MARs: Multi-view Attention Regularizations for Patch-based Feature Recognition of Space Terrain

Timothy Chase Jr and Karthik Dantu

University at Buffalo
{tbchase,kdantu}@buffalo.edu

Abstract. The visual detection and tracking of surface terrain is required for spacecraft to safely land on or navigate within close proximity to celestial objects. Current approaches rely on template matching with pre-gathered patch-based features, which are expensive to obtain and a limiting factor in perceptual capability. While recent literature has focused on in-situ detection methods to enhance navigation and operational autonomy, robust description is still needed. In this work, we explore metric learning as the lightweight feature description mechanism and find that current solutions fail to address inter-class similarity and multi-view observational geometry. We attribute this to the view-unaware attention mechanism and introduce Multi-view Attention Regularizations (MARs) to constrain the channel and spatial attention across multiple feature views, regularizing the *what* and *where* of attention focus. We thoroughly analyze many modern metric learning losses with and without MARs and demonstrate improved terrain-feature recognition performance by upwards of 85%. We additionally introduce the Luna-1 dataset, consisting of Moon crater landmarks and reference navigation frames from NASA mission data to support future research in this difficult task. Luna-1 and source code are publicly available at https://droneslab.github.io/mars/.

Keywords: Multi-view Metric Learning, Attention Regularization

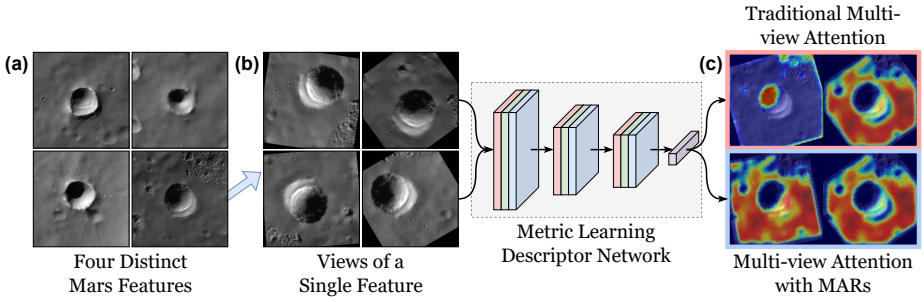


Fig. 1: Patch-based features of space terrain exhibit extreme inter-class similarity and varying multi-view observations, which is difficult for metric learning to discern where attention focus is disparate. We propose Multi-view Attention Regularizations (MARs) to alleviate this issue and drive the attention of arbitrary viewpoints together.

1 Introduction

Exploring deep space objects such as planets, comets, and asteroids involves ambitious and increasingly complex scientific pursuits. It has also been one of the earliest real-world applications of robotic autonomy. Advanced missions strive for spacecraft to land on or maneuver within close proximity to surfaces of highly irregular terrain and varying topography, which poses a significant challenge to spacecraft navigation as communication latency is often too great to permit any Earth-based assistance through radiometric tracking, real-time planning and control, or precise GPS positioning. More recently, these challenges are being addressed through the optical tracking of prominent surface terrain features to provide Terrain Relative Navigation (TRN). This approach has been validated on recent flagship missions including the landing of the Mars Perseverance Rover [56] and the collection of asteroid regolith by OSIRIS-REx [81]. With compute power limited by radiation-tolerant hardware, current approaches to TRN are template matching and correlation techniques using patch-based features (called landmarks) on static navigation maps that are collected and constructed *a priori* [56, 81, 109]. The set of landmarks and the underlying map require extensive pre-navigation costs and effort to obtain and develop. In the case of OSIRIS-REx, an estimated USD 68.5 million (roughly 25% of the nine-year operations budget) was spent performing sufficient reconnaissance to gather and refine this data over a 1.5-year period [80, 97].

To reduce costs and accelerate mission timelines, it would be beneficial to detect and track these landmarks at navigation time similar to Simultaneous Localization and Mapping (SLAM) systems on Earth; a formulation that would also permit generalization to unseen and unexpected scenarios such as planetary weather [46] or asteroid ejection events [65]. SLAM is incredibly challenging to perform during TRN as space environments are generally unstructured, where low lighting and similarity in feature spaces create ambiguity and a lack of re-identifiability [38–40]. The use of learning-based solutions to overcome these challenges is possible with the recent inception of rad-hard inference accelerators [31, 37, 42], which enable in-situ terrain detection methods [14, 30, 66, 69, 90].

Robust description of these detections remains an open problem. On Earth, this is similar to representation learning for tasks such as fine-grained classification, visual place recognition, and person re-identification. At the core of these applications is an objective to learn discriminative image embeddings for efficient similarity computation and downstream retrieval, which is commonly facilitated by metric learning. Compared to Earth-based recognition, however, landmark recognition in space is more nuanced where only one subclass of geological terrain is considered (e.g., *crater*) with possibly thousands of individual instances to discern against (e.g., *crater 570* vs *crater 1181*). This is more fine-grained than even the most challenging of traditional benchmarks (e.g., CUB-200 [104]), demonstrated by Figure 1-a. Apart from individual discernability, the terrain on the surface of planetary bodies, moons, and asteroids can vary widely in appearance from one observation to the next (Figure 1-b), which is difficult for metric learning to reason about on its own (Figure 1-c).

In this work, we examine metric learning as it relates to landmark description during spacecraft TRN. We identify shortcomings in modern metric learning losses and consider poor performance an effect of the view-unaware attention mechanism included in modern architectures. We introduce Multi-view Attention Regularizations (MARs) to bolster recognition accuracy, training network attention to be implicitly view-aware and improving embedding distinguishability. Through additional similarity spaces, we constrain the *what* and *where* of attention information to enforce the consistency of focus between arbitrary feature views. Our approach is extensively validated on Earth, Mars, and Moon landmarks, where we introduce a photo-realistic dataset in the latter case. Experimental results demonstrate the effectiveness of our MARs learning constraint where attention between views is heavily correlated and recognition performance is greatly improved. Overall, we make the following contributions in this paper:

- We study metric learning as the patch-based landmark descriptor for spacecraft navigation and perform extensive studies over traditional methods. We demonstrate shortcomings with terrestrial-based solutions and show correlations between the view-unaware attention mechanism and poor recognition performance in single-shot networks. To the best of our knowledge, this is the first study of its kind.
- We introduce Multi-view Attention Regularizations (MARs), a novel learning constraint to enforce the consistency of channel and spatial attention focus between arbitrary feature views.
- We release a new dataset, Luna-1, consisting of Moon crater landmarks and representative navigation frames using real-world NASA data, facilitating experimentation with multi-view and patch-based recognition systems in space navigation settings.
- We demonstrate the utility of our MARs method, achieving state-of-the-art single-shot landmark description results on Earth, Mars, and Moon environments. Furthermore, we qualitatively showcase improved multi-view attention alignment using MARs.

2 Related Work

Spacecraft Terrain Relative Navigation: Landmarks used for Terrain Relative Navigation (TRN) are collected *a priori* through extensive surveying of the target body and crafted offline by human ground operators. RElative Terrain Imaging NAvigation (RETINA) [109] and Natural Feature Tracking (NFT) [81] are current asteroid-focused TRN methods that create 3D Digital Terrain Models (DTMs) by Stereophotoclinometry (SPC) [35]. Visually prominent areas on the DTM are identified by hand, which are extracted as 2D image templates and uploaded to the spacecraft. Onboard, these templates (i.e., landmarks) are regenerated in SPC fashion to adjust shading based on the predicted illumination conditions of the surface. Navigation frames are then searched for correspondence by traditional image processing algorithms. The Mars Perseverance Landing Vision System (MP-LVS [56]) deployed a similar technique during the landing

phase of the Mars 2020 mission, which hand-picked landmarks on landing site survey imagery captured by other orbiting spacecraft at Mars.

Current TRN solutions include many shortcomings that are a detriment to mission cost, complexity, and time-to-science. The amount of pre-navigation imagery required is immense, and the subsequent time needed to hand-pick which "features-to-track" is extensive. The total number of features used by the system is incredibly sparse due to the level of human involvement, which drastically reduces perceptual capability and prevents reasoning over unseen areas. Onboard rendering of predicted landmark appearances severely limits frame rate, which can jeopardize spacecraft safety during critical phases of the mission. For example, the deployment of NFT on OSIRIS-REx executed at 0.0083 FPS, or one frame every two minutes, as it made contact with the surface [81].

