are reported separately for the Marmara and Accra MD events. Bold values indicate the best-performing scores per metric.

Search Scope	**Mean Position**	**Top-k Accuracy (%)**				
		1	**3**	**5**	**10**	**50**
Global	15.5	37.5	56.5	65.7	75.7	92.3
Local	12.6	37.0	56.5	65.8	76.4	93.5
Accra	12.8	32.7	54.7	64.4	73.4	92.8
Marmara	12.4	41.2	58.3	67.2	79.3	94.2
Drift-bound	**1.9**	**62.2**	**88.4**	**94.0**	**98.8**	**99.2**
Accra ($0.9m/s$)	2.2	55.0	84.9	91.0	98.2	98.9
Marmara ($0.4m/s$)	1.6	69.4	91.9	97.0	99.5	99.5

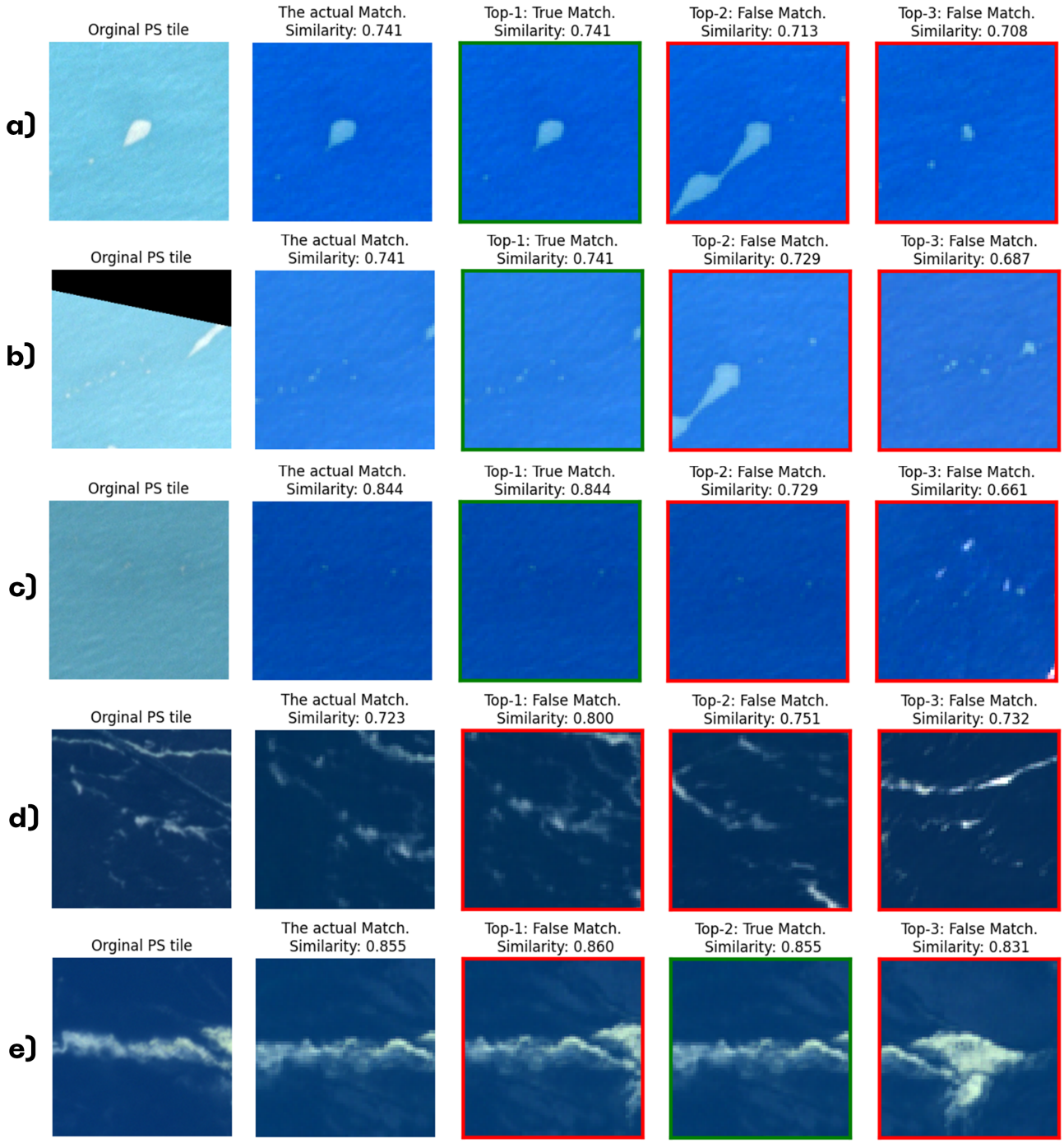

Figure 7: Examples of top-3 match predictions from the global S2 tile test set for given PS query tiles. Note that the false positives share high visual similarity with the actual match tiles. a-c are from the Accra MD events, d and e are from the Marmara MD event.

ues for the drift-bound candidate set restrictions significantly enhanced the model's retrieval performance. The model ranked the true match at an average position of 1.9, compared to 12.6 in the local regime. Top-1 accuracy improved from 37.0% (local regime) to 62.2%, and top-3, top-5, top-10 and top-50 accuracies increased by 31.9%, 28.2%, 22.4% and 5.7%, respectively. These improvements are evident in Figures 9 and 10, which visualise top-1 predictions under the local and drift-bound regimes. In the local regime, many top-1 predictions are pointed to MD patches at distances unfeasible to be reached by passive drift. In contrast, drift-bound predictions were more localised. However, some confusion remained, especially in clustered MD patches, where the top-1 prediction points

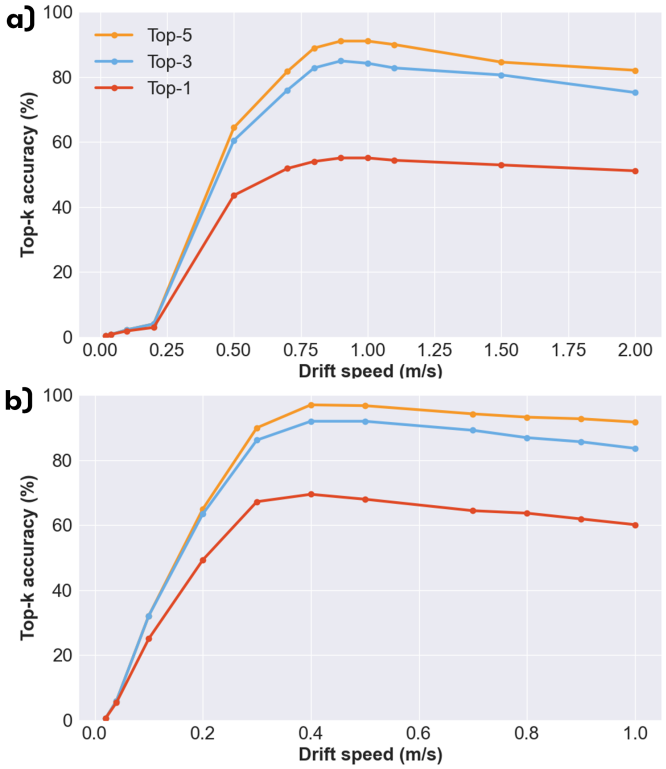

Figure 8: Top-1, top-3, and top-5 retrieval accuracies under varying drift maximum drift speed regimes for: a) the Accra MD event; b) Marmara MD events.

to an MD patch near the correct match, resulting in an incorrect top-1 prediction despite spatial proximity to the correct match.

When examining MD events individually, the Marmara MD event test set yielded higher retrieval scores than the Accra set. Under the local regime, top-1 accuracies were 41.2% (Marmara) and 32.7% (Accra), increasing to 69.4% and 55.0%, respectively, under the drift-bound regime. As shown in Figures 9 and 10, the model successfully identified several matches on the first attempt in both study events. However, prediction errors are still evident, consistent with overall top-1 scores below 70%.

7. Discussion

This study aimed to design an optimal architecture, optimisation and training strategies, as well as implementation procedures for the MD patch matching problem across PS and S2 double acquisition imagery. The proposed Double Acquisition NN (as described in Section 3) emerged through a series of experiments that provided insights into the architectural and implementation requirements and the inherent challenges of the MD patch matching task.

7.1. Framework Performance Analysis (RQ1)

Experimental results indicate that MD patch matching across PS-S2 double acquisitions aligns more closely with a multi-modal matching task. Consequently, performance benefits from using platform-specific encoders that account for the modality gap between platforms. However, this gap can be effectively bridged through robust feature extraction, without the need for an explicit similarity estimation; simple cosine similarity proves sufficient. The MD patch matching performance is enhanced when the model is trained in a retrieval setting that leverages batches for negative sampling, to introduce many negative pairs. However, this presents a significant class imbalance, necessitating careful selection of loss functions to ensure reliable optimisation. The final, proposed Double Acquisition NN adopts a CLIP-inspired framework configuration.

Effects of Shared Backbone Architectures

The superior performance of the proposed Double Acquisition NN with platform-specific encoders and the relative advantage of SiamNN and supervised SimCLR variants using individual encoders over their shared-weight counterparts suggests that MD patch matching resembles a multi-modal similarity estimation task.

Although both inputs are images and not fundamentally different data types, like image-text pairs, they differ in spatial resolution (3 m vs up to 60 m) and spectral information (4 vs 13 spectral channels), making them heterogeneous in terms of information content. These differences likely limit the effectiveness of shared-weight encoder architectures in extracting consistent and useful features across platforms. Moreover, the small pre-encoder modules used in this study (single or 3-layer CNNs) may have been insufficient to extract modality-invariant subfeatures, which could then benefit from a shared-weight backbone.

Although both PS and S2 platforms can achieve comparable results for the same tasks (e.g., mountain pine mapping (Rösch et al., 2022)), the features used to achieve these results may originate from different aspects of the data. Platform-specific encoders can better exploit these intricacies, yielding more robust features that align more effectively in a shared latent space. The findings of this study support multi-modality presence and the choice of individual encoders for MD patch matching.

Training Objective and Retrieval Task Alignment

The poor retrieval performance with relatively high classification accuracies of the SiamNN framework variants could have been anticipated. The original SiamNN framework was designed for a classification rather than a retrieval task (Bromley et al., 1993). In this study, the SiamNN models, particularly the 3-layer and Individual encoder variants, performed the classification task reasonably well. This performance indicates that the models could distinguish positive and negative pairs in a balanced dataset.

