Acta Astronautica 210 (2023) 393–410

Contents lists available at ScienceDirect

Acta Astronautica

journal homepage: www.elsevier.com/locate/actaastro

AstroVision: Towards autonomous feature detection and description for missions to small bodies using deep learning

Travis Driver [a,*], Katherine A. Skinner [b], Mehregan Dor [a], Panagiotis Tsiotras [a]

[a] Georgia Institute of Technology, Atlanta, GA, USA
[b] University of Michigan, Ann Arbor, MI, USA

ARTICLE INFO

Keywords:
Keypoint detection
Feature description
Feature tracking
Deep learning
Computer vision
Spacecraft navigation
Small bodies

ABSTRACT

Missions to small celestial bodies rely heavily on optical feature tracking for characterization of and relative navigation around the target body. While deep learning has led to great advancements in feature detection and description, training and validating data-driven models for space applications is challenging due to the limited availability of large-scale, annotated datasets. This paper introduces AstroVision, a large-scale dataset comprised of 115,970 densely annotated, real images of 16 different small bodies captured during past and ongoing missions. We leverage AstroVision to develop a set of standardized benchmarks and conduct an exhaustive evaluation of both handcrafted and data-driven feature detection and description methods. Next, we employ AstroVision for end-to-end training of a state-of-the-art, deep feature detection and description network and demonstrate improved performance on multiple benchmarks. The full benchmarking pipeline and the dataset will be made publicly available to facilitate the advancement of computer vision algorithms for space applications.

1. Introduction

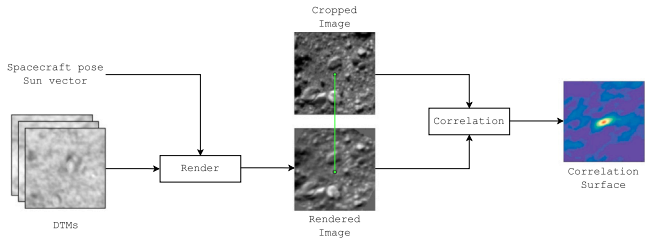
There has been an increasing interest in missions to small bodies (e.g., asteroids, comets) due to their great scientific value, with four currently in operation (OSIRIS-REx, Hayabusa2, Lucy, DART) and two scheduled to launch over the next year (Psyche, Janus). In addition to planetary protection [1] and resource utilization [2,3], small bodies are believed to be remnants from the solar system's formation, and studying their composition could provide insight into the solar system's evolution and the origins of organic life on Earth [4].

Feature tracking is an integral component of current small body shape reconstruction and relative navigation methodologies. However, the current state-of-the-practice relies heavily on humans-in-the-loop. Specifically, human operators on the ground manually identify salient surface features from images acquired during an extensive characterization phase, where the definition of saliency usually undergoes multiple iterations [5]. Extracted features are then combined with *a priori* global shape and spacecraft pose (position and orientation) estimates and used to iteratively construct a collection of digital terrain maps (DTMs), local topography and albedo maps, through a method known as stereophotoclinometry (SPC) [6]. DTM construction typically involves extensive human-in-the-loop verification and carefully designed image acquisition plans to achieve optimal results [7,8]. These topographic features, along with global shape models, are critical for precision navigation and orbit determination for ground-based maneuvering and planning during data acquisition phases [9]. Moreover, upon satisfying strict accuracy and resolution requirements, a catalog of DTMs can be uplinked to the spacecraft and correlated with onboard images to produce an onboard navigation solution for execution of safety-critical maneuvers [10], e.g., during the OSIRIS-REx Touch-And-Go (TAG) sample collection event [5]. While this manual approach has achieved much success, its reliance on extensive human involvement for extended durations limits mission capabilities and increases operational costs [11–13].

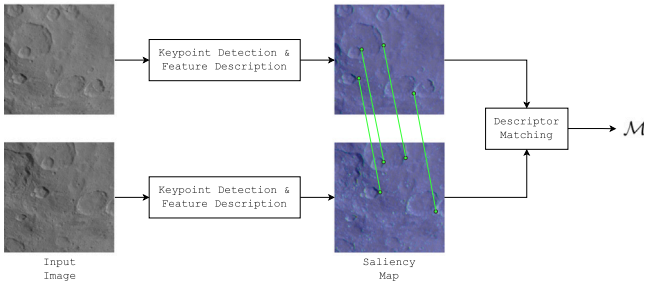
While autonomous feature tracking methods have been investigated to reduce reliance on current human-in-the-loop practices for missions to small bodies [14,15], these works have focused exclusively on traditional *handcrafted* features (e.g., SIFT [16]). More recently, feature detection and description methods that leverage deep *convolutional neural networks* (CNNs) have been shown to significantly outperform handcrafted methods when applied to terrestrial imagery, especially in scenarios involving considerable change in illumination, scale, and perspective [17–20]. However, transferring recent advances in deep learning to small body science applications is challenging due to the unavailability of relevant, annotated data [21]. To the best of our knowledge, there exists no large-scale, annotated dataset comprised entirely of *real* small body images. Indeed, previous work has relied

* Corresponding author.
E-mail address: travisdriver@gatech.edu (T. Driver).

https://doi.org/10.1016/j.actaastro.2023.01.009
Received 29 July 2022; Received in revised form 2 December 2022; Accepted 10 January 2023
Available online 11 January 2023



(a) **DTM-based feature tracking.** DTMs are rendered by leveraging *a priori* spacecraft pose and Sun vector information, along with a photometric model, which is subsequently correlated with the input image to register a match. Adapted from [27].