Terrestrial Recognition: The front-end vision in current TRN systems can be radically improved by leveraging rad-hard accelerators and object detection-style observation methods discussed in section 1; although a robust description technique is required to close the loop. Earth-based tasks such as fine-grained visual classification (FGVC), visual place recognition (VPR), and person re-identification (Re-ID) intrinsically demonstrate this capability and reason over similar challenges, including high intra-class and low inter-class variances [5, 20, 73], multi-view observations [7, 48, 98, 121], and appearance change over time [2, 48]. Nevertheless, there are considerable challenges in adopting the current literature. Modern solutions to FGVC, VPR, and Re-ID are focused on description and retrieval problems at internet-scale [17,102,116,118,120] and consequently have become more involved than a single-stage network. These methods employ multiple forward passes [1, 28], region proposals [43, 88, 93, 114, 122], model fusions [41, 74, 84, 91, 92, 99, 113, 119], multi-stage re-rankings [4, 70, 112], and high-parameter transformer models [3, 18, 24, 26, 32, 59, 68, 83, 89, 95, 111]. As such, there is a primary concern about the physical execution of these techniques onboard resource-limited spaceflight computers [10, 36, 107]. Large models that cannot fit within accelerator caches must be executed in a hybrid manner, where model parameters are streamed from the host processor to the accelerator during inference. This has a detrimental effect on execution time [13] and requires careful consideration, given that cache sizes in the current generation of spacecraft accelerators are small (e.g., 8 MB in [42]).

Furthermore, TRN landmark recognition requires more granular reasoning than FGVC, VPR, and Re-ID, akin to frame-to-frame feature matching problems in SLAM. Recognition in FGVC, VPR, and Re-ID is performed by recalling instances from a pre-seeded database by global description [2, 7], where any viewpoint and domain generalization is generally a byproduct of learning with extremely large datasets [2,6,34] or the aggregation of large datasets [59]; a technique that is not currently adoptable due to the lack of space landmark datasets (two at the time of writing including the proposed Luna-1). Additionally, the *sequential* nature of the TRN task needs consideration, where any recognition database is populated as samples are encountered instead of recalling against the entire population upfront (the effects of which have not been studied previously).

Futhermore, the appearance differences between instances of a geological space-terrain feature (e.g., crater) are generally more subtle than traditional FGVC and Re-ID datasets making discernability more challenging [64, 104, 117].

We identify metric learning as the core facilitator of discriminative representation learning in terrestrial tasks and examine its capabilities to permit lightweight, onboard-executable single-shot description networks for spacecraft TRN. Observing terrain features from a remote-sensing platform exhibits complex transformation spaces, however, which must be taken into consideration.

Viewpoint Challenges and Attention: During TRN the observed target body is rotating and revolving distinctly from the spacecraft leading to an unconstrained appearance change in landmark illumination, translation, and rotation over time. Such a transformation space is generally uncommon in the literature (Re-ID would not expect a person observation to be upside-down for example [117]), and modern metric learning losses do not permit invariancy to these transformations directly. The convolutional layers used in modern networks are known to be equivariant to translations over the input image [22, 47], but are not naturally equivariant to rotations. Explicit in-network modifications for adding rotation equivariance have recently been explored including steerable filters [22, 103], multi-orientation feature extractions [27, 67], and alternative coordinate systems [33, 51, 55, 75]. Equivariant properties can be additionally learned [11, 77], which may be advantageous as a supplement to explicit mechanisms or when explicit mechanisms are themselves undesirable [87]. Learning equivariance is popular in the literature through batch-sampling, mining, and augmentation approaches [12, 16, 44, 71, 105, 115]. The remote sensing literature has studied similar techniques with the fusion of pre-trained group convolutions [21] and probabilistic formulations of metric space locations [58], although they are restrictive in their reasoning through trainings with pre-rotated data.

The explicit encoding of equivariant properties into the attention mechanism has recently been explored [9, 15, 54]. At large, however, analyzing *learned* attention equivariance as it compares to these mechanisms (or the combination thereof) has not been studied previously. The Self-supervised Equivariant Attention Mechanism (SEAM) [101] is one of the only works that target attention-equivariance learning directly through self-supervised regularization. Multi-view attention similarity learning such as the Contrastive Attention Map Loss (CAML) [62] has shown impressive equivariant properties as a byproduct of contrastive learning over attention maps. Although, integration of these methods within metric learning frameworks is a challenging task as SEAM requires Class Activation Maps (CAMs) and CAML targets foreground/background feature separation using image statistics from segmentation labels.

3 Methodology

Prior work demonstrates that attention has a large influence on recognition performance in multi-view settings, but the extent of this influence concerning equivariant properties (either encoded or learned) is unclear. Equivariance

does not guarantee that attention, being a strictly learnable mechanism, will be identical between multiple views of the same feature; it only *suggests* that it should be similar. An alignment of attention focus should lessen the downstream recognition difficulty, maximizing separability and view-dependent groupings in the embedding space, although such a constraint is not readily formulated in current multi-view metric learning pipelines. We suggest that any attention disagreement must be directly accounted for during the training, and propose a soft learning constraint to rectify any variance. This concept forms the basis of our proposed Multi-view Attention Regularizations (MARs), described in this section. We first introduce our learning framework and baseline network architecture in subsection 3.1 and subsection 3.2. We then detail our constraint for aligning attention and frame the overall learning objective in subsection 3.3.

3.1 Learning Framework

The framework for data augmentation and batch formation plays a critical role in multi-view similarity learning [12, 16, 44, 115], where we start by following the popular SimCLR [16] method. SimCLR aims to maximize the learned representation similarity between augmented views of the same input. With training batch size B, we begin by sampling a minibatch of $B/2$ samples where each sample x gets augmented by two distinct transformation operations to produce new views $x_1 = t_1(x)$ and $x_2 = t_2(x)$ where t_1 and t_2 are sampled from the same family of augmentations $\mathcal{T}$. $\mathcal{T}$ is a composition function of three image transformations that include a random brightness adjustment, rotation, and translation. An encoder network $f(\cdot)$ is applied to the augmented data to extract intermediate representations $h_1 = f(x_1)$ and $h_2 = f(x_2)$. These representations are in turn mapped to the metric space through projection head $g(\cdot)$ to yield embeddings $z_1 = g(h_1)$ and $z_2 = g(h_2)$. Given this (z_1, z_2) positive pair, the other $2(B/2 - 1)$ embeddings in the minibatch are considered negative samples. The batch of z embeddings is fed to any applicable metric learning loss $\mathcal{L}_{ML}$, as is the traditional metric learning process. To assist with the inter-class granularity of landmark recognition we additionally employ hard sample mining in traditional multi-similarity (MS) [100] fashion to yield ap, p, an, n batch-indices where ap, p represent anchor-positives and positives (simply the indices of the twice augmented images) and an, n the indices of embeddings deemed similar by the MS metric but have different instance labels.

3.2 Network Architectures

With inspiration from large-scale Earth-based recognition networks [50, 60, 106] we employ a ResNeXt-101 [110] architecture with Squeeze-and-Excitation (SE) [53] attention as the baseline for encoder network $f(\cdot)$. Encoder $f(\cdot)$ is the primary bottleneck to onboard execution performance as it will hold the most parameters, and we select ResNeXt-101 as a middle ground between discriminative representation power and model size. Furthermore, we elect to stay on the larger

end of model size in contrast to the ResNet-50 class [49] to isolate representation power and examine the effects of different attention and equivariance setups. Our embedding projection head $g(\cdot)$ is a smaller network consisting of a Generalized Mean Pooling (GeM) [85] layer followed by a linear (512), batch norm, and PReLU activation. In contrast to Earth-based recognition, we perform no functions other than a single shot $f(\cdot)$ and $g(\cdot)$ to descript instances.