However, when deployed in a retrieval setting, these model variants struggled to identify the correct match among many negative candidates. This struggle may

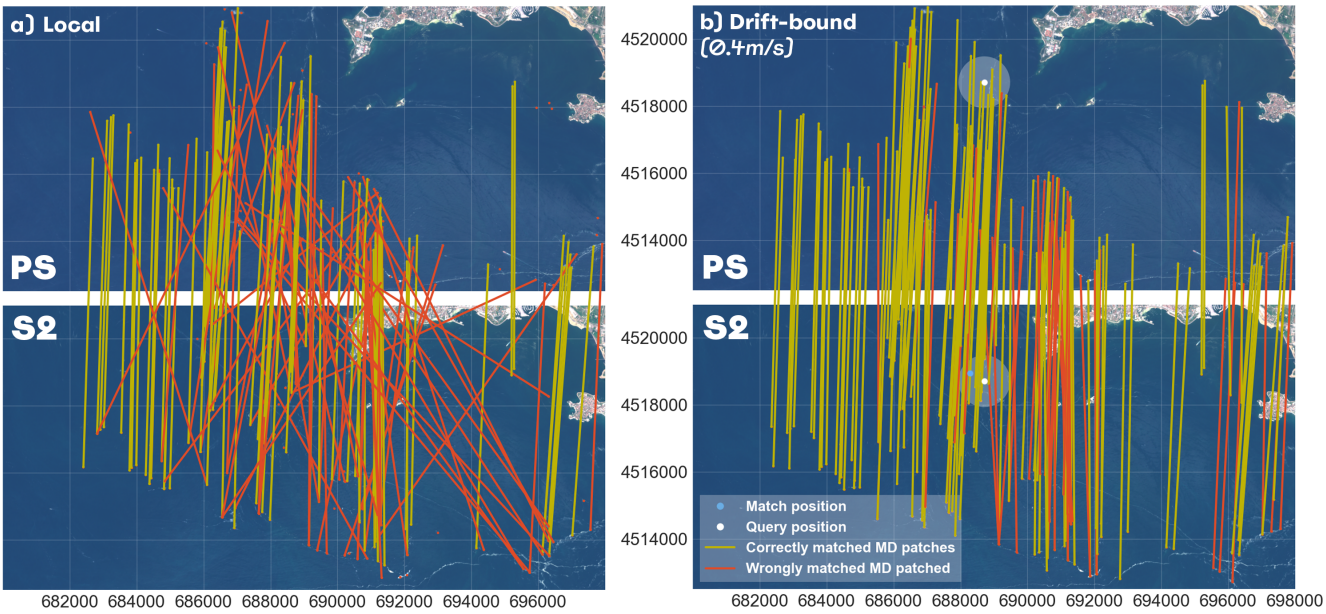

Figure 9: Comparison of top-1 predictions for: a) local Marmara MD event-bound, and b) drift-bound scopes. Lines connect each MD patch in the PS scene with its top-1 match in the corresponding S2 scene. Note that red dots without lines indicate top-1 predictions that were wrongly predicted and did not fit into the extent of the figure. In b), the white circle indicates an example of search scope for a given query (white dot), while the blue dot is the true match position.

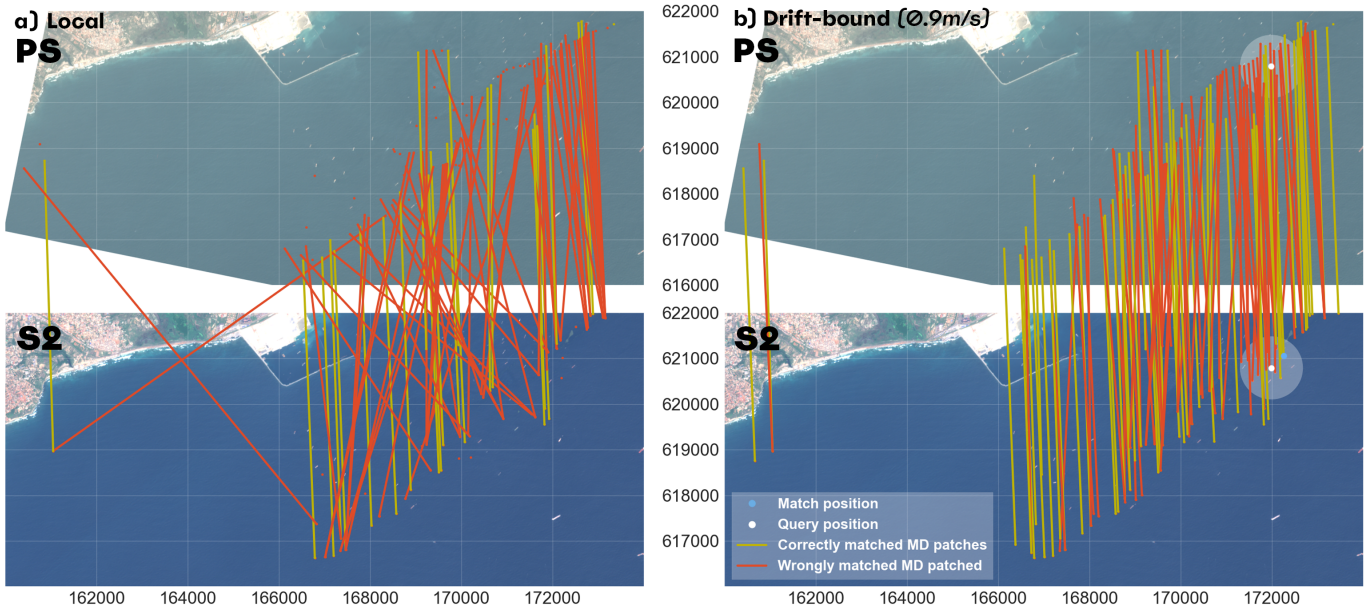

Figure 10: Comparison of top-1 predictions for: a) local Accra MD event-bound, and b) drift-bound scopes. Lines connect each MD patch in the PS scene with its top-1 match in the corresponding S2 scene. Note that red dots without lines indicate top-1 predictions that were wrongly predicted and did not fit into the extent of the figure. In b), the white circle indicates an example of search scope for a given query (white dot), while the blue dot is the true match position.

be attributed to their balanced training configuration: batches comprised 64 positive and 64 negative pairs. In contrast, remaining frameworks, namely DRSL, supervised SimCLR, and the proposed framework, adopted retrieval-like optimisation strategies and leveraged the whole batch to generate all possible negative pairs, making up as many as approximately 16,200 per batch.

Training with more negatives likely enabled the latter models to learn more discriminative representations, allowing easier positive and negative identification. These models had significantly lower average negative similarities (Figure 4), suggesting they were more confident in rejecting negative pairs. In contrast, SiamNN variants, trained with limited negatives, appear to have predicted similarity

scores that were not discriminative enough for the retrieval task.

A potential solution could have been inspired by Zheng et al. (2017), who trained their SiamNN with a gradual increase in negative pairs per epoch to avoid biases towards positive pairs in a retrieval setting.

Altogether, these findings highlight the importance of selecting an appropriate design for model architecture and training with the target task in mind. They also showcase how classification metrics can be misleading when evaluating models intended for retrieval applications.

Importance of Learning Relational Similarity

Despite using batch-leveraged negative sampling, the DRSL model underperformed relative to the other retrieval-like frameworks (i.e., CLIP and supervised SimCLR). Several possible factors may have contributed to this.

First, the DRSL model did not appear to converge within the 200 allocated training epochs (evident from training loss curves; not shown). This training behaviour may be attributed to a larger number of trainable parameters introduced by the additional FNNs, potentially hindering gradient flow and slowing convergence.

Second, DRSL was trained using the MSE loss, which is known to be sensitive to imbalanced datasets and tends to be biased towards the majority class (Wang et al., 2016). In this case, the large number of negative pairs likely led to overly conservative similarity predictions, particularly for positive pairs (as seen in the low average similarity value for positive pairs in Figure 4).

Although DRSL may have benefited from longer training and a more considerate design, such as a smaller batch size (as used in the original DRSL study (Wang et al., 2021)), selective negative sample mining, or cost-sensitive loss (Wang et al., 2016), the results state: despite relying on a more straightforward cosine similarity metric, the Double Acquisition NN outperformed the DRSL, which adopted a learnable similarity metric; and suggests that for multi-platform MD patch matching, a strong feature extractor and a simple similarity metric are sufficient to find modality-invariant space and effectively use it for similarity estimations.

Contrastive Learning with Batch-Negative Sampling Analysis

The strongest-performing frameworks across retrieval metrics, namely, the proposed Double Acquisition NN with CLIP loss and supervised SimCLR-Individual with all NT-Xent loss variants, were based on the InfoNCE loss. Their comparable performance, albeit with differences, suggests that contrastive optimisation with batch-based negative sampling is particularly well-suited for MD patch matching.

Loss Function Considerations

The similar results across NT-Xent loss variants were somewhat unexpected. The NT-Xent-Full variant incorporated additional negative pairs from the same platform, increasing the number of negative pairs to over 65,000 per batch. However, the similar performance results suggest that distinguishing PS-PS and S2-S2 negatives was irrelevant for the PS-S2 matching task. One possible explanation is that the model learned to differentiate PS and S2 MD patches based on embedded platform differences and used it as a shortcut, never considering such pairs as positives. Further experiments would be required to confirm this hypothesis.

The NT-Xent-In and NT-Xent-Out loss variants did not show a consistent advantage. These losses are designed for scenarios where multiple positive matches are possible per query (Hoffmann et al., 2022). In this study, each PS query MD patch typically had only one or at most four positive S2 tile matches, potentially limiting the benefits of such losses.

The CLIP loss was not tested independently of its architecture, making it challenging to interpret its potential advantages over the NT-Xent variants in isolation. However, some key distinctions between these losses are worth noting. NT-Xent-Full loss employed a broader negative pair definition, while the NT-Xent-In and -Out variants optimised the network only in the PS-to-S2 direction. In contrast, the CLIP loss is optimised in both PS-to-S2 and S2-to-PS directions. This symmetric optimisation strategy may have partially contributed to the observed performance improvement of the CLIP-based framework.

Another difference lies in the temperature scaling approach. All SimCLR NT-Xent variants used a fixed temperature parameter, while the CLIP-based proposed model adopted a trainable temperature, removing the need to tune it manually. The temperature was initialised in both frameworks at $\tau = 0.07$. During training, this temperature gradually decreased to approximately 0.05 in the proposed model. Although small, this temperature difference sharpened the softmax distribution, resulting in higher penalties for the hard negatives (i.e., negative samples predicted to be highly similar to the anchor). As noted by Wang and Liu (2021), lower temperatures shift the model's focus toward learning to discriminate the most challenging negatives rather than all negative samples. Lower temperatures potentially encouraged the proposed Double Acquisition NN model to learn more fine-grained distinctions between hard negatives and improve its ability to differentiate MD patch pairs.

Architectural Considerations

Beyond differences in InfoNCE-based loss implementation, architectural design may have also contributed to the performance differences between the CLIP-based Double Acquisition NN and the supervised SimCLR-Individual

frameworks. While both employed platform-specific encoders and batch-based contrastive learning, they differed in their projection schemes.

Supervised SimCLR used a non-linear, shared-weight projection head, whereas the proposed framework used individual linear projection layers in each branch. Sharing the projection head may have been suboptimal given the use of individual encoders, as it forced learning a modality-invariant projection scheme, potentially complicating alignment. Moreover, the non-linear projection head in SimCLR was originally designed to support transformation-invariant feature learning (Chen et al., 2020). However, in the present study, PS-S2 matching was essentially an image-to-image matching task without transformation-equivalent distortions: there were no large-scale rotations or random spatial shifts, such as those produced by random cropping. Consequently, a non-linear projection head may have introduced unnecessary complexity. CLIP study findings support this interpretation: researchers found linear projection layers sufficient for multi-modal alignment (Radford et al., 2021).

The CLIP-based framework consistently outperformed the supervised SimCLR variants. While several factors may explain this difference, the CLIP framework was ultimately preferred for the MD patch matching due to its performance enhancements and stronger conceptual alignment with its original design. The CLIP framework was specifically designed for multi-modal matching tasks. In contrast, the supervised SimCLR implementation was adapted from a single-modality contrastive learning framework, and its best implementation for MD patch matching had two individual encoders (deviating from the original SiamNN-like configuration). As such, the CLIP-based design had the conceptual advantage of being deliberately created to deal with multi-modal problems.