(b) **Keypoint-based feature tracking.** Keypoints, extracted from each images' saliency map, and their associated descriptors abstract away the image, and tracking is performed by matching local descriptors between images.

Fig. 1. Feature tracking paradigms.

entirely on simulated data [22–24], small sets (i.e., <150 images) of manually annotated real imagery [25], or datasets restricted to a single body [26]. Moreover, operation in space presents a unique set of environmental (e.g., dynamic hard lighting, self-similar surface features) and operational (e.g., significant scale and perspective change during approach) challenges that are likely not adequately captured in available datasets based on terrestrial imagery.

This paper presents *AstroVision*, a large-scale dataset comprised of 115,970 densely annotated, real images of 16 different small bodies from both legacy and ongoing deep space missions to bridge the *terrestrial-to-extraterrestrial* domain gap and facilitate the study of deep learning for autonomous navigation in the vicinity of a small body. The contributions of this paper are as follows: (i) AstroVision is a *first-of-a-kind* dataset for vision-based tasks in the vicinity of a small body with special emphasis on feature tracking applications; (ii) we perform an exhaustive evaluation of both *handcrafted* and *data-driven* keypoint detection and feature description pipelines under challenging conditions on real imagery; (iii) we employ AstroVision for end-to-end training of a state-of-the-art, deep feature detection and description network and demonstrate improved performance with respect to our benchmarks. We make our dataset, benchmarking pipeline, and trained models publicly available at https://github.com/astrovision.

2. Background

In the following subsections, we detail the feature tracking process (Section 2.1) and feature-based pose estimation methodologies (Section 2.2). For completeness, we also provide a brief overview of structure-from-motion in Section 2.3.

2.1. Feature tracking

Robust tracking of salient image features is a critical component of current small body relative navigation methods, as the apparent displacement of tracked features between images can be leveraged to estimate the relative pose of the spacecraft as it moves around the body. In the context of optical feature tracking, saliency typically refers to the ability to detect and precisely localize the feature under multiple viewing conditions (i.e., *repeatability*) and to the distinctiveness of the feature to ensure accurate matching between images (i.e., *reliability*) [18,19]. The current state-of-the-practice for small body feature tracking leverages high-fidelity DTMs of salient surface regions as local feature representations, which require extensive human involvement and mission operations planning for accurate construction [7,8]. Criteria for selecting salient features typically undergo multiple iterations through testing and development of the DTMs [5]. Next, each DTM is combined with *a priori* estimates of the spacecraft's pose and Sun pointing vector, along with a photometric model, to yield a photorealistic rendering of the DTM with respect to the input image. Finally, tracking is performed by comparing the rendering against the input image near the expected feature location using normalized cross-correlation, where a match is declared if a significant correlation peak is detected [5,28]. This process is illustrated in Fig. 1(a). The relative pose of the spacecraft when the image was taken can be computed using the registered matches and the *a priori* DTM position estimates. Therefore, this DTM-based method relies on the fidelity of the *a priori* data products and can only be utilized after the target body has been adequately observed and reconstructed at the required resolutions [10].

In this work we instead investigate approaches to feature tracking that rely on *autonomous* keypoint detection and feature description. Consider two images $I : \Omega \to \mathbb{R}$ and $I' : \Omega' \to \mathbb{R}$ with pixel domains $\Omega \subset \mathbb{R}^2$ and $\Omega' \subset \mathbb{R}^2$, respectively. *Keypoints* $\mathbf{p}_k \in \Omega$ ($\mathbf{p}'_{k'} \in \Omega'$) localize salient regions in the image, which are typically extracted from a saliency map $S : \Omega \to \mathbb{R}$. Saliency can be predefined (e.g., corners) and localized using image filtering methods or learned from data (see Section 3.1).

Feature description is the task of forming a latent representation of the local image data at detected keypoints, where the latent representation commonly takes the form of a d-dimensional vector $\mathbf{d}_k \in \mathbb{R}^d$ referred to as the *descriptor* associated with the keypoint $\mathbf{p}_k$. Consider, for instance, corresponding keypoints $\{\mathbf{p}_k\}_{k\in K}$ and $\{\mathbf{p}'_{k'}\}_{k'\in K'}$ with correspondences defined by $\mathcal{M} := \{(k, \tau(k)) \mid \tau : K \leftrightarrow K'\}$. The overarching goal of feature description is to compute descriptors such that

$$d(\mathbf{d}_l, \mathbf{d}'_{l'}) < \min\left(\min_{k\neq l} d(\mathbf{d}_k, \mathbf{d}'_{l'}), \min_{k'\neq l'} d(\mathbf{d}_l, \mathbf{d}'_{k'})\right) \tag{1}$$

for all $(l, l') \in \mathcal{M}$, where $d(\cdot,\cdot)$ is some distance metric. In words, feature description seeks to assign a descriptor to each keypoint such that descriptors of corresponding keypoints are closer together than those of other non-corresponding keypoints. Common metrics $d(\cdot,\cdot)$ include the Euclidean distance, or the Hamming distance for binary descriptors [29]. We give an overview of different keypoint detection and feature description methodologies based on both handcrafted filtering approaches and deep learning in Section 3.1.