Encoding Rotational Equivariance: Augmentations applied in $\mathcal{T}$ mimic the unconstrained landmark appearance change found in spacecraft TRN (assuming the spacecraft is in a non-geosynchronous position relative to the target body). As the pose of the target body will be changing independently of the spacecraft we cannot assume rotated landmark views will be limited to anything less than a full 360 degree of change. Although data augmentation attempts to implicitly teach the network to be robust, reasoning over this level of extreme rotation remains a challenging property to learn. As such, we additionally seek to study the benefits of explicit rotational equivariance integration in $f(\cdot)$.

RIC-CNN [75] develops a convolutional operation (the Rotation-Invariant Coordinate Convolution, RIC-C) based on a novel coordinate system that permits this equivariance as a replacement to standard convolutional layers. RIC-C extends the idea of deformable convolutions [23] and does not require any transformation of the representation space of input images or intermediate features. This property enacts a simple and efficient implementation, which we leverage in this work by replacing all standard convolution operations in $f(\cdot)$ with RIC-C layers. For brevity, we refer interested readers to [75] for the full account of the coordinate system and RIC-C operation.

Spatial Attention: SE attention improves the interdependencies within feature maps by assigning weights to each channel and selecting the most relevant for a given input. This type of attention is focused on relevancy *between* features alone (channel) and carries no understanding of relevancy *within* an individual feature (spatial). We assume spatial attention has a critical role in multi-view metric learning and introduce this in $f(\cdot)$. Coordinate Attention (CA [52]) provides spatial awareness through distinctive pooling operations in the height and width dimensions while preserving the channel dimensionality. This is in contrast to other spatial attention techniques such as the Convolutional Block Attention Module (CBAM [108]) that collapse channel information via pooling before learning spatial weight factors. We modify $f(\cdot)$ by replacing SE attention with CA.

3.3 Forming Attention Similarity Constraints

The inclusion of explicit rotational equivariant properties through RIC-C layers and the ability to attend spatially with CA is the basis for which we explore our proposed MARs constraint. During the training procedure, we seek to drive both the *what* (channel) and the *where* (spatial) elements focused by the attention mechanism together, without explicitly assuming that one view is correct in either of these aspects. This alignment is thus a moving target, where it is imperative to impose a soft constraint between them. In other words, it is undesirable

to calculate a strict differentiation between attention maps at any point during training to avoid a collapse in attention information. The constraint should prioritize that the attention maps from each view evolve similarly over time.

Pose Normalization and Channel Reduction: To facilitate this constrained evolution we propose to introduce regularization terms by embedding attention into additional metric spaces. Let A_i be the set of attention maps output from ResNeXt block $i \in N$ from $f(\cdot)$ where N is the number of these blocks. For each positive pair in the training batch, we have multi-view attention maps A_{i1} and A_{i2}. For each ResNeXt block outputting A_i, we output inverse transformation t_i^{-1} where the translation parameters are adjusted relative to the spatial resolution of A_i. We apply the inverse transformation to normalize the translation and orientation (pose) of each attention map to equal that of the input image, yielding pose-normalized attention maps $\hat{A}_{i1} = t_{i1}^{-1}(A_{i1})$ and $\hat{A}_{i2} = t_{i2}^{-1}(A_{i2})$. To embed attention into additional metric spaces we employ mini variants of the projection head $g(\cdot)$, which do not include any linear layers for dimensionality reduction. Instead, we first reduce the channel dimension of $\hat{A}_i \in \mathbb{R}^{C\times H\times W}$ through a 1x1 convolution $\text{Conv}_i^1(\cdot)$ with reduction factor r to yield $\hat{A}_i^r \in \mathbb{R}^{C/r\times H\times W}$. This process prevents obscurification, keeps the data correlated, and reduces learnable parameters.

Channel and Spatial Attention Embeddings: For positive and pose-normalized attention pairs $(\hat{A}_{i1}^r, \hat{A}_{i2}^r)$ we utilize the mini channel-wise (c) projection head $gc_i(\cdot)$ to produce channel attention embeddings $zc_{i1} = gc_i(\hat{A}_{i1}^r)$ and $zc_{i2} = gc_i(\hat{A}_{i2}^r)$. GeM pooling collapses the spatial dimensions to yield an embedding with length given by C/r where C is the channel dimension of the current ResNeXt block i. For spatial attention embeddings, we first perform height and width pooling (similar to CA) on $\hat{A}_i^r$. Specifically, given height (y) and width (x) pooling operators $\text{Ypool}(\cdot)$ and $\text{Xpool}(\cdot)$ we produce intermediate representations $hy_i = \text{Ypool}(\hat{A}_i^r)$ and $hx_i = \text{Xpool}(\hat{A}_i^r)$. These representations are input to mini spatial projection heads $gy_i(\cdot)$ and $gx_i(\cdot)$ to yield height and width embeddings $zy_i = gy_i(hy_i)$ and $zx_i = gx_i(hx_i)$. The mini projection heads $gc_i(\cdot)$, $gy_i(\cdot)$, and $gx_i(\cdot)$ do not share any parameters and are instantiated once per block $i \in N$. This allows distinct regularization on attention maps with the same channel-spatial resolution as well as calculating accurate batch-norm statistics that are channel, spatial-height, and spatial-width disparate.

Multi-view Attention Regularizations (MARs): Once embedded, we regulate the channel and spatial attention focus using a cosine similarity loss:

$$\mathcal{L}_{\text{cs}}(z_1, z_2) = 1 - \frac{z_1 \cdot z_2}{\|z_1\|_2 \cdot \|z_2\|_2} \tag{1}$$

given embeddings z_1 and z_2. We define a channel-wise MARs ($\mathcal{L}_{\text{ChMARs}}$) as the cosine similarity between channel attention embeddings:

$$\mathcal{L}_{\text{ChMARs}}(\hat{A}_{i1}^r, \hat{A}_{i2}^r) = \mathcal{L}_{\text{cs}}(zc_{i1}, zc_{i2}) \tag{2}$$

given positive pair, pose-normalized and dimensionality reduced attention maps $\hat{A}_{i1}^r$ and $\hat{A}_{i2}^r$. Likewise, we define a spatial-wise MARs ($\mathcal{L}_{\text{SpMARs}}$) as the cosine

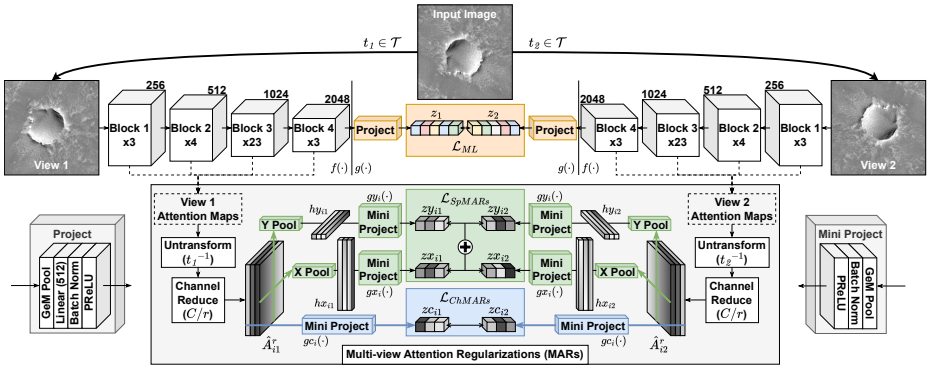


Fig. 2: Framework of the proposed Multi-view Attention Regularizations (MARs). MARs aligns the *what* (channel) and *where* (spatial) focus of attention between multiple patch-feature views using distinct metric spaces.