7.2. Impact of Hyperparameter Choices (RQ2)

Experimental results suggest that the proposed Double Acquisition NN benefits from moderately large batch sizes only when training is extended, likely due to high visual similarity between MD patches. The augmentation strategies tested in this study were detrimental to the model training, presumably due to the introduction of unrepresentative data variations. The relatively small dataset size likely constrains the potential effectiveness of deeper base encoder architectures, such as ResNet-50. Instead, the best-performing configuration used shallower ResNet-18 encoders pre-trained by SSL4EO, applied to each platform branch separately. This performance gain can be attributed to the already available feature extraction capabilities in pre-trained weights, facilitating more efficient fine-tuning for the MD patch matching task. Further refinement of SSL4EO weights for the marine domain through continued self-supervised SimCLR pre-training may have been misconfigured, producing a catastrophic forgetting effect, rendering refined weights worse than the original SSL4EO weights for the downstream task.

Batch Size Limitations

Typically, contrastive learning approaches benefit from large batch sizes or memory banks, which introduce more negative samples and promote robust feature learning by enhancing the model's discriminative abilities (Chen et al., 2020; He et al., 2020; Mitrovic et al., 2020).

In a restricted training duration regime, the benefits of larger batch sizes appeared to be inverted, possibly due to the nature of the collected dataset: MD patches often share similar shapes, sizes, and spectral characteristics (Figure 7 shows an example of top-3 predictions with high visual similarity). The visual similarity between MD patches makes them nearly indistinguishable, even for humans. Consequently, larger-batch models (i.e., with more negative pairs) were penalised more heavily for assigning high similarity scores to visually similar negatives. Such discouragement of high-value predictions resulted in more conservative predictions for positive pairs and a reduced average similarity range (Figure 5). This narrower prediction range is suboptimal for candidate ranking, as it diminishes the model's ability to distinguish between visually similar positive and negative samples. Mitrovic et al. (2020) highlighted that excessive amounts of negatives degrade the model performance, while Kalantidis et al. (2020) noted the importance of the quality of hard-negative samples over large quantities. Therefore, a more careful selection of negative samples may be required to benefit from a large-batch training.

The proposed Double Acquisition NN benefited from a large batch size only when trained for a longer duration. A potential explanation could be attributed to the positive-to-negative pair ratio per epoch. A 32-batch variant in each epoch was supplied with 2,166 positive and approximately 66,500 negative samples (a ratio of 1:30). In contrast, the 128-batch variant received the same amount of positives, together with approximately 274,700 negatives (a ratio of 1:125). This increased amount of negative samples in the larger-batch-size training settings may have diluted the positive signal and thus required longer training for the model to achieve comparable discriminative abilities. However, these improved results relative to smaller batch sizes suggest that with the longer training, the model was able to benefit from the increased batch size and the consequent number of negative pairs to learn to differentiate similarly looking MD patches better.

It is worth noting that training such larger-batch models for more allocated epochs significantly increased the training time. Therefore, the findings showcase a trade-off between performance and efficiency.

Risk of Unrealistic Augmentations

The finding that no augmentation outperformed all other strategies suggests that even mild augmentations introduced data alterations not encountered at the inference. Augmentations are beneficial only when they simulate realistic variations to virtually expand the training dataset

and allow the model to benefit from increased generalisation capacity. Conversely, unrealistic variations hinder the model's ability to learn robust features.

Performance degradation under harsh augmentations was not unexpected, as 90° rotations for MD patches are very rare, and were never observed during the annotation process. Mild augmentations (e.g., minor rotations) were assumed to be realistic based on RGB visual validation (see Figure H.1a). However, the independent application of a randomised augmentation strategy to each tile in a pair likely increased intra-pair differences, complicating the matching task. The brightness and contrast transformations, applied in both the medium and harsh augmentation strategies, were only visually validated in RGB spectral channels, which may have introduced unrealistic histogram shifts in the non-visible light spectrum. Altogether, these augmentations may have forced the model to learn representations that do not align with the nature of actual data.

The low performance scores from the spectral channel shuffling experiment suggest that forcing the model to rely solely on spatial information is ineffective, and spectral information is critical for matching MD patches. However, varied spatial resolution across S2 spectral channels may have introduced spatial artefacts, such as MD patch shape blurring upon shuffling (see reduced shape definition for the first S2 tile in Figure H.1d) and limited the model's ability to rely on spatial information. Additionally, the experimental model was initialised with SSL4EO weights, likely tuned to spectral information for feature extraction due to the critical role of the spectral signature in any general RS task. Therefore, the spectral channel shuffling may have had detrimental effects on available feature extraction capabilities.

In contrastive representation learning, the model's ability to learn better representations highly depends on a sufficiently difficult matching task. For self-supervised image-matching-based learning approaches, harsh augmentations are often used to create a challenging task and avoid learning identity mappings (Chen et al., 2020). However, the matching objective in this study differs from a typical image-matching task. Here, the MD patch pairs are not augmented views but rather two different captures of the same MD object. It can be argued that the inherent platform (modality) shift between PS and S2 already creates a challenging matching task. Any additional augmentations may have exaggerated these differences, making a matching task no longer relevant for learning robust features.

These findings align with previous studies, such as Wang and Qi (2022), which noted that too extreme augmentations interfere with effective representation learning if not carefully designed. The results of this study confirm the importance of investigating augmentations in domain-specific settings (Wang et al., 2022).

Encoder Depth Limitations

The deeper ResNet-50 architecture was less effective than the shallower ResNet-18 as a base encoder in the proposed Double Acquisition NN. One plausible explanation is the influence of pre-trained weight initialisation. Although both ResNet variants were initialised using weights from the SSL4EO study (pre-trained with MoCo on the same dataset), different depths and independent pre-training may have resulted in different initialisation positions within the solution space. It is possible that the ResNet-18 weights were more favourable for transfer to the MD patch matching task.

Some evidence of weight initialisation bias is apparent in performance differences observed in ResNet-50 variants with different initialisations (MoCo and DINO). Shekhar et al. (2023) reported a different observation: in their study, different contrastive SSL methods, including MoCo and DINO, learned to extract similar features across architectures like ResNet and Vision Transformers. However, researchers did not compare different ResNet depths. Therefore, architectural complexity could still be a factor in the downstream performance.

The inconsistencies with prior literature and lower performances of ResNet-50, regardless of the two tested initialisations, relative to ResNet-18, suggest that, even if weight initialisation bias was present, there were other underlying limitations of the ResNet-50 architecture for the MD patch matching task.

A more probable explanation is that ResNet-50 required a larger training dataset than was available. It is well established that deeper networks, capable of learning more complex representations, require larger datasets and longer training. Training such networks with a small dataset may lead to overfitting (Sun et al., 2017). Shallower networks offer a more robust alternative in limited data regimes by reducing the risk of over-parameterisation (Raghu et al., 2019; Brigato and Iocchi, 2021). Similar findings were reported by Du et al. (2017), who found that a shallower ResNet was more effective in a SiamNN architecture for handwriting matching, although researchers did not elaborate on the reason.

Overall, these findings align with the well-established notion that, in limited data regimes, the choice of model architecture plays a critical role in overall performance (Brigato and Iocchi, 2021). Given the scarcity of MD time-aware annotation data, the model architecture remains a key factor in model success.

Importance of Pre-training

The notable performance improvement of the proposed Double Acquisition NN when initialised with SSL4EO weights for each encoder branch may be attributed to the general feature extraction capabilities learned during pre-training, facilitating a smoother transfer to the MD patch matching task. Such an outcome was expected, as it is well-known that SSL-pre-trained models, when fine-tuned for the downstream task, require less labelled data and

converge faster and often achieve comparable or enhanced performance compared to models trained from scratch. Wang et al. (2022, 2023) found that fine-tuned SSL models consistently outperformed their supervised counterparts across various RS tasks, particularly in low-data regimes.

The underperformance of the configuration in which SSL4EO weights were applied only to the S2 branch may be explained by the model's overall objective to align PS and S2 features. In this configuration, the PS branch was forced to learn features aligned with those extracted from the pre-trained S2 encoder. This objective is particularly challenging, given the differences between PS and S2 imagery and the substantially smaller training dataset size relative to the scale of SSL4EO pre-training.

In contrast, the model in which both encoder branches were initialised with the same S2-based pre-trained weights achieved consistently strong results, suggesting that SSL4EO weights, although pre-trained on S2 data, are also beneficial for encoding PS imagery in the MD patch matching task context. One interpretation is that SSL4EO pre-trained models are not strictly platform-specific and can be transferred to multi-platform tasks, at least for the PS platform. Alternatively, the modality gap between PS and S2 may be smaller than previously discussed.

A key consideration for successfully transferring SSL pre-trained models to downstream tasks is the representativeness of the pre-training dataset relative to the target task (Wang et al., 2022). The original SSL4EO models were trained on RS imagery containing seasonal representations of various land cover types, but with limited ocean coverage (Wang et al., 2023). While this broad RS-domain pre-training was proven beneficial, it raises the question of whether models pre-trained on marine-specific data would yield further performance enhancements for marine-domain tasks. Currently, there are no publicly available ResNet-based weights for the marine domain. However, recent initiatives, such as Corley and Robinson (2024), have begun exploring marine-domain pre-training for transformer architectures.

Contrary to expectations, additional refinement of SSL4EO weights through self-supervised SimCLR pre-training did not improve downstream performance. On the contrary, results suggest the presence of catastrophic forgetting. It seems that the initial weight update in the first epochs of pre-training instantly degraded the representation extraction capacity of the original SSL4EO weights.

Catastrophic forgetting typically occurs when a model is subsequently trained on another distinct task (McCloskey and Cohen, 1989). However, in this study, self-supervised pre-training was not intended to shift the task but to refine the SSL4EO weights for the marine domain. The observed weight degradation may be attributed to either task incoherence introduced by the pretext task or too-radical optimisation procedures that shifted the initial feature space.

Task incoherence may have stemmed from an inappropriate augmentation strategy used in self-supervised SimCLR. The harsh augmentations may not have been sufficiently challenging or informative in an intra-platform setting. For example, ocean surface textures remained unaffected under harsh transformations (Figure H.1c), potentially creating shortcuts for positive pair identification. As a result, each branch encoder may have learned branch-specific feature extractions that were not valuable in a multi-platform setting, where, for example, water patterns rarely match. Such findings suggest a case of a too-specific pretext task resulting in reduced generalisation capacity (Chen et al., 2020; Wang et al., 2022).

In addition, hyperparameters such as learning rate and temperature may have caused an abrupt shift from the learned useful representation space. Chen et al. (2020) noted the importance of performance drop without proper temperature scaling. A more carefully tuned hyperparameter design may have mitigated the risk of catastrophic forgetting. Nonetheless, this would not address the previously discussed issue of pretext-task misalignment or eliminate the present study's limitations for *supervised* contrastive learning, which are also relevant for *self-supervised* pre-training: a small dataset size and lack of diverse negatives..

7.3. Impact of Candidate Selection Constraints (RQ3)

Many visually similar MD patches are, in fact, false positives when considered in spatial context. Therefore, the retrieval task of matching MD patches extends beyond designing an effective similarity ranking model. For the proposed Double Acquisition NN to be successful in applied scenarios, knowledge-informed candidate selection constraints are needed to minimise the inclusion of such false positives into the candidate set to reduce the complexity of the retrieval task and improve the accuracy in applied scenarios.

Retrieval Task Complexity

The relatively low performance of the model under the global regime does not necessarily indicate model failure; instead, it must be interpreted in the context of the task's inherent complexity. Assigning high similarity values to visually similar but spatially incorrect MD patches is not technically incorrect from a visual similarity standpoint. However, because MD patch matching also requires spatial correctness, such predictions ultimately become false positives in applied scenarios, reducing practical accuracy.