similarity between Y-pooled and X-pooled attention embeddings:

$$\mathcal{L}_{\text{SpMARs}}(\hat{A}^r_{i1}, \hat{A}^r_{i2}) = \mathcal{L}_{\text{cs}}(zy_{i1}, zy_{i2}) + \mathcal{L}_{\text{cs}}(zx_{i1}, zx_{i2}) \tag{3}$$

and our combined MARs regularization loss by:

$$\mathcal{L}_{\text{MARs}}(\hat{A}^r_{i1}, \hat{A}^r_{i2}) = \gamma_{Ch}\mathcal{L}_{\text{ChMARs}}(\hat{A}^r_{i1}, \hat{A}^r_{i2}) + \gamma_{Sp}\mathcal{L}_{\text{SpMARs}}(\hat{A}^r_{i1}, \hat{A}^r_{i2}) \tag{4}$$

where γ_{Ch} and γ_{Sp} are weight parameters that control the influence of channel and spatial attention alignment respectively. With augmented image batch X and mined indices (ap, p, an, n) our complete learning objective is given as:

$$\mathcal{L}(X, (ap, p, an, n)) = \mathcal{L}_{\text{ML}}(g(f(X)), (ap, p, an, n)) + \sum_{i=1}^{N \in f(\cdot)} \mathcal{L}_{\text{MARs}}(\hat{A}^r_{ap}, \hat{A}^r_{p}) \tag{5}$$

where

$$\hat{A}^r_{ap} = \text{Conv}^1_i(t^{-1}_{ap}(f_i(X_{ap}))), \quad \hat{A}^r_{p} = \text{Conv}^1_i(t^{-1}_{p}(f_i(X_{p}))) \tag{6}$$

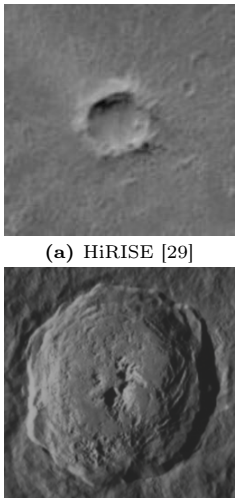
with X_{ap} and X_p the anchor-positive and positive pair images and $f_i(\cdot)$ the i'th block in $f(\cdot)$ that outputs attention maps A_i. Our end-to-end pipeline with MARs regularization is shown in Figure 2.

4 Evaluation

We wish to study lightweight single-shot TRN landmark description using modern metric learning both with and without MARs as well as the effect of different attention and equivariant mechanisms. We first discuss the datasets used for experimentation in this section, followed by a description of our experiments, implementation details, and analysis of the results.

4.1 Datasets

We leverage three datasets of Mars, Moon, and Earth landmark images. HiRISE [29] contains 700 Mars crater images and is the only real-world dataset available at the time of writing. We further introduce Luna-1, a 5,067 sample Moon crater dataset generated in the Blender 3D software [8] with real-world NASA data products. Luna-1 additionally contains 2,161 emulated navigation frames from a Lunar Reconnaissance Orbiter (LRO) three-orbit reference navigation sequence. An example landmark image from HiRISE and Luna-1 is shown in Figure 3. Additional Luna-1 details and visualizations can be found in the supplementary. For Earth landmarks, we utilize the stadium class from RESISC45 [19], a terrestrial remote sensing scene classification dataset with 700 samples. We refer to HiRISE, Luna-1, and RESISC45 as *Mars Crater*, *Moon Crater*, and *Earth Stadium* respectively. For all datasets, we partition two instance-distinct groups for training and testing such that each group contains half of the available images (as is standard in the literature). In the case of Luna-1, we ensure all craters seen during the navigation sequence are added to the test set before this partitioning.



(a) HiRISE [29]

(b) Luna-1

Fig. 3: Mars (a), Moon (b) landmark examples.

4.2 Experiments

We perform two experiments that emulate landmark recognition behavior during TRN, including a sequential, incremental recall experiment (*Incremental Recall@1*) and an object detection-style description experiment on the navigation frames from Luna-1 (*Moon Navigation*). Additionally, we perform a traditional *Recall@1* (gallery size one) as well as a Luna-1 relocalization experiment (*Moon Lost-in-Space*). Details of these experiments are provided below. Additional results including model execution times on spacecraft hardware and MARs training curves can be viewed in the supplementary.

Incremental Recall@1: The embedding database starts empty and test-partition landmarks are randomly selected. Each landmark is augmented by a transform sampled from $\mathcal{T}$. Embeddings are generated single-shot from the model and the database is searched for correspondence. Embeddings are stored in the database if no match is found. We compute Recognition Accuracy (RA) as the percentage of correct matches relative to the total number of landmark matches (either correct, incorrect, or missed). Missed matches are landmarks that were added to the database more than once (i.e., duplicate embeddings). The RA formulation is given in Equation 7. To provide multiple observations we repeat each landmark in the test partition twice.

$$RA = \left(\frac{\text{Correct Matches}}{\text{Correct Matches} + \text{Incorrect Matches} + \text{Missed Matches}}\right) * 100 \quad (7)$$

Moon Navigation: Luna-1 navigation frames are iterated sequentially, where each frame comes paired with ground-truth bounding box annotations of visible craters. For each frame, we first perform non-maximum suppression (NMS) to emulate the use of an object detector (akin YOLO [86]). Landmarks are given by cropping the resulting set of bounding boxes which are in turn augmented by a random transform sampled from $\mathcal{T}$. Embeddings are generated single-shot by the model and RA performance is measured identically to the Incremental Recall@1 experiment. Models trained on Mars Crater are not considered in this experiment due to the domain shift between Mars and Moon. However, one may expect a level of feature generality on crater landmarks from any environment and we report such results in the supplementary.

Moon Lost-in-Space: This experiment emulates the kidnapped robot problem in traditional robotics literature. The embedding database is first seeded with all crater landmarks seen during the first orbit of the Luna-1 navigation, where landmarks are detected and augmented identically to the Moon Navigation experiment. Frames from the last orbit are then randomly selected and the RA is reported by matching computed embeddings to those in the database. The database is not updated throughout the experiment outside of the initial seed. Similar to Moon Navigation we only consider models trained on Moon Crater data here, and report a Mars Crater training study in the supplementary. Furthermore, an ablation study over singular transformation types in the family $\mathcal{T}$ for this experiment as well as Moon Navigation is given in the supplementary.

4.3 Implementation Details

To determine the effectiveness of metric learning for robust landmark description, it is imperative to understand two primary conditions for TRN: (i) recognition over time with many similar terrain features encountered sequentially, and (ii) the unique transformation space in remote sensing. Therefore, we frame this study as a measure of modern metric learning invariancy and discriminative properties under these conditions and elect not to compare against fully-fledged Earth-based systems that are unsuitable for onboard spaceflight. Additionally, we seek to understand the influence of MARs on various metric learning losses and the effect of different attention and equivariant setups.

Table 1: Evaluated variants of the baseline model. *conv2d:* PyTorch convolution. *RIC:* Rotation Invariant Convolution [75]. *SE:* Squeeze-Excitation attention [53]. *CA:* Coordinate Attention [52].

Name	Conv.	Att.	Loss
conv2d SE	conv2d	SE	$\mathcal{L}_{\text{ML}}$ Only
RIC CA	RIC	CA	$\mathcal{L}_{\text{ML}}$ Only
MARs	RIC	CA	$\mathcal{L}_{\text{MARs}}$

We evaluate three variants of the baseline model which are described in Table 1. Each model effectively adds a level of equivariance (and in theory, robustness to challenging multi-view appearance change) from the last; i.e., *conv2d SE* learned equivariance only, *RIC CA* learned and explicit equivariance, *MARs* learned and explicit equivariance with attention constraints. We study the effects of each model across nine discriminative learning losses ($\mathcal{L}_{\text{ML}}$) found recent in the literature, including Circle Loss [94], Direction-Regularized Multi-Similarity

(DR-MS) [76], NTXent [82], PNP [72], Proxy Anchor [63], ProxyNCA++ [96], Subcenter ArcFace [25], Supervised Contrastive (SupCon) [61], and Proxy Synthesis [45]. We train all models for 150 epochs using a batch size of 32 on Earth Stadium and Mars Crater and 128 on Moon Crater.

We use the PyTorch Metric Learning (PML) [79] library for MS miner implementation as well as all $\mathcal{L}_{\text{ML}}$ losses except ProxyNCA++ and Proxy Synthesis, in which we use the paper implementations. The Faiss [57] library is used as the embedding database in all experiments. We use cosine similarity for database comparison and define a matching threshold of 0.9. If multiple embeddings are retrieved above this threshold we consider the largest one a match. In Moon Navigation an NMS threshold of 0.5 is used based on the YOLO default. The p parameter in GeM layers is learnable with an initial value of 3. The reduction factor r is set to 4. All $\mathcal{L}_{\text{MARs}}$ models have γ_{Ch} and γ_{Sp} parameters set to 0.15.

4.4 Experimental Results

Table 2: Recall@1 (gallery size one) and Incremental Recall@1 recognition accuracy for all models and $\mathcal{L}_{\text{ML}}$ losses. Bold values signify the highest performing model for each $\mathcal{L}_{\text{ML}}$, while underlined values show the best model/$\mathcal{L}_{\text{ML}}$ variant on each dataset.

$\mathcal{L}_{\text{ML}}$	**Recall@1**									**Incremental Recall@1**								
	Earth Stadium			Mars Crater			Moon Crater			Earth Stadium			Mars Crater			Moon Crater		
	conv2d SE	RIC CA	MARs (Ours)	conv2d SE	RIC CA	MARs (Ours)	conv2d SE	RIC CA	MARs (Ours)	conv2d SE	RIC CA	MARs (Ours)	conv2d SE	RIC CA	MARs (Ours)	conv2d SE	RIC CA	MARs (Ours)
Circle [94]	86.29	**93.71**	90.29	59.71	**92.29**	80.57	**98.38**	97.75	96.93	4.79	5.11	**5.13**	3.59	**60.81**	12.56	43.36	**60.09**	30.46
DR-MS [76]	86.57	88.86	**89.71**	66.57	**84.29**	62.57	**96.14**	85.97	96.02	4.33	5.27	**54.32**	4.04	**48.48**	4.04	62.55	2.47	**64.57**
NTXent [82]	**95.43**	91.71	94.29	66.57	83.43	**91.71**	98.03	98.98	**99.57**	**12.23**	7.75	8.59	4.49	**34.10**	34.01	69.85	77.06	**81.69**
PNP [72]	80.57	79.43	**85.14**	43.71	**80.86**	68.00	75.18	88.65	**94.09**	4.64	**5.27**	4.66	3.14	**16.70**	5.14	10.24	15.54	**40.84**
Proxy Anchor [63]	90.57	99.71	**100.00**	56.86	84.86	**98.57**	99.96	–	**100.00**	4.48	72.86	**78.06**	3.44	10.45	**71.10**	94.56	–	**94.78**
ProxyNCA++ [96]	95.71	96.29	**99.71**	63.14	**99.43**	78.57	99.65	**98.78**	95.47	4.33	4.64	**12.00**	4.04	**7.71**	4.79	56.03	**71.23**	70.27
Subcenter ArcFace [25]	77.71	**87.71**	83.43	40.86	62.29	**79.14**	–	–	**94.05**	**4.49**	4.48	4.19	3.29	4.99	**38.69**	–	–	**20.45**
SupCon [61]	76.57	92.00	**95.71**	74.29	**94.86**	91.43	97.08	**99.21**	98.98	4.19	5.80	**46.43**	4.65	**57.52**	49.19	16.73	79.42	**84.11**
Proxy Synthesis [45]	94.86	95.14	**99.71**	76.86	74.57	**99.14**	99.68	**99.96**	99.84	4.33	**39.41**	22.57	4.03	4.97	**35.14**	**91.40**	64.71	17.47

Recall@1: Table 2 (left) displays results for Recall@1. Firstly, including explicit rotation equivariance and spatial attention (RIC CA) leads to improvements on many $\mathcal{L}_{\text{ML}}$, suggesting that learning transformation robustness alone is not enough and a combination of learned and explicit equivariance is necessary. On Mars Crater data, MARs leads to substantial improvements on certain $\mathcal{L}_{\text{ML}}$ such as NTXent, Proxy Anchor, Subcenter ArcFace, and Proxy Synthesis which were improved by roughly 10%, 16%, 27%, and 33% respectively compared to RIC CA. This is evidence that attention similarity heavily influences $f(\cdot)$ feature selection and results in more discriminative embeddings. Conv2d SE and RIC CA variants for Subcenter ArcFace on Moon Crater see a failure to converge during training while MARs variants do not. This is significant as it demonstrates a boost in representation power that can enable $\mathcal{L}_{\text{ML}}$ losses that would otherwise fail. Overall, a MARs model variant is best-in-class for Earth Stadium and Moon Carter, and competitive (<1% difference) on Mars Crater.

Benefits are not guaranteed as we observe lower accuracy than RIC CA with MARs for certain $\mathcal{L}_{\text{ML}}$ such as DR-MS on Mars Crater, which sees a performance decrease of roughly 26%. This reveals a correlation between attention alignment and embedding separability that is $\mathcal{L}_{\text{ML}}$ specific. Knowing how attention information ultimately presents itself in the embedding projected by $g(\cdot)$ is

not obvious, which may indicate incompatible $\mathcal{L}_{\text{ML}}$ where attention information is ultimately obscured and not readily distinguishable in the $\mathcal{L}_{\text{ML}}$ space.

Incremental Recall@1: Recognition accuracy for the Incremental Recall@1 experiment is shown in Table 2 (right). We see very low accuracy with conv2d SE on Earth Stadium and Mars Crater, which signals poor representation power and an indiscernible metric space on these smaller datasets. Encoding rotational equivariance in RIC CA helps alleviate this issue in cases such as Proxy Anchor on Earth Stadium and many $\mathcal{L}_{\text{ML}}$ on Mars Crater. MARs has a profound impact in cases where RIC CA offers little to no benefit, such as Proxy Anchor on Mars Crater where we see a roughly 85% improvement over RIC CA. On Moon Crater, MARs attention constraints improve $6/9$ $\mathcal{L}_{\text{ML}}$ losses and is competitive on Subcenter ArcFace ($< 1\%$ difference), indicating benefits with more training data. Furthermore, we see a similar pattern of behavior to the Recall@1 experiment where the performance of RIC CA and MARs varies wildly across $\mathcal{L}_{\text{ML}}$ losses, as shown by Proxy Synthesis on Mars/Moon Crater. Overall, Proxy Anchor with MARs is best-in-class on every dataset for this experiment.

Moon Navigation and Lost-in-Space: Table 3 displays results for Moon Navigation (left) and Moon Lost-in-Space (right). For Moon Navigation, conv2d SE retains maximum performance on $4/9$ $\mathcal{L}_{\text{ML}}$ losses while RIC CA achieves the highest accuracy only on SupCon loss, supporting the theory that fully explicit equivariance may be undesirable in cases with more training data, and a partial (i.e., learned) equivariance is better suited due to the sufficiency of the learning framework to distinguish low-level features under transformation [87]. Nevertheless, MARs is quite competitive in instances where RIC CA has worse performance than conv2d SE showing that the additional *learned* multi-view attention constraint counteracts the negative effects of the explicit properties. In total, MARs achieves the highest performance on $4/9$ $\mathcal{L}_{\text{ML}}$ and obtains best-in-class with Proxy Anchor. Accuracy is relatively high for all methods on Moon Lost-in-Space, which could be an artifact of the high-framerate navigation sequence that fills the embedding database with many duplicate (although augmented) craters throughout the first orbit. MARs demonstrates performance increases on $4/9$ $\mathcal{L}_{\text{ML}}$ losses once again with only NTXent and Subcenter ArcFace being common among both experiments.

Table 3: Accuracy for Moon Navigation (left) and Moon Lost-in-Space (right).

$\mathcal{L}_{\text{ML}}$	**Moon Navigation**			**Moon Lost-in-Space**		
	conv2d SE	RIC CA	MARs (Ours)	conv2d SE	RIC CA	MARs (Ours)
Circle [94]	**58.07**	37.97	38.46	94.03	**96.68**	92.31
DR-MS [76]	**37.69**	3.12	36.68	86.34	88.06	**90.05**
NTXent [82]	48.25	32.00	**57.68**	94.83	83.16	**96.29**
PNP [72]	14.34	23.34	**24.66**	61.41	**77.98**	75.46
Proxy Anchor [63]	64.17	–	**66.31**	**97.21**	–	96.02
ProxyNCA++ [96]	**58.27**	53.92	35.87	**94.69**	93.24	91.38
Subcenter ArcFace [25]	–	–	**40.63**	–	–	**81.17**
SupCon [61]	17.92	**42.28**	37.50	89.39	90.32	**90.58**
Proxy Synthesis [45]	**61.26**	60.53	32.67	**96.42**	93.77	36.87

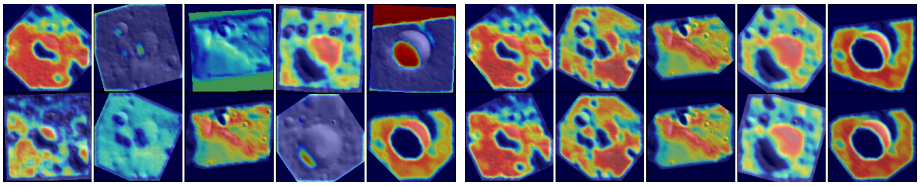
Table 4: MARs γ ablation study.