As observed during the annotation process and from visualised top-3 model predictions (Figure 7), MD patches often exhibit high visual similarity. This similarity makes the matching task particularly complex, especially under the global regime, where the model is equipped to retrieve the correct match from a set of 661 candidates, many of which are nearly indistinguishable in shape, size, or composition. This resemblance is especially pronounced for MD patches originating from the same MD event.

An additional source of complexity in distinguishing MD patches arises from platform limitations. Due to the

limited spectral information in PS imagery, it can be assumed that the PS encoder primarily relies on shape attributes for feature extraction. Conversely, the reduced spatial resolution of S2 imagery limits the visibility of fine-grained shape details. As a result, this could have complicated the cross-resolution MD matching and may have caused the reduced retrieval performance.

Some clustered annotations further amplify the retrieval complexity. In some cases, clustered MD patches were annotated individually, resulting in tiles that are only slightly offset in their centre coordinates but visually near-identical (e.g., Figure 7c and e). Since the model is requested to match MD patches centred within the tile, this annotation type increases the likelihood of confusion during retrieval.

One potential mitigation would be to space out annotations to reduce ambiguity over which MD patch is centred in the tile. However, in the intended application of the MD patch matching module in the automated tracking pipeline (see Figure Appendix A for such system composition), the MD detectors are expected to identify all visible MD patches, including those in clusters. Therefore, avoiding such annotations would not reflect real-world applications, and such a solution would be suboptimal.

Performance Improvement Through Drift-Informed Restrictions

Reducing candidate sets is a common strategy in information retrieval and recommendation systems to enhance retrieval efficiency and accuracy. For example, Borisyuk et al. (2016) showed how machine learning-based filtering of LinkedIn job advertisement candidates improved retrieval efficiency and the recommendation relevance.

Limiting the candidate scope from local to drift-bound regimes nearly doubled the top-1 accuracy of the Double Acquisition NN. Therefore, a similar principle of reducing the candidate set size to simplify the retrieval task applies to MD patch matching.

These performance enhancements largely stem from excluding visually similar but spatially implausible MD patches from the top-k subsets. This exclusion allows the model to focus on contextually realistic candidates rather than being forced to differentiate spatially distant but highly similar MD patches. Additionally, in some drift-informed cases, the candidate set may have been restricted to the true match alone, effectively making the retrieval task trivial. Although these cases simplify the retrieval task, they remain a valid representation of a realistic, well-informed candidate selection. In addition, these cases were rare due to the tendency of MD patches to occur in clusters (Figures 9 and 10), often resulting in multiple candidates within the drift-informed search area.

Increased performance through the drift-bound regime aligns with the initial intuition that MD patches drift within a limited spatial area. This assumption was partially inspired by manual MD tracking (e.g., estimating the expected MD drift trajectory for search efforts in multi-temporal scenes; Weiß et al. (2022)) and emergency response efforts (e.g., back-tracking MH370 flight debris to the possible crash location; Durgadoo et al. (2019)), both of which leveraged ocean drift models to constrain the spatial extent of the search area based on expected drift behaviour.

This study used maximum drift speed as a more easily estimated and applicable proxy to restrict the candidate search space. This approach avoids the need for complex and accurate ocean drift models, often unreliable in coastal areas (Pereiro et al., 2018). In applied scenarios, these drift speed estimates can be derived from windage coefficients combined with ocean current and wind data (Pereiro et al., 2018; Durgadoo et al., 2019; Li et al., 2020). However, drift-informed candidate selection is susceptible to the accuracy of these estimates. In the case of underestimation, the true match may be excluded from the candidate set, forcing the model to retrieve the most visually similar false positive. In contrast, overestimating maximum drift speed may increase the candidate set size and the likelihood of including other visually similar MD patches, thereby increasing task complexity. As such, uncertainty in drift-speed estimation remains a limiting factor in deploying the Double Acquisition NN in applied scenarios.

Additional Candidate Selection Filters

While the drift-bound regime substantially improved retrieval performance, the top-1 accuracy remained imperfect. A plausible explanation could be the clustered annotation problem introduced earlier. As the drift-bound regime restricted the candidate set to spatially neighbouring MD patches, these clustered annotations increased the likelihood of the model retrieving a visually similar neighbouring MD patch rather than the true match. This may help explain the observed mean position of 1.9, which indicates that while the correct match was usually among the top predictions, it was not always the first.

Additional strategies for candidate filtering could target this problem and reduce false positives, especially in scenarios where prior knowledge of drift patterns is unavailable or unreliable. Several potential candidate filters are:

- A logic-based spatial filter, applied in the absence of drift knowledge. Since MD patches drift passively, there is a physical limit beyond which natural drift becomes infeasible.
- A drift behaviour filter, which exploits drift patterns without relying on drift models. Candidates can be re-ranked based on how well their predicted drift trajectory (distance and direction) aligns with the dominant drift patterns observed among neighbouring MD patch pairs. Figures 9 and 10 show almost parallel alignment of drift trajectories for *correct* top-1 predictions, whereas *incorrect* predictions often show

directional inconsistencies. However, this filtering approach may not always be reliable, as non-parallel, eddy-induced trajectories were observed during the annotation process.

- A low-confidence match filter, which discards or flags predictions for manual validation, when, for example, the top-1 candidate has a low estimated similarity, or top-k predictions have nearly equal estimated similarities, indicating high confusion between most likely candidates.

- A greedy-match filter, which iteratively matches the most likely pairs and removes the retrieved candidate from the candidate set. This approach could potentially reduce confusion between clustered near-duplicates.

While none of these strategies were tested in this study, they illustrate the potential benefits of incorporating simple yet informed candidate selection to increase model performance further.

Model Generalisation Concerns

The comparable performance observed across global and local regimes suggests that the model learned to identify and restrict top candidates to a local scope. This outcome may be attributed to the specifics of each MD-event and data-induced explanations.

The Marmara MD event was a mucilage bloom, visually exhibiting more intricate structures (Figure 7d and e). In contrast, the Accra MD patches, composed of macroalgae and spume, had more uniform round shapes and tended to occur in clusters (Figure 7a, b and c). These visual differences, potentially including spectral signature differences, may have made it easier for the model to distinguish Marmara patches from Accra. The intricate and easier-to-differentiate MD formations in Marmara may also explain its higher, relative to Accra, per-event performance scores.

From a data perspective, two factors have to be accounted for. First, the Marmara MD event constituted the majority of training tiles, likely biasing the model to be more discriminative toward mucilage structures present in this event. Second, the dataset split was spatially but not temporally independent. The latter may have allowed the model to exploit event-specific artefacts such as histogram shifts, potentially caused by environmental conditions (e.g., water and atmospheric conditions, or daylight illumination). These observations raise concerns about the model's ability to generalise to unseen MD events. Such generalisation challenges in the marine domain were also noted by Carmo et al. (2021), who found that the performance of marine floating object detection models was highly dependent on the training scenes.

8. Conclusions

The growing urgency to mitigate marine plastic pollution calls for reliable methods to monitor and track its presence over time, enabling the development of knowledge-based solutions. This study addresses the tracking need by compiling the first annotation dataset of PS-S2 MD patch pairs and developing the Double Acquisition NN—a model designed to match MD patches captured by the two platforms within a one-hour interval.

The Double Acquisition NN employs two platform-specific ResNet-18 encoders, each followed by individual linear projection layers. It is trained using a supervised contrastive loss. A systematic examination of multiple similarity estimation frameworks and hyperparameter configurations led to the following key findings:

- Platform-specific encoders outperform shared ones, confirming a modality gap between PS and S2 platforms.

- Frameworks designed for retrieval tasks (e.g., CLIP) outperformed frameworks designed for classification problems (e.g., SiamNN), primarily due to the larger number of negative samples used for training.

- In the MD patch matching task, simple similarity metrics combined with strong feature extraction outperform more complex, explicitly learned relational similarity networks.

- Moderate batch sizes are optimal in limited data regimes; larger batches overly penalise visually similar negatives, leading to overly conservative prediction and reduction of the model's distinguishing abilities.

- The tested augmentation strategies degraded model performance due to poor design, highlighting the importance of domain-informed augmentation strategies.

- Shallower encoders (i.e., ResNet-18) yielded more robust features than deeper ones (i.e., ResNet-50) under limited data regimes.

- SSL4EO pre-trained weights improved model performance in the marine domain. However, poorly designed additional self-supervised weight refinement led to catastrophic forgetting, highlighting the importance of pretext task alignment and careful implementation.

The proposed Double Acquisition NN evaluation under different candidate selection regimes revealed that:

- MD patch matching is a complex task, due to high visual similarity between MD patches and the need for spatial correctness.

- Restricting the candidate search space based on knowledge is critical to improve applied retrieval performance.
- However, the performance gains from candidate selection are limited by the accuracy of the prior knowledge used.

Beyond their original purposes, both the dataset and model offer potential for broader MD and plastic waste monitoring context applications. The newly collected dataset can support platform-specific tasks (e.g., MD detection in PS or S2 imagery) and is particularly valuable for studying the temporal dynamics of MD. The combination of spatial and temporal information in annotations has potential for deriving MD drift coefficients and calibrating and validating drift models.

The Double Acquisition NN could support two use cases. First, it could be a verification tool for identifying plastic debris in PS MD patches when the presence is confirmed in corresponding S2 MD patches through spectral signature analysis. Second, the model may be configured for PS-PS MD patch matching due to its individual encoder architecture, enabling more frequent and temporally extended tracking.

This study contributes a multi-temporal dataset and a retrieval model, which are key components for developing an automated MD tracking system. Future work could focus on integrating the proposed Double Acquisition NN with MD detectors and more robust filtering mechanisms, forming a fully automated tracking pipeline. Another promising direction lies in improving the retrieval accuracy of the proposed Double Acquisition NN by incorporating more robust feature extraction strategies, achieved through the collection of larger datasets and large-scale training, or by leveraging the capabilities of more complex encoder architectures and marine domain pre-trained models.

Acknowledgements

The author thanks Marc Rußwurm for supervision, insightful discussion and comments, suggestions on alternative methods and continuous support; and Emanuele Dalsasso for supervision, insightful discussion and comments and for sharing valuable literature recommendations.

Declaration of Generative AI

During the preparation of this work, the author used generative AI to troubleshoot code issues and initialise visualisation design templates. DeepL and ChatGPT were used as writing tools to improve academic tone, readability, and grammatical accuracy. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication. All methodology and analysis are the work of the author or inspired by cited literature.

References

Ahn, W.J., Ko, K.S., Lim, M.T., Pae, D.S., Kang, T.K., 2023. Multiple object tracking using re-identification model with attention module. Applied Sciences 13, 4298. doi:https://doi.org/10.3390/app13074298.

Andrady, A.L., 2015. Persistence of Plastic Litter in the Oceans. Springer International Publishing. chapter 3. p. 57–72. doi:https://doi.org/10.1007/978-3-319-16510-3_3.

Bajon, R., Huck, T., Grima, N., Maes, C., Blanke, B., Richon, C., Couvelard, X., 2023. Influence of waves on the three-dimensional distribution of plastic in the ocean. Marine Pollution Bulletin 187, 114533. doi:https://doi.org/10.1016/j.marpolbul.2022.114533.

Barnes, D.K.A., Galgani, F., Thompson, R.C., Barlaz, M., 2009. Accumulation and fragmentation of plastic debris in global environments. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 1985–1998. doi:https://doi.org/https://doi.org/10.1098/rstb.2008.0205.