γ_{Ch}	γ_{Sp}	Recall@1	Incremental Recall@1	Moon Navigation	Moon Lost-in-Space
0.0	0.3	**100**	59.429	**48.204**	89.594
0.15	0.15	98.571	**71.105**	38.301	89.335
0.3	0.0	69.43	4.18	1.93	62.62
1	1	96.571	61.517	23.593	**89.733**

Ablation Study, γ Parameter: Table 4 gives an ablation study over the γ parameters in MARs. We train four combinations of γ_{Ch} and γ_{Sp} using Proxy

Anchor $\mathcal{L}_{\text{ML}}$ on Mars Crater, where we can see a clear sensitivity. Spatial attention has the biggest impact where $\gamma_{Sp} = 0$ reduces accuracy in all experiments, indicating the importance of spatial attention alignment in multi-view learning. The conservative 0.15 for both γ_{Ch} and γ_{Sp} gives utility to the channel component, as we see the best results on Recall@1 and Incremental Recall@1 accuracy (and only slightly less accuracy on Moon Lost-in-Space). A perhaps surprising result is the difference in performance (or lack thereof) between low parameter settings (0.15) and unweighted settings ($\gamma_{Ch} = \gamma_{Sp} = 1$), implying that the magnitude of $\mathcal{L}_{\text{MARs}}$ has little effect on the optimization.

Qualitative Analysis: Figure 4 visualizes attention focus via the EigenCAM [78] algorithm, comparing RIC CA and MARs ($\gamma_{Sp} = 0.3, \gamma_{Ch} = 0$) for Proxy Anchor on Mars Crater. MARs shows a near-identical magnitude EigenCAM between each positive pair of pose-normalized landmarks where RIC CA focuses on disparate regions. Additional examples and animations of attention evolution during the training can be viewed in the supplementary.



(a) RIC CA **(b)** MARs

Fig. 4: Attention visualizations with EigenCAM [78] on Mars Crater trained with Proxy Anchor $\mathcal{L}_{\text{ML}}$.

5 Conclusion

The utility of metric learning as a single-shot landmark description technique for spacecraft TRN was thoroughly explored in this work. We demonstrated that metric learning alone cannot adequately descript fine-grained instances of celestial terrain given multi-view observations and complex transformation spaces. We show that traditional workarounds such as equivariant convolutional layers are in many cases still insufficient. We identify shortcomings with the view-unaware attention mechanism and proposed Multi-view Attention Regularizations (MARs) to regulate attention focus between views. MARs enacts a soft learning constraint that prevents attention collapse, effectively driving the *what* and *where* elements of attention together and eases the downstream separability task. We demonstrated the utility of our method through rigorous and comprehensive experimentation, where we showed regular improvements to a wide range of metric learning losses by upwards of 85% on navigation-style tasks. We additionally introduced the Luna-1 dataset to facilitate more active research in TRN landmark recognition, consisting of photo-realistic Moon crater landmarks and paired navigation images using real-world NASA data.

References

1. Ali, S., Sullivan, J., Maki, A., Carlsson, S.: A baseline for visual instance retrieval with deep convolutional networks. In: Proceedings of International Conference on Learning Representations (2015)
2. Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Mixvpr: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2998–3007 (2023)
3. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval. arXiv preprint arXiv:2304.05884 (2023)
4. Bai, S., Bai, X.: Sparse contextual activation for efficient visual re-ranking. IEEE Transactions on Image Processing **25**(3), 1056–1069 (2016)
5. Bera, A., Wharton, Z., Liu, Y., Bessis, N., Behera, A.: Sr-gnn: Spatial relation-aware graph neural network for fine-grained image categorization. IEEE Transactions on Image Processing **31**, 6017–6031 (2022)
6. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888 (2022)
7. Berton, G., Trivigno, G., Caputo, B., Masone, C.: Eigenplaces: Training viewpoint robust models for visual place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11080–11090 (2023)
8. Blender Online Community: Blender - A 3D Modelling and Rendering Package. Blender Foundation (2018), `http://www.blender.org`
9. Bowles, M., Bromley, M., Allen, M., Scaife, A.: E (2) equivariant self-attention for radio astronomy. arXiv preprint arXiv:2111.04742 (2021)
10. Brewer, C., Franconi, N., Ripley, R., Geist, A., Wise, T., Sabogal, S., Crum, G., Heyward, S., Wilson, C.: Nasa spacecube intelligent multi-purpose system for enabling remote sensing, communication, and navigation in mission architectures. In: Proceedings of the Small Satellite Conference. No. SSC20-VI-07, AIAA/USU (2020)
11. Bruintjes, R.J., Motyka, T., van Gemert, J.: What affects learned equivariance in deep image recognition models? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4838–4846 (2023)
12. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
13. Chase, T., Goodwill, J., Dantu, K., Wilson, C.: Profiling vision-based deep learning architectures on nasa spacecube platforms. In: 2024 IEEE Aerospace Conference. pp. 1–16 (2024). `https://doi.org/10.1109/AERO58975.2024.10521096`
14. Chase Jr, T., Gnam, C., Crassidis, J., Dantu, K.: You only crash once: Improved object detection for real-time, sim-to-real hazardous terrain detection and classification for autonomous planetary landings. arXiv preprint arXiv:2303.04891 (2023)
15. Chen, N., Villar, S.: Se (3)-equivariant self-attention via invariant features. In: Machine Learning for Physics NeurIPS Workshop (2022)
16. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)

17. Chen, W., Liu, Y., Wang, W., Bakker, E.M., Georgiou, T., Fieguth, P., Liu, L., Lew, M.S.: Deep learning for instance retrieval: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7270–7292 (2023). https://doi.org/10.1109/TPAMI.2022.3218591
18. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15050–15061 (2023)
19. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
20. Chou, P.Y., Kao, Y.Y., Lin, C.H.: Fine-grained visual classification with high-temperature refinement and background suppression. arXiv preprint arXiv:2303.06442 (2023)
21. Chung, H., Nam, W.J., Lee, S.W.: Rotation invariant aerial image retrieval with group convolutional metric learning (2020)
22. Cohen, T.S., Welling, M.: Steerable cnns. arXiv preprint arXiv:1612.08498 (2016)
23. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks (2017)
24. Darvish, M., Pouramini, M., Bahador, H.: Towards fine-grained image classification with generative adversarial networks and facial landmark detection. In: 2022 International Conference on Machine Vision and Image Processing (MVIP). pp. 1–6. IEEE (2022)
25. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 741–757. Springer (2020)
26. Diao, Q., Jiang, Y., Wen, B., Sun, J., Yuan, Z.: Metaformer: A unified meta framework for fine-grained recognition. arXiv preprint arXiv:2203.02751 (2022)
27. Dieleman, S., De Fauw, J., Kavukcuoglu, K.: Exploiting cyclic symmetry in convolutional neural networks. In: International conference on machine learning. pp. 1889–1898. PMLR (2016)
28. Do, T.T., Cheung, N.M.: Embedding based on function approximation for large scale image search. IEEE transactions on pattern analysis and machine intelligence **40**(3), 626–638 (2017)
29. Doran, G., Lu, S., Mandrake, L., Wagstaff, K.: Mars orbital image (HiRISE) labeled data set version 3 (2019). https://doi.org/10.5281/zenodo.2538136, https://doi.org/10.5281/zenodo.2538136
30. Downes, L., Steiner, T.J., How, J.P.: Deep learning crater detection for lunar terrain relative navigation. In: AIAA Scitech 2020 Forum (2020). https://doi.org/10.2514/6.2020-1838, https://arc.aiaa.org/doi/abs/10.2514/6.2020-1838
31. Dunkel, E., Swope, J., Towfic, Z., Chien, S., Russell, D., Sauvageau, J., Sheldon, D., Romero-Cañas, J., Espinosa-Aranda, J.L., Buckley, L., Hervas-Martin, E., Fernandez, M., Knox, C.: Benchmarking deep learning inference of remote sensing imagery on the qualcomm snapdragon and intel movidius myriad x processors onboard the international space station. In: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 5301–5304 (2022). https://doi.org/10.1109/IGARSS46834.2022.9884906
32. Ermolov, A., Mirvakhabova, L., Khrulkov, V., Sebe, N., Oseledets, I.: Hyperbolic vision transformers: Combining improvements in metric learning. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7409–7419 (2022)
33. Esteves, C., Allen-Blanchette, C., Zhou, X., Daniilidis, K.: Polar transformer networks. arXiv preprint arXiv:1709.01889 (2017)
34. Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., Chen, D.: Unsupervised pre-training for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14750–14759 (2021)
35. Gaskell, R.W., Barnouin-Jha, O.S., Scheeres, D.J., Konopliv, A.S., Mukai, T., Abe, S., Saito, J., Ishiguro, M., Kubota, T., Hashimoto, T., Kawaguchi, J., Yoshikawa, M., Shirakawa, K., Kominato, T., Hirata, N., Demura, H.: Characterizing and navigating small bodies with imaging data. Meteoritics & Planetary Science **43**(6) (2008)
36. Geist, A., Brewer, C., Davis, M., Franconi, N., Heyward, S., Wise, T., Crum, G., Petrick, D., Ripley, R., Wilson, C., et al.: Spacecube v3. 0 nasa next-generation high-performance processor for science applications. In: Proceedings of the Small Satellite Conference. No. SSC19-XII-02, AIAA/USU (2019)
37. Geist, A., Crum, G., Brewer, C., Afanasev, D., Sabogal, S., Wilson, D., Goodwill, J., Marshall, J., Perryman, N., Franconi, N., et al.: Nasa spacecube next-generation artificial-intelligence computing for stp-h9-scenic on iss. In: Proceedings of the Small Satellite Conference. No. SSC23-P1-32, AIAA/USU (2023)
38. Gentil, C.L., Vayugundla, M., Giubilato, R., Sturzl, W., Vidal-Calleja, T., Triebel, R.: Gaussian process gradient maps for loop-closure detection in unstructured planetary environments. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 1895–1902 (2020)
39. Giubilato, R., Gentil, C.L., Vayugundla, M., Schuster, M.J., Vidal-Calleja, T., Triebel, R.: Gpgm-slam: a robust slam system for unstructured planetary environments with gaussian process gradient maps. arXiv preprint arXiv:2109.06596 (2021)
40. Giubilato, R., Sturzl, W., Wedler, A., Triebel, R.: Challenges of slam in extremely unstructured environments: The dlr planetary stereo, solid-state lidar, inertial dataset. IEEE Robotics and Automation Letters **7**, 8721–8728 (2022)
41. Gong, Y., Huang, L., Chen, L.: Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method. arXiv preprint arXiv:2101.08533 (2021)
42. Goodwill, J., Crum, G., MacKinnon, J., Brewer, C., Monaghan, M., Wise, T., Wilson, C.: Nasa SpaceCube Edge TPU SmallSat Card for Autonomous Operations and Onboard Science-Data Analysis. Small Satellite Conference (2021)
43. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 241–257. Springer (2016)
44. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
45. Gu, G., Ko, B., Kim, H.G.: Proxy synthesis: Learning with synthetic classes for deep metric learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1460–1468 (2021)

46. Guzewich, S.D., Lemmon, M., Smith, C.L., Martínez, G., de Vicente-Retortillo, Á., Newman, C.E., Baker, M., Campbell, C., Cooper, B., Gómez-Elvira, J., Harri, A.M., Hassler, D., Martin-Torres, F.J., McConnochie, T., Moores, J.E., Kahanpää, H., Khayat, A., Richardson, M.I., Smith, M.D., Sullivan, R., de la Torre Juarez, M., Vasavada, A.R., Viúdez-Moreiras, D., Zeitlin, C., Zorzano Mier, M.P.: Mars Science Laboratory Observations of the 2018/Mars Year 34 Global Dust Storm. Geophysical Research Letters **46**(1), 71–79 (Jan 2019). https://doi.org/10.1029/2018GL080839
47. Habte, S.B., Ibenthal, A., Bekele, E.T., Debelee, T.G.: Convolution filter equivariance/invariance in convolutional neural networks: A survey. In: Pan African Conference on Artificial Intelligence. pp. 191–205. Springer (2022)
48. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14141–14152 (2021)
49. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
50. Henkel, C., Singer, P.: Supporting large-scale image recognition with out-of-domain samples. arXiv preprint arXiv:2010.01650 (2020)
51. Henriques, J.F., Vedaldi, A.: Warped convolutions: Efficient invariance to spatial transformations. In: International Conference on Machine Learning. pp. 1461–1469. PMLR (2017)
52. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design (2021)
53. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks (2019)
54. Hutchinson, M.J., Le Lan, C., Zaidi, S., Dupont, E., Teh, Y.W., Kim, H.: Lietransformer: Equivariant self-attention for lie groups. In: International Conference on Machine Learning. pp. 4533–4543. PMLR (2021)
55. Jiang, R., Mei, S.: Polar coordinate convolutional neural network: From rotation-invariance to translation-invariance. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 355–359. IEEE (2019)
56. Johnson, A.E., Aaron, S.B., Chang, J., Cheng, Y., Montgomery, J.F., Mohan, S., Schroeder, S., Tweddle, B.E., Trawny, N., Zheng, J.X.: The Lander Vision System for Mars 2020 Entry Descent and Landing. AAS Guidance Navigation and Control Conference (2017)
57. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
58. Kang, J., Fernandez-Beltran, R., Wang, Z., Sun, X., Ni, J., Plaza, A.: Rotation-invariant deep embedding for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022). https://doi.org/10.1109/TGRS.2021.3088398
59. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. IEEE Robotics and Automation Letters (2023)
60. Kha Vu, C.: Deep metric learning: A (long) survey. https://hav4ik.github.io/articles/deep-metric-learning-survey (2021), accessed: 2023-11-11
61. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)

62. Ki, M., Uh, Y., Choe, J., Byun, H.: Contrastive attention maps for self-supervised co-localization. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2783–2792 (2021). https://doi.org/10.1109/ICCV48922.2021.00280
63. Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning (2020)
64. Krause, J., Deng, J., Stark, M., Fei-Fei, L.: Collecting a large-scale dataset of fine-grained cars (2013)
65. Lauretta, D.S., Hergenrother, C.W., Chesley, S.R., Leonard, J.M., Pelgrift, J.Y., Adam, C.D., Asad, M.A., Antreasian, P.G., Ballouz, R.L., Becker, K.J., Bennett, C.A., Bos, B.J., Bottke, W.F., Brozović, M., Campins, H., Connolly, H.C., Daly, M.G., Davis, A.B., de León, J., DellaGiustina, D.N., d'Aubigny, C.Y.D., Dworkin, J.P., Emery, J.P., Farnocchia, D., Glavin, D.P., Golish, D.R., Hartzell, C.M., Jacobson, R.A., Jawin, E.R., Jenniskens, P., Kidd, J.N., Lessac-Chenen, E.J., Li, J.Y., Libourel, G., Licandro, J., Liounis, A.J., Maleszewski, C.K., Manzoni, C., May, B., McCarthy, L.K., McMahon, J.W., Michel, P., Molaro, J.L., Moreau, M.C., Nelson, D.S., Owen, W.M., Rizk, B., Roper, H.L., Rozitis, B., Sahr, E.M., Scheeres, D.J., Seabrook, J.A., Selznick, S.H., Takahashi, Y., Thuillet, F., Tricarico, P., Vokrouhlický, D., Wolner, C.W.V.: Episodes of particle ejection from the surface of the active asteroid (101955) bennu. Science **366**(6470), eaay3544 (2019). https://doi.org/10.1126/science.aay3544, https://www.science.org/doi/abs/10.1126/science.aay3544
66. Lee, C.: Automated crater detection on mars using deep learning. Planetary and Space Science **170**, 16–28 (2019). https://doi.org/https://doi.org/10.1016/j.pss.2019.03.008, https://www.sciencedirect.com/science/article/pii/S0032063318303945
67. Li, J., Yang, Z., Liu, H., Cai, D.: Deep rotation equivariant network. Neurocomputing **290**, 26–33 (2018)
68. Li, S., Sun, L., Li, Q.: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1405–1413 (2023)
69. Li, W., Zhou, B., Hsu, C.Y., Li, Y., Ren, F.: Recognizing terrain features on terrestrial surface using a deep learning model: An example with crater detection. In: Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery. p. 33–36. GeoAI '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3149808.3149814, https://doi.org/10.1145/3149808.3149814
70. Li, X., Larson, M., Hanjalic, A.: Pairwise geometric matching for large-scale object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5153–5161 (2015)
71. Li, Y., Yang, M., Zhang, Z.: A survey of multi-view representation learning. IEEE transactions on knowledge and data engineering **31**(10), 1863–1883 (2018)
72. Li, Z., Min, W., Song, J., Zhu, Y., Kang, L., Wei, X., Wei, X., Jiang, S.: Rethinking the optimization of average precision: only penalizing negative instances before positive ones is enough. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1518–1526 (2022)
73. Liu, D., Zhao, L., Wang, Y., Kato, J.: Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. Pattern Recognition **140**, 109550 (2023)