Basu, B., Sannigrahi, S., Sarkar Basu, A., Pilla, F., 2021. Development of novel classification algorithms for detection of floating plastic debris in coastal waterbodies using multispectral sentinel-2 remote sensing imagery. Remote Sensing 13, 1598. doi:https://doi.org/10.3390/rs13081598.

Béchaz, M., 2024. Modeling and predicting oceanic drifts of plastic waste with physical and data-driven approaches. Master's thesis. École Polytechnique Fédérale de Lausanne.

Biermann, L., Clewley, D., Martinez-Vicente, V., Topouzelis, K., 2020. Finding plastic patches in coastal waters using optical satellite data. Scientific Reports 10. doi:https://doi.org/10.1038/s41598-020-62298-z.

Borisyuk, F., Kenthapadi, K., Stein, D., Zhao, B., 2016. Casmos: A framework for learning candidate selection models over structured queries and documents, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. p. 441–450. doi:https://doi.org/10.1145/2939672.2939718.

Bosi, S., Broström, G., Roquet, F., 2021. The role of stokes drift in the dispersal of north atlantic surface marine debris. Frontiers in Marine Science 8. doi:https://doi.org/10.3389/fmars.2021.697430.

Brigato, L., Iocchi, L., 2021. A close look at deep learning with small data, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. p. 2490–2497. doi:https://doi.org/10.1109/icpr48806.2021.9412492.

Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., Lecun, Y., Moore, C., Sackinger, E., Shan, R., 1993. Signature Verification Using a "Siamese" Time Delay Neural Network. International Journal of Pattern Recognition and Artificial Intelligence 07, 669–688. doi:https://doi.org/10.1142/s0218001493000339.

Carbery, M., O'Connor, W., Palanisami, T., 2018. Trophic transfer of microplastics and mixed contaminants in the marine food web and implications for human health. Environment International 115, 400–409. doi:https://doi.org/10.1016/j.envint.2018.03.007.

Carmo, R., Mifdal, J., Rußwurm, M., 2021. Detecting macro floating objects on coastal water bodies using sentinel-2 data, in: OCEANS 2021: San Diego – Porto, IEEE. p. 1–7. doi:https://doi.org/10.23919/oceans44145.2021.9705668.

Chamas, A., Moon, H., Zheng, J., Qiu, Y., Tabassum, T., Jang, J.H., Abu-Omar, M., Scott, S.L., Suh, S., 2020. Degradation rates of plastics in the environment. ACS Sustainable Chemistry and Engineering 8, 3494–3511. doi:https://doi.org/10.1021/acssuschemeng.9b06635.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 URL: https://arxiv.org/abs/2002.05709, arXiv:2002.05709.

Chicco, D., 2020. Siamese Neural Networks: An Overview. Springer US. chapter 3. p. 73–94. doi:https://doi.org/10.1007/978-1-0716-0826-5_3.

Corley, I., Robinson, C., 2024. Hydro foundation model. URL: https://github.com/isaaccorley/hydro-foundation-model.

Cózar, A., Aliani, S., Basurko, O.C., Arias, M., Isobe, A., Topouzelis, K., Rubio, A., Morales-Caselles, C., 2021. Marine Litter Windrows: A Strategic Target to Understand and Manage the Ocean Plastic Pollution. Frontiers in Marine Science 8, 571796. doi:https://doi.org/10.3389/FMARS.2021.571796/BIBTEX.

Cózar, A., Arias, M., Suaria, G., Viejo, J., Aliani, S., Koutroulis, A., Delaney, J., Bonnery, G., Macías, D., de Vries, R., Sumerot, R., Morales-Caselles, C., Turiel, A., González-Fernández, D., Corradi, P., 2024. Proof of concept for a new sensor to monitor marine litter from space. Nature Communications 2024 15:1 15, 1–12. doi:https://doi.org/10.1038/s41467-024-48674-7.

Dalsasso, E., Russwurm, M., Donner, C., Vries, R.d., Volpi, M., Tuia, D., 2025. A cross-sensor approach for marine litter detection with self-supervised learning. EGU25 doi:https://doi.org/10.5194/EGUSPHERE-EGU25-8279.

Du, W., Fang, M., Shen, M., 2017. Siamese Convolutional Neural Networks for Authorship Verification, pp. 1–8. URL: http://cs231n.stanford.edu/reports/2017/pdfs/801.pdf.

Duarte, M.M., Azevedo, L., 2023. Automatic detection and identification of floating marine debris using multispectral satellite imagery. IEEE Transactions on Geoscience and Remote Sensing 61, 1–15. doi:https://doi.org/10.1109/TGRS.2023.3283607.

Durgadoo, J.V., Biastoch, A., New, A.L., Rühs, S., Nurser, A.J., Drillet, Y., Bidlot, J.R., 2019. Strategies for simulating the drift of marine debris. Journal of Operational Oceanography 14, 1–12. doi:https://doi.org/10.1080/1755876x.2019.1602102.

Eriksen, M., Cowger, W., Erdle, L.M., Coffin, S., Villarrubia-Gómez, P., Moore, C.J., Carpenter, E.J., Day, R.H., Thiel, M., Wilcox, C., 2023. A growing plastic smog, now estimated to be over 170 trillion plastic particles afloat in the world's oceans—urgent solutions required. PLOS ONE 18, e0281596. doi:https://doi.org/10.1371/journal.pone.0281596.

Gall, S., Thompson, R., 2015. The impact of debris on marine life. Marine Pollution Bulletin 92, 170–179. doi:https://doi.org/10.1016/j.marpolbul.2014.12.041.

Garaba, S.P., Aitken, J., Slat, B., Dierssen, H.M., Lebreton, L., Zielinski, O., Reisser, J., 2018. Sensing ocean plastics with an airborne hyperspectral shortwave infrared imager. Environmental Science and Technology doi:https://doi.org/10.1021/acs.est.8b02855.

Gerin, R., Poulain, P.M., Besiktepe, S.T., Zanasca, P., 2013. On the surface circulation of the marmara sea as deduced from drifters. TURKISH JOURNAL OF EARTH SCIENCES 22, 919–930. doi:https://doi.org/10.3906/yer-1202-8.

Goddijn-Murphy, L., Martínez-Vicente, V., Dierssen, H.M., Raimondi, V., Gandini, E., Foster, R., Chirayath, V., 2024. Emerging technologies for remote sensing of floating and submerged plastic litter. Remote Sensing 16, 1770. doi:https://doi.org/10.3390/rs16101770.

Goddijn-Murphy, L., Peters, S., van Sebille, E., James, N.A., Gibb, S., 2018. Concept for a hyperspectral remote sensing algorithm for floating marine macro plastics. Marine Pollution Bulletin 126, 255–262. doi:https://doi.org/10.1016/j.marpolbul.2017.11.011.

He, H., Chen, M., Chen, T., Li, D., 2018. Matching of remote sensing images with complex background variations via siamese convolutional neural network. Remote Sensing 10, 355. doi:https://doi.org/10.3390/rs10020355.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 9729–9738. doi:https://doi.org/10.1109/cvpr42600.2020.00975.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 1–12. doi:https://doi.org/10.1109/cvpr.2016.90.

Hoffmann, D.T., Behrmann, N., Gall, J., Brox, T., Noroozi, M., 2022. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. Proceedings of the AAAI Conference on Artificial Intelligence 36, 897–905. doi:https://doi.org/10.1609/aaai.v36i1.19972.

Hu, C., 2022. Remote detection of marine debris using sentinel-2 imagery: A cautious note on spectral interpretations. Marine Pollution Bulletin 183, 114082. doi:https://doi.org/10.1016/j.marpolbul.2022.114082.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. doi:https://doi.org/10.48550/ARXIV.1502.03167.

Jambeck, J.R., Geyer, R., Wilcox, C., Siegler, T.R., Perryman, M., Andrady, A., Narayan, R., Law, K.L., 2015. Plastic waste inputs from land into the ocean. Science 347, 768–771. doi:https://doi.org/10.1126/science.1260352.

Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D., 2020. Hard negative mixing for contrastive learning. doi:https://doi.org/10.48550/ARXIV.2010.01028.

Karakuş, O., 2023. On advances, challenges and potentials of remote sensing image analysis in marine debris and suspected plastics monitoring. Frontiers in Remote Sensing 4. doi:https://doi.org/10.3389/frsen.2023.1302384.

Kikaki, A., Karantzalos, K., Power, C.A., Raitsos, D.E., 2020. Remotely sensing the source and transport of marine plastic debris in bay islands of honduras (caribbean sea). Remote Sensing 12, 1727. doi:https://doi.org/10.3390/rs12111727.

Kikaki, K., Kakogeorgiou, I., Hoteit, I., Karantzalos, K., 2024. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. ISPRS Journal of Photogrammetry and Remote Sensing 210, 39–54. doi:https://doi.org/10.1016/j.isprsjprs.2024.02.017.

Lavender, S., 2022. Detection of waste plastics in the environment: Application of copernicus earth observation data. Remote Sensing 14, 4772. doi:https://doi.org/10.3390/rs14194772.

Lebreton, L., Andrady, A., 2019. Future scenarios of global plastic waste generation and disposal. Palgrave Communications 5. doi:https://doi.org/10.1057/s41599-018-0212-7.

Lebreton, L.C.M., van der Zwet, J., Damsteeg, J.W., Slat, B., Andrady, A., Reisser, J., 2017. River plastic emissions to the world's oceans. Nature Communications 8. doi:https://doi.org/10.1038/ncomms15611.

Lermusiaux, P.F.J., Marshall, J., Peacock, T., Noble, C., Doshi, M., Kulkarni, C.S., Gupta, A., Haley, P.J., Mirabito, C., Trotta, F., Levang, S.J., Flierl, G.R., 2019. Plastic pollution in the coastal oceans: Characterization and modeling, in: OCEANS 2019 MTS/IEEE SEATTLE, IEEE. p. 1–10. doi:https://doi.org/10.23919/oceans40490.2019.8962786.

Li, Y., Zhang, H., Tang, C., 2020. A review of possible pathways of marine microplastics transport in the ocean. Anthropocene Coasts 3, 6–13. doi:https://doi.org/10.1139/anc-2018-0030.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110. doi:https://doi.org/10.1023/b:visi.0000029664.99615.94.

Matthews, J.P., Ostrovsky, L., Yoshikawa, Y., Komori, S., Tamura, H., 2017. Dynamics and early post-tsunami evolution of floating marine debris near fukushima daiichi. Nature Geoscience 10, 598–603. doi:https://doi.org/10.1038/ngeo2975.

Maximenko, N., Corradi, P., Law, K.L., Van Sebille, E., Garaba, S.P., Lampitt, R.S., Galgani, F., Martinez-Vicente, V., Goddijn-Murphy, L., Veiga, J.M., Thompson, R.C., Maes, C., Moller, D., Löscher, C.R., Addamo, A.M., Lamson, M.R., Centurioni, L.R., Posth, N.R., Lumpkin, R., Vinci, M., Martins, A.M., Pieper, C.D., Isobe, A., Hanke, G., Edwards, M., Chubarenko, I.P., Rodriguez, E., Aliani, S., Arias, M., Asner, G.P., Brosich, A., Carlton, J.T., Chao, Y., Cook, A.M., Cundy, A.B., Galloway, T.S., Giorgetti, A., Goni, G.J., Guichoux, Y., Haram, L.E., Hardesty, B.D., Holdsworth, N., Lebreton, L., Leslie, H.A., Macadam-Somer, I., Mace, T., Manuel, M., Marsh, R., Martinez, E., Mayor, D.J., Le Moigne, M., Molina Jack, M.E., Mowlem, M.C., Obbard, R.W., Pabortsava, K., Robberson, B., Rotaru, A.E., Ruiz, G.M., Spedicato, M.T., Thiel, M., Turra, A., Wilcox, C., 2019. Toward the integrated marine debris observing system. Frontiers in Ma-

rine Science 6. doi:https://doi.org/10.3389/fmars.2019.00447.

McCloskey, M., Cohen, N.J., 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. Elsevier. p. 109–165. doi:https://doi.org/10.1016/s0079-7421(08)60536-8.

Mifdal, J., Longépé, N., Rußwurm, M., 2021. Towards detecting floating objects on a global scale with learned spatial features using sentinel 2. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-3–2021, 285–293. doi:https://doi.org/10.5194/isprs-annals-v-3-2021-285-2021.

Mikeli, P., Kikaki, K., Kakogeorgiou, I., Karantzalos, K., 2022. How challenging is the discrimination of floating materials on the sea surface usin high resolution multispectral satellite data? The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B3-2022, 151–157. doi:https://doi.org/10.5194/isprs-archives-xliii-b3-2022-151-2022.

Mitrovic, J., McWilliams, B., Rey, M., 2020. Less can be more in contrastive learning, in: Zosa Forde, J., Ruiz, F., Pradier, M.F., Schein, A. (Eds.), Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops, PMLR. pp. 70–75. URL: https://proceedings.mlr.press/v137/mitrovic20a.html.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. doi:https://doi.org/10.48550/ARXIV.1807.03748.

Pereiro, D., Souto, C., Gago, J., 2018. Calibration of a marine floating litter transport model. Journal of Operational Oceanography 11, 125–133. doi:https://doi.org/10.1080/1755876x.2018.1470892.

Plastics Europe, 2024. Plastics – the fast facts 2024. URL: https://plasticseurope.org/knowledge-hub/plastics-the-fast-facts-2024/. fact Sheet.

Politikos, D.V., Adamopoulou, A., Petasis, G., Galgani, F., 2023. Using artificial intelligence to support marine macrolitter research: A content analysis and an online database. Ocean and Coastal Management 233, 106466. doi:https://doi.org/10.1016/j.ocecoaman.2022.106466.

Qiu, K., Ai, Y., Tian, B., Wang, B., Cao, D., 2018. Siamese-resnet: Implementing loop closure detection based on siamese network, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE. p. 716–721. doi:https://doi.org/10.1109/ivs.2018.8500465.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. doi:https://doi.org/10.48550/ARXIV.2103.00020.

Raghu, M., Zhang, C., Kleinberg, J.M., Bengio, S., 2019. Transfusion: Understanding transfer learning with applications to medical imaging. CoRR abs/1902.07208. URL: http://arxiv.org/abs/1902.07208, arXiv:1902.07208.

Rösch, M., Sonnenschein, R., Buchelt, S., Ullmann, T., 2022. Comparing planetscope and sentinel-2 imagery for mapping mountain pines in the sarntal alps, italy. Remote Sensing 14, 3190. doi:https://doi.org/10.3390/rs14133190.

Rußwurm, M., Venkatesa, S.J., Tuia, D., 2023. Large-scale detection of marine debris in coastal areas with sentinel-2. iScience 26, 108402. doi:https://doi.org/10.1016/j.isci.2023.108402.

Salgado-Hernanz, P.M., Bauzà, J., Alomar, C., Compa, M., Romero, L., Deudero, S., 2021. Assessment of marine litter through remote sensing: recent approaches and future goals. Marine Pollution Bulletin 168, 112347. doi:https://doi.org/10.1016/j.marpolbul.2021.112347.

Sannigrahi, S., Basu, B., Basu, A.S., Pilla, F., 2022. Development of automated marine floating plastic detection system using sentinel-2 imagery and machine learning models. Marine Pollution Bulletin 178, 113527. doi:https://doi.org/10.1016/j.marpolbul.2022.113527.

van Sebille, E., Wilcox, C., Lebreton, L., Maximenko, N., Hardesty, B.D., van Franeker, J.A., Eriksen, M., Siegel, D., Galgani, F., Law, K.L., 2015. A global inventory of small floating plastic debris. Environmental Research Letters 10, 124006. doi:https://doi.org/10.1088/1748-9326/10/12/124006.

Shekhar, S., Bordes, F., Vincent, P., Morcos, A., 2023. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations. doi:https://doi.org/10.48550/ARXIV.2304.13089, arXiv:2304.13089.

Shen, A., Zhu, Y., Angelov, P., Jiang, R., 2024. Marine debris detection in satellite surveillance using attention mechanisms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17, 4320–4330. doi:https://doi.org/10.1109/jstars.2024.3349489.

Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 843–852. doi:https://doi.org/10.1109/iccv.2017.97.

Tang, S., Zhou, S., Kang, W., Wu, Q., Deng, F., 2019. Finger vein verification using a siamese cnn. IET Biometrics 8, 306–315. doi:https://doi.org/10.1049/iet-bmt.2018.5245.

The Guardian, 2018. Water rises to three feet in st mark's basilica after venice floods. The Guardian URL: https://www.theguardian.com/weather/2018/oct/30/water-rises-to-three-feet-in-st-marks-basilica-after-venice-floods. accessed: 2024-10-19.

Themistocleous, K., Papoutsa, C., Michaelides, S., Hadjimitsis, D., 2020. Investigating detection of floating plastic litter from space using sentinel-2 imagery. Remote Sensing 12, 2648. doi:https://doi.org/10.3390/rs12162648.

Topouzelis, K., Papageorgiou, D., Karagaitanakis, A., Papakonstantinou, A., Arias Ballesteros, M., 2020. Remote sensing of sea surface artificial floating plastic targets with sentinel-2 and unmanned aerial systems (plastic litter project 2019). Remote Sensing 12, 2013. doi:https://doi.org/10.3390/rs12122013.

Topouzelis, K., Papageorgiou, D., Suaria, G., Aliani, S., 2021. Floating marine litter detection algorithms and techniques using optical remote sensing data: A review. Marine Pollution Bulletin 170, 112675. doi:https://doi.org/10.1016/j.marpolbul.2021.112675.

Topouzelis, K., Papakonstantinou, A., Garaba, S.P., 2019. Detection of floating plastics from satellite and unmanned aerial systems (plastic litter project 2018). International Journal of Applied Earth Observation and Geoinformation 79, 175–183. doi:https://doi.org/10.1016/j.jag.2019.03.011.

Wang, F., Liu, H., 2021. Understanding the behaviour of contrastive loss, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. p. 2495–2504. doi:https://doi.org/10.1109/cvpr46437.2021.00252.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J., 2016. Training deep neural networks on imbalanced data sets, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE. p. 4368–4374. doi:https://doi.org/10.1109/ijcnn.2016.7727770.

Wang, X., Hu, P., Zhen, L., Peng, D., 2021. Drsl: Deep relational similarity learning for cross-modal retrieval. Information Sciences 546, 298–311. doi:https://doi.org/10.1016/j.ins.2020.08.009.

Wang, X., Qi, G.J., 2022. Contrastive learning with stronger augmentations. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–12doi:https://doi.org/10.1109/tpami.2022.3203630.

Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X., 2022. Self-supervised learning in remote sensing: A review. IEEE Geoscience and Remote Sensing Magazine 10, 213–247. doi:https://doi.org/10.1109/mgrs.2022.3198244.

Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X., 2023. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. URL: https://arxiv.org/abs/2211.07044, arXiv:2211.07044.

Waring, R., Harris, R., Mitchell, S., 2018. Plastic contamination of the food chain: A threat to human health? Maturitas 115, 64–68. doi:https://doi.org/10.1016/j.maturitas.2018.06.010.

Weiß, T., Bochow, M., Ghezzo, M., Cester, I., 2022. Detection and tracking of large marine litter based on high-resolution remote sensing time series, machine learning, and ocean current

modelling. Executive Summary Report. Technical Report. EuropeanSpaceAgency. URL: https://nebula.esa.int/content/detection-and-tracking-large-marine-litter-based-high-resolution-remote-sensing-time-series.

Yagci, A.L., Colkesen, I., Kavzoglu, T., Sefercik, U.G., 2022. Daily monitoring of marine mucilage using the modis products: a case study of 2021 mucilage bloom in the sea of marmara, turkey. Environmental Monitoring and Assessment 194. doi:https://doi.org/10.1007/s10661-022-09831-x.

Zhang, Y., Wu, P., Xu, R., Wang, X., Lei, L., Schartup, A.T., Peng, Y., Pang, Q., Wang, X., Mai, L., Wang, R., Liu, H., Wang, X., Luijendijk, A., Chassignet, E., Xu, X., Shen, H., Zheng, S., Zeng, E.Y., 2023. Plastic waste discharge to the global ocean constrained by seawater observations. Nature Communications 14. doi:https://doi.org/10.1038/s41467-023-37108-5.

Zheng, Z., Zheng, L., Yang, Y., 2017. A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications 14, 1–20. doi:https://doi.org/10.1145/3159171.

Appendix A. Proposed Automated MD Tracking System

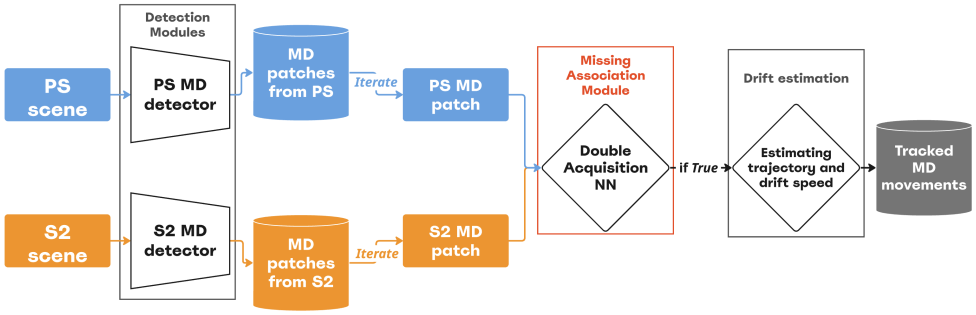

Figure A.1: Automated Marine Debris (MD) tracking system, inspired by Ahn et al. (2023). In this system, satellite scenes are subjected to MD patch detectors, which extract spatial information about detected objects into MD patch databases. Each detected object in a PlanetScope (PS) scene is then compared against candidates, each detected object in the corresponding Sentinel-2 (S2) scene, using the Double Acquisition Neural Network (NN), developed by this study. If two MD patches are identified to be matching, their drift trajectory and velocity can be estimated and recorded in the final database of tracked MD movements.

Appendix B. Dataset Structure and Composition

Table B.1 provides an overview of the marine debris (MD) events and the corresponding Scene IDs of PlanetScope (PS) and Sentinel-2 (S2) imagery. These scenes were used for manual MD double acquisition annotations and subsequent tile generation.

The collected PS-S2 double acquisition MD patch annotations, published alongside this study, are in a structured database format. The database consists of three tables: Scene, Patch, and Match (see Figure B.1). Together, these three tables provide spatial and temporal information for each MD patch and information about its source scene.

- The Scene table contains metadata about the scenes in which MD patches were annotated.
- The Patch table contains all annotation information for MD patches from both PS and S2 platforms. A patch entry has MD patch centre coordinates, provided in the coordinate reference system (CRS), defined by its corresponding scene. Each MD patch entry references its parent scene, allowing the acquisition information and platform metadata to be traced.
- The Match table links two entries from the Patch table into a positive match - a pair composed of an MD patch annotation in PS and the corresponding patch annotation in S2.