74. Liu, Y., Guo, Y., Wu, S., Lew, M.S.: Deepindex for accurate and efficient image retrieval. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 43–50 (2015)
75. Mo, H., Zhao, G.: Ric-cnn: rotation-invariant coordinate convolutional neural network. Pattern Recognition **146**, 109994 (2024)
76. Mohan, D.D., Sankaran, N., Fedorishin, D., Setlur, S., Govindaraju, V.: Moving in the right direction: A regularization for deep metric learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14579–14587 (2020). https://doi.org/10.1109/CVPR42600.2020.01460
77. Motyka, T.: Learned equivariance in convolutional neural networks (2022)
78. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 international joint conference on neural networks (IJCNN). pp. 1–7. IEEE (2020)
79. Musgrave, K., Belongie, S.J., Lim, S.N.: Pytorch metric learning. ArXiv **abs/2008.09164** (2020)
80. National Aeronautics and Space Administration, University of Arizona, Lockheed Martin: Osiris-rex operations timeline. https://www.asteroidmission.org/asteroid-operations/, accessed: 2023-06-01
81. Norman, C., Miller, C., Olds, R., Mario, C., Palmer, E., Barnouin, O., Daly, M., Weirich, J., Seabrook, J., Bennett, C., et al.: Autonomous navigation performance using natural feature tracking during the osiris-rex touch-and-go sample collection event. The Planetary Science Journal **3**(5), 101 (2022)
82. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
83. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
84. Ozaki, K., Yokoo, S.: Large-scale landmark retrieval/recognition under a noisy and diverse dataset. arXiv preprint arXiv:1906.04087 (2019)
85. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1655–1668 (2019). https://doi.org/10.1109/TPAMI.2018.2846566
86. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2016)
87. Romero, D.W., Lohit, S.: Learning partial equivariances from data. Advances in Neural Information Processing Systems **35**, 36466–36478 (2022)
88. Salvador, A., Giró-i Nieto, X., Marqués, F., Satoh, S.: Faster r-cnn features for instance search. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 9–16 (2016)
89. Shabanov, A., Tarasov, A., Nikolenko, S.: Stir: Siamese transformer for image retrieval postprocessing. arXiv preprint arXiv:2304.13393 (2023)
90. Silburt, A., Ali-Dib, M., Zhu, C., Jackson, A., Valencia, D., Kissin, Y., Tamayo, D., Menou, K.: Lunar crater identification via deep learning. Icarus **317**, 27–38 (2019). https://doi.org/https://doi.org/10.1016/j.icarus.2018.06.022, https://www.sciencedirect.com/science/article/pii/S0019103518301386
91. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
92. Srivastava, S., Sharma, G.: Omnivec: Learning robust representations with cross modal sharing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1236–1248 (2024)

93. Sun, S., Zhou, W., Tian, Q., Li, H.: Scalable object retrieval with compact image representation from generic object regions. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **12**(2), 1–21 (2015)
94. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6398–6407 (2020)
95. Tan, F., Yuan, J., Ordonez, V.: Instance-level image retrieval using reranking transformers. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 12105–12115 (2021)
96. Teh, E.W., DeVries, T., Taylor, G.W.: Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 448–464. Springer (2020)
97. The Planetary Society: Cost of osiris-rex. `https://www.planetary.org/space-policy/cost-of-osiris-rex`, accessed: 2023-06-01
98. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8933–8940 (2019)
99. Wang, Q., Lai, J., Yang, Z., Xu, K., Kan, P., Liu, W., Lei, L.: Improving cross-dimensional weighting pooling with multi-scale feature fusion for image retrieval. Neurocomputing **363**, 17–26 (2019)
100. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5022–5030 (2019)
101. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12275–12284 (2020)
102. Wei, X.S., Song, Y.Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S.: Fine-grained image analysis with deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(12), 8927–8948 (2021)
103. Weiler, M., Cesa, G.: General e (2)-equivariant steerable cnns. Advances in neural information processing systems **32** (2019)
104. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Tech. Rep. CNS-TR-201, Caltech (2010), `/se3/wp-content/uploads/2014/09/WelinderEtal10_CUB-200.pdf`, `http://www.vision.caltech.edu/visipedia/CUB-200.html`
105. Weng, L.: Contrastive representation learning. `https://lilianweng.github.io/posts/2021-05-31-contrastive/#parallel-augmentation` (2021), accessed: 2023-11-11
106. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2575–2584 (2020)
107. Wilson, C., George, A.: Csp hybrid space computing. Journal of Aerospace Information Systems **15**(4), 215–227 (2018)
108. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module (2018)

109. Wright, C.A., Eepoel, J.V., Liounis, A.J., Shoemaker, M.A., Deweese, K., Getzandanner, K.M.: Relative Terrain Imaging Navigation (RETINA) Tool for the Asteroid Redirect Robotic Mission (ARRM). AAS Guidance Navigation and Control Conference (2016)
110. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks (2017)
111. Xu, Q., Wang, J., Jiang, B., Luo, B.: Fine-grained visual classification via internal ensemble learning transformer. IEEE Transactions on Multimedia (2023)
112. Yang, F., Matei, B., Davis, L.S.: Re-ranking by multi-feature fusion with diffusion for image retrieval. In: 2015 IEEE Winter Conference on Applications of Computer Vision. pp. 572–579. IEEE (2015)
113. Yang, F., Li, J., Wei, S., Zheng, Q., Liu, T., Zhao, Y.: Two-stream attentive cnns for image retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1513–1521 (2017)
114. Yu, T., Wu, Y., Bhattacharjee, S., Yuan, J.: Efficient object instance search using fuzzy objects matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
115. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
116. Zhang, X., Wang, L., Su, Y.: Visual place recognition: A survey from deep learning perspective. Pattern Recognition **113**, 107760 (2021). https://doi.org/https://doi.org/10.1016/j.patcog.2020.107760, https://www.sciencedirect.com/science/article/pii/S003132032030563X
117. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Computer Vision, IEEE International Conference on (2015)
118. Zheng, L., Yang, Y., Tian, Q.: Sift meets cnn: A decade survey of instance retrieval. IEEE transactions on pattern analysis and machine intelligence **40**(5), 1224–1244 (2017)
119. Zheng, L., Zhao, Y., Wang, S., Wang, J., Tian, Q.: Good practice in cnn feature transfer. arXiv preprint arXiv:1604.00133 (2016)
120. Zhou, W., Li, H., Tian, Q.: Recent advance in content-based image retrieval: A literature survey. arXiv preprint arXiv:1706.06064 (2017)
121. Zhu, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Zheng, W., Sun, X.: Viewpoint-aware loss with angular regularization for person re-identification (2019)
122. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 391–405. Springer (2014)