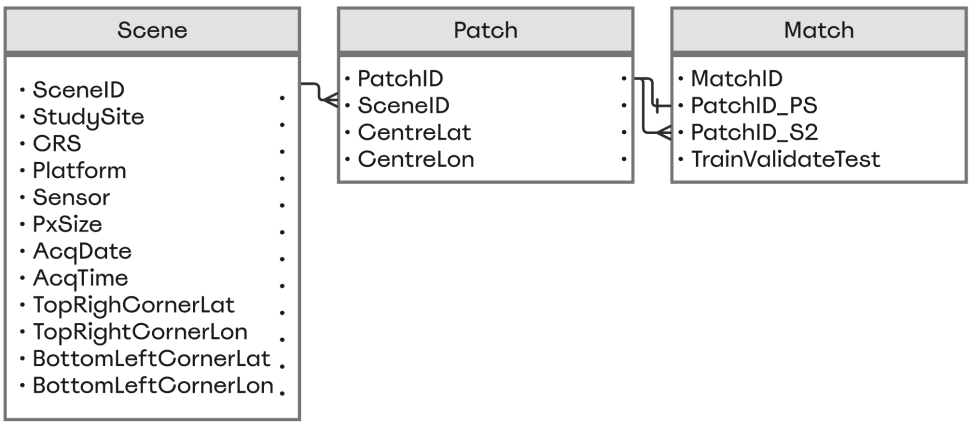

Figure B.1: Structure of the database containing information about the collected dataset. The Scene table includes: scene ID, study site name, local UTM CRS, platform source (PS or S2), sensor (for PS scenes), acquisition date and time, and bounding box coordinates. The Patch table includes: MD patch annotation coordinates and reference to the associated scene. The Match table links each MD patch annotation from PS with the corresponding MD patch annotation from S2. It provides information on whether a match is used for training, validation, or testing.

Table B.1: MD events and the corresponding Scene IDs used in this study for annotating MD pairs and generating tiles.

MD Event	**Date**	**S2 Scene IDs**	**PS Scene IDs**
Bay Islands, Honduras	2017-10-09	20171009T161341	20171009_153515_0f25, 20171009_153514_0f25, 20171009_153609_103a, 20171009_153610_103a, 20171009_153729103c
Venice, Italy	2018-06-30	20180630T100029	20180630_093120_1038, 20180630_093121_1038, 20180630_093122_1038, 20180630_093123_1038, 20180630_093124_1038, 20180630_093125_1038, 20180630_093126_1038
Calabria, Italy	2018-10-22	20181022T094029	20181022_085429_0f2b, 20181022_085430_0f2b, 20181022_085431_0f2b, 20181022_091141_0e20, 20181022_091552_1002, 20181022_091553_1002
Accra, Ghana	2018-10-31	20181031T101139	20181031_095646_101b, 20181031_095847_0f43, 20181031_095848_0f43, 20181031_095850_0f43, 20181031_095925_103b, 20181031_095926_103b
Venice, Italy	2018-10-31	20181031T10113	20181031_093318_0e3a, 20181031_093404_1004, 20181031_093405_1004, 20181031_093406_1004, 20181031_093407_1004, 20181031_093408_1004
Lagos, Nigeria	2019-01-01	20190101T100411	20190101_094700_1011, 20190101_095118_0f2a
Durban, South Africa	2019-04-24	20190424T073619	20190424_073843_1105, 20190424_080100_0f2d
Thassos, Greece	2021-04-30	20210430T090549	20210430_082515_04_242d
Marmara, Turkey	2021-05-19	20210519T084601	20210519_080717_05_2440, 20210519_081237_40_106c, 20210519_081238_92_106c, 20210519_081240_43_106c, 20210519_081241_94_106c, 20210519_081243_46_106c, 20210519_081345_38_2251, 20210519_081347_67_2251, 20210519_083059_1039, 20210519_083338_100a, 20210519_083339_100a, 20210519_083340_100a

Appendix C. PlanetScope Scene Pre-Processing

PS scenes, provided the *Planet*, are measured in units of radiance ($Wm^{-2}sr^{-1}$), representing top-of-atmosphere radiance (TOARad), quantifying the amount of light captured over an area covered by each pixel. However, TOARad measurements are influenced not only by surface reflection qualities but also by illumination conditions. To account for varied lightning conditions due to different acquisition times and locations, TOARad values were converted to top-of-atmosphere reflectance (TOARef), which measures the ratio between reflected and incident radiation. The *Planet* provides the necessary conversion indices and the requested scenes for each spectral channel, enabling TOARad conversion to TOARef following Eq. C.1.

$$\textbf{PS scene}_{\text{ TOARef}} = \begin{bmatrix} i_B \\ i_G \\ i_R \\ i_{NIR} \end{bmatrix} \times \textbf{PS scene}_{\text{ TOARad}} \tag{C.1}$$

where i_B, i_G, i_R, and i_{NIR} are scaling factors for the red, blue, green, and near-infrared spectral channels, respectively, individually for each PS scene.

Appendix D. Spatial Dataset Split

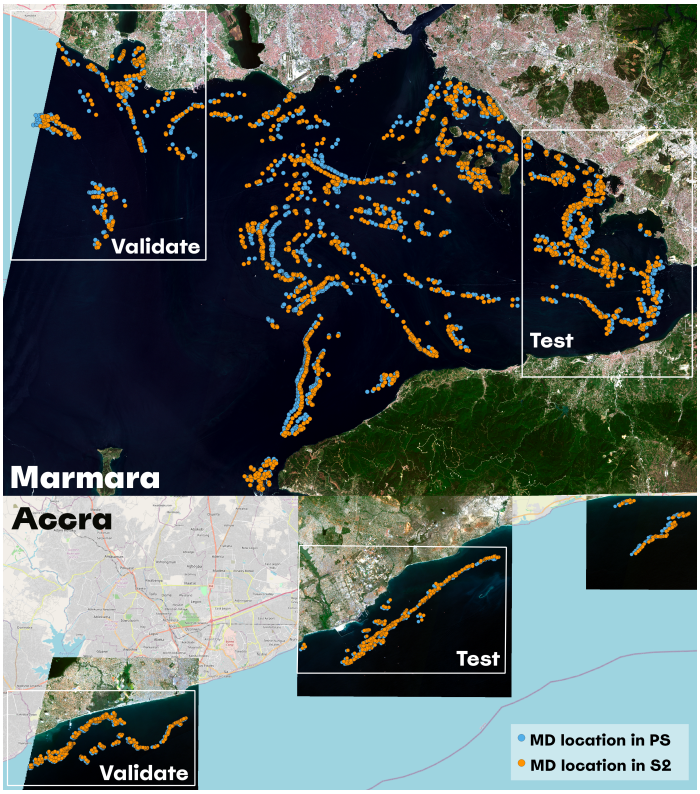

Figure D.1: Spatial data split for Marmara and Accra MD events, showing spatially independent regions used for model training, validation, and testing (remaining).

Appendix E. Evaluation Dataset Compositions

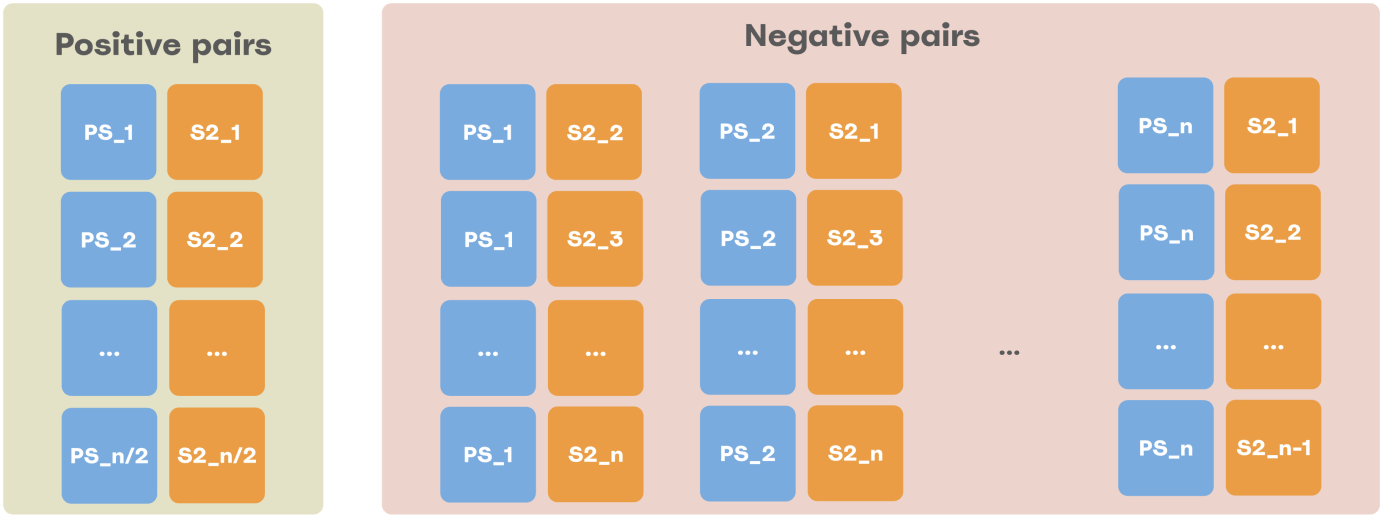

Figure E.1: Composition of the dataset used for retrieval-based evaluation. The dataset comprises all possible, annotated positive pairs (left) and negative pairs (right): each PS tile is paired with all non-matching S2 tiles. For example, PS_1 is paired with $S2_2$ through $S2_n$. This dataset composition allows for global top-k retrieval evaluation.

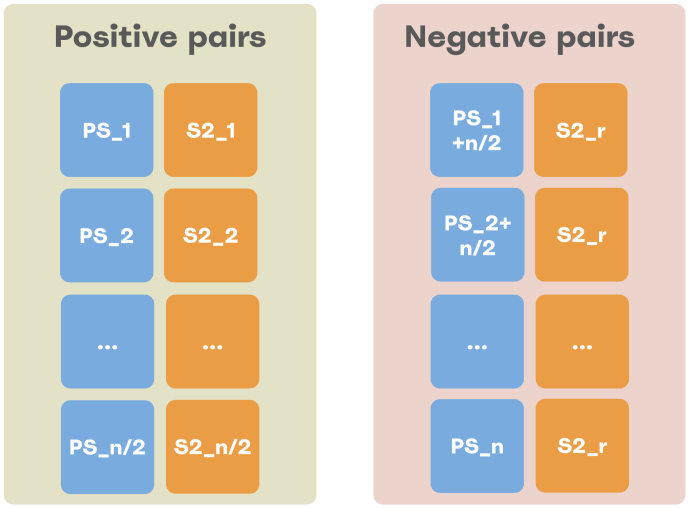

Figure E.2: Composition of the balanced dataset. The dataset contains 50% positive pairs (left), where each PS MD patch tile is paired with its corresponding S2 tile. The remaining 50% are negative pairs (right), formed by pairing the second half of the PS tiles with randomly shuffled S2 tiles, indexed by $r \in [1, n]$. Here, n denotes the total number of tile pairs in the dataset.

Appendix F. Siamese Neural Network Configurations

Three architectural configurations were tested in this study to address the mismatch in input dimensions and implement the Siamese Neural Network (SiamNN) framework for architecture and training.

The first **SiamNN-Single** architectural configuration was composed of simple pre-encoders (Figure 3 and Table F.1) consisting of a convolution layer with `kernel_size=4` and `stride=4`, followed by an Average Pooling layer with `kernel_size=2` and `stride=2` for PS branch, which produced middle-level output with a size 64 by 64 px and 128 channels. To obtain matching dimensions S2 middle-level outputs, an S2 pre-encoder was a single convolution layer with `kernel_size=1`. These outputs were then forwarded through a shared-weight backbone, where the first layer of the shared-weight backbone was adjusted to accommodate 128 input channels, while the last fully connected layer was replaced again with identity mapping, similarly to the proposed Double Acquisition NN ResNet modifications (see section 3.1).

The second **SiamNN-3-block** architecture was identical to the first one, but simple pre-encoders were replaced with 3-layer Convolutional Neural Networks (CNNs) (Table F.1). These CNNs comprised three convolutional layers with batch normalisations and ReLU activations in between. Additionally, the PS CNN pre-encoder had an average pooling layer in the second layer, to reduce its output size and allow standardisation of input dimensionality for both PS and S2 input tiles.

The last **SiamNN-Individual** configuration was identical to the proposed Double Acquisition NN architecture with two branch-specific encoders, but without the projection layers.

Table F.1: Pre-Encoder Architecture Comparisons. 2D convolutional layer (Conv2D), Batch Normalisation layer (BatchNorm), Rectified Linear Unit activation layer (ReLU), Average Pooling layer (AvgPool), input channels (in), output channels (out), kernel size (k), padding (p), stride (s)

Block	**PS Pre-encoder**		**S2 Pre-encoder**	
	Single	**3-layer CNN**	**Single**	**3-layer CNN**
1	Conv2D(in=4, out=128, k=4, s=4) AvgPool(2,2)	Conv2D(in=4, out=32, k=3, p=1) BatchNorm(32) ReLU()	Conv2D(in=13, out=128, k=1, s=1)	Conv2D(in=13, out=32, k=3, p=1) BatchNorm(32) ReLU()
2		Conv2D(32, 64, 3, 1) BatchNorm(64) ReLU() AvgPool(2,2)		Conv2D(32, 64, 3, 1) BatchNorm(64) ReLU()
3		Conv2D(64, 128, 3, 1) BatchNorm(128) ReLU()		Conv2D(64, 128, 3, 1) BatchNorm(128) ReLU()

Appendix G. Deep Relational Similarity Learning Configuration

The Deep Relational Similarity Learning (DRSL) configuration for MD patch matching was modelled after Wang et al. (2021) as a two-branch network (Figure G.1). Both the PS and S2 branches utilised branch-specific ResNet-18 architectures for image encoders. The first convolutional layers were adjusted to intake 4 and 13 spectral channels, respectively, and the final fully connected layers were replaced with identity layers. Each ResNet encoder was followed by a branch-specific Fully Connected Neural Network (FNN), composed of three linear layers with batch normalisations and Rectified Linear Unit (ReLU) activations applied between the layers and after the last layer. Following Wang et al. (2021), the linear layers were mapping output vectors from 512 to 1024, then to 1024, and finally to 300 dimensions.

The 300-dimensional outputs from both branches were concatenated to form a 600-dimensional vector and passed to a Relational NN. This network was composed of 3 linear layers with batch normalisations and ReLU activations between them. The linear layers performed 600-1024-1024-1 mapping. The final mapping was a pairwise relational similarity prediction, which, during the training, was optimised to 0 and 1 labels by employing Mean Squared Error (MSE) loss.

The 300-dimensional outputs of both branches were concatenated to form a 600-dimensional vector, which was passed to a relational neural network (Relational NN). This network consisted of three linear layers with batch normalisation and ReLU activations between them. The layers performed a dimensionality mapping of 600-1024-1024-1.

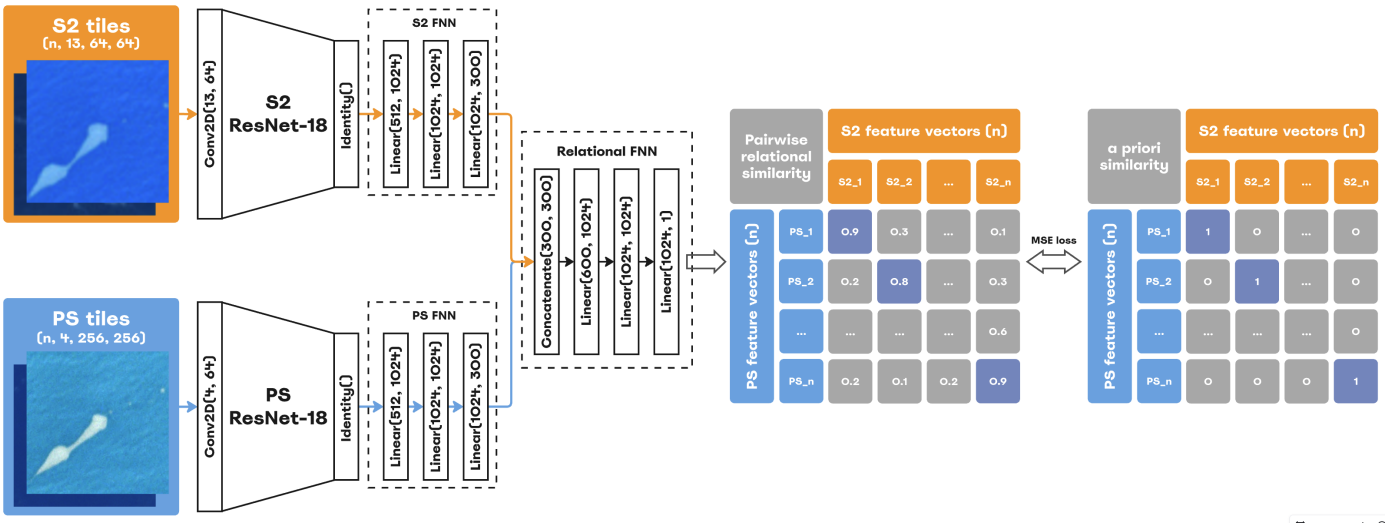

Figure G.1: The implementation of Deep Relational Similarity Learning (DRSL) for the MD patch matching task across PS and S2 double acquisitions. FNN-Fully connected Neural Network.

Appendix H. Augmentation Strategies

This study explored four augmentation strategies: mild, medium, harsh, and spectral channel-shuffling (Figure H.1). Based on visual inspection, the augmentation strategies were adjusted for PS and S2 to produce a similar range of transformations.

- In the **mild** augmentation strategy (Figure H.1 (a)) tiles were randomly rotated by up to ±20°, and scaled by a random factor ranging from 0.5 to 1.3 times. Elastic transformations and Gaussian blur were also applied randomly, along with the addition of Gaussian noise.
- In the **medium** augmentation strategy, the random rotation range was increased to ±45°, and the scaling factor range was extended to 0.35-1.5. Additionally, brightness and contrast adjustments were applied with random factors in the ranges of 0.6–1.4 and 0.5–3, respectively.
- In the **harsh** augmentation strategy, the random rotation range was further expanded to ±90°, and the scaling factor was kept the same as in the medium augmentation strategy. While the brightness and contrast factors remained unchanged from the medium strategy, brightness was adjusted for each spectral channel individually, with a random offset of ±0.2 for S2 and ±0.05 for PS channels.
- In the **spectral-channel-shuffling** augmentation strategy, the only augmentation strategy applied was the random spectral channel re-arrangement.

It is worth noting that directly extracted tiles (referred to as pre-tiles) from PS and S2 scenes had a larger extent of 900 × 900 m (300 × 300 px for PS and 90 × 90 px for S2). Each pre-tile was centred on the MD patch, while the additional inclusive padding allowed for data augmentations that alter tile extent and may otherwise introduce no-data values. Pre-tiles were centre-cropped to the previously mentioned final input sizes for all model training and inference procedures: 256 × 256 pixels for PS and 64 × 64 pixels for S2.

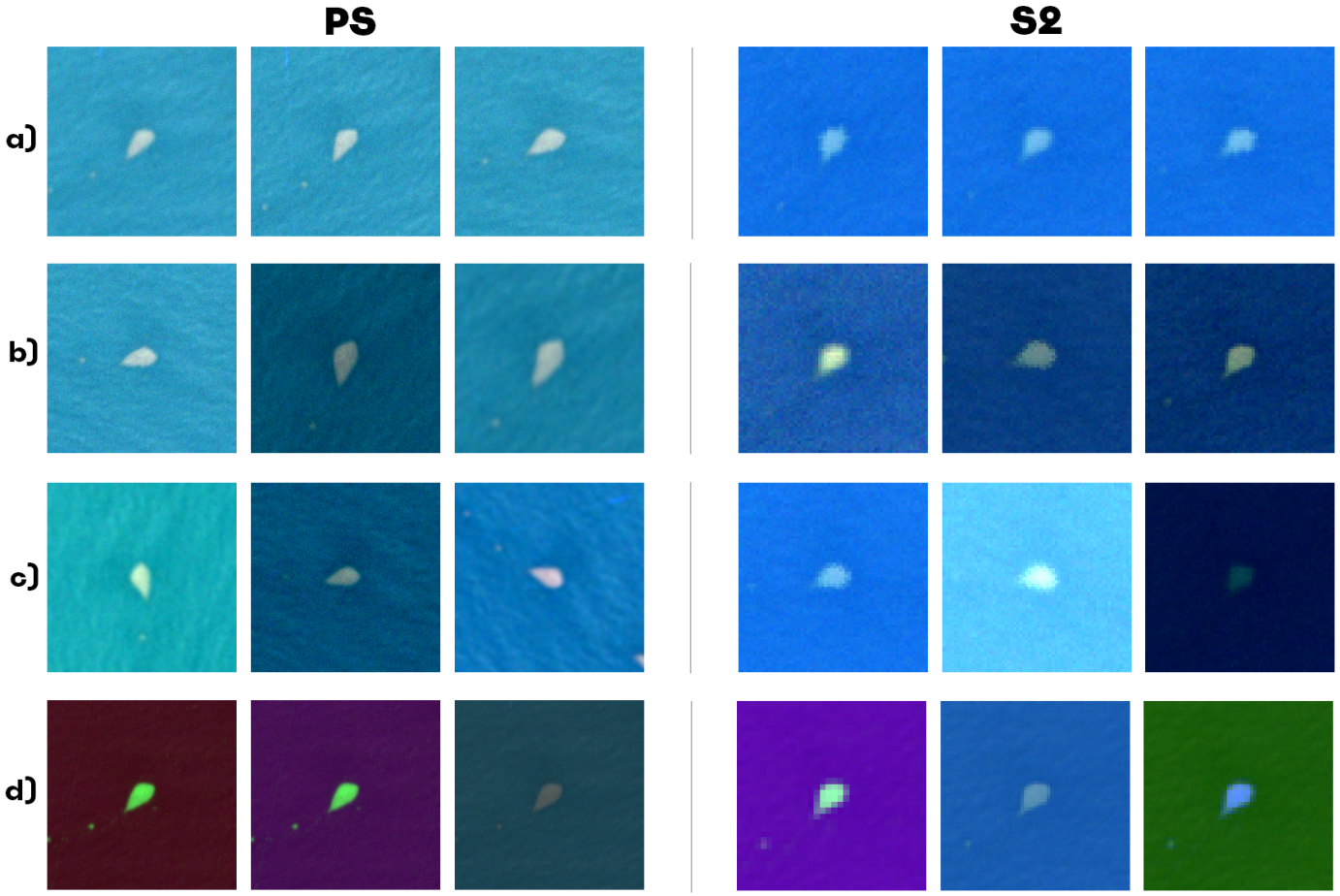

Figure H.1: RGB Examples of PS and S2 tiles with: a) mild, b) medium, c) harsh and d) spectral-channel-shuffling augmentation strategies.