Finally, feature tracking is conducted through detection of keypoints and matching of their corresponding descriptors between images. The objective defined in (1) elicits a straightforward descriptor matching criterion referred to as *mutual nearest-neighbors* (MNN):

$$\tilde{\mathcal{M}} := \left\{(l, l') \mid d(\mathbf{d}_l, \mathbf{d}'_{l'}) < \min_{k\neq l} d(\mathbf{d}_k, \mathbf{d}'_{l'})\right\} \bigcap \left\{(l, l') \mid d(\mathbf{d}_l, \mathbf{d}'_{l'}) < \min_{k'\neq l'} d(\mathbf{d}_l, \mathbf{d}'_{k'})\right\}. \tag{2}$$

In this work we leverage MNN with the Euclidean distance metric for feature matching between images. This keypoint-based tracking process is illustrated in Fig. 1(b). Exploiting recently developed matching approaches based on deep learning [30] will be the subject of future work.

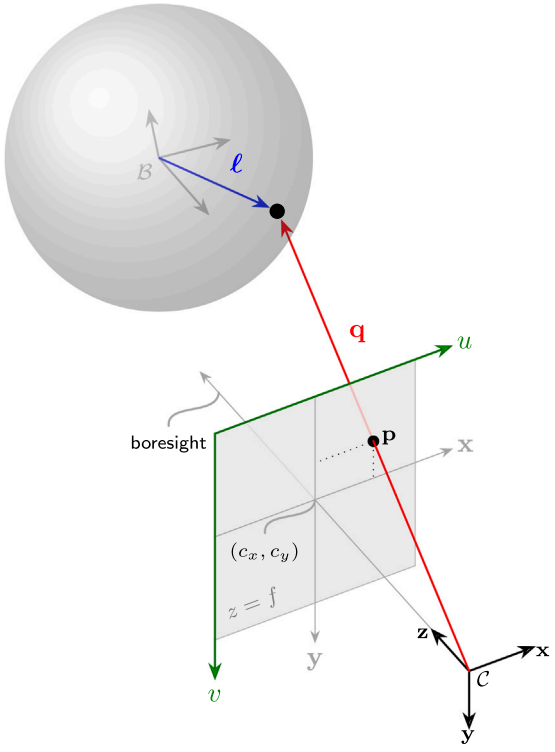


Fig. 2. Camera model geometry.

2.2. Feature-based pose estimation

Consider a spacecraft equipped with a monocular camera navigating around a target small body. The relative pose between cameras can be estimated by tracking the apparent motion of salient surface *landmarks* between images. Formally, let $\mathcal{B}$ denote some body-fixed frame of the small body with origin B, and let $\mathcal{C}_i$ denote the camera frame at time index i with origin C_i. Moreover, let $\boldsymbol{\ell}_k^{\mathcal{B}} = \begin{bmatrix} \ell_{x,k}^{\mathcal{B}} & \ell_{y,k}^{\mathcal{B}} & \ell_{z,k}^{\mathcal{B}} \end{bmatrix}^\top \in \mathbb{R}^3$ denote the vector from B to the kth surface landmark expressed in $\mathcal{B}$, let $\mathbf{q}_k^{\mathcal{C}_i} = \begin{bmatrix} q_{x,k}^{\mathcal{C}_i} & q_{y,k}^{\mathcal{C}_i} & q_{z,k}^{\mathcal{C}_i} \end{bmatrix}^\top \in \mathbb{R}^3$ denote the vector from C_i to the kth landmark expressed in $\mathcal{C}_i$, and let $\mathbf{p}_k^{(i)} = \begin{bmatrix} u_k^{(i)} & v_k^{(i)} \end{bmatrix}^\top \in \mathbb{R}^2$ denote the 2D image coordinates of the kth landmark observed by camera $\mathcal{C}_i$, i.e., the keypoint.

A landmark can be *forward-projected* onto the image plane via

$$\begin{aligned} \underline{\mathbf{p}}_k^{(i)} = \Pi\left(\boldsymbol{\ell}_k^{\mathcal{B}}, T_{\mathcal{C}_i\mathcal{B}}; K\right) &= \frac{1}{d_k^{\mathcal{C}_i}} \left[K \mid \mathbf{0}^{3\times 1}\right] T_{\mathcal{C}_i\mathcal{B}} \underline{\boldsymbol{\ell}}_k^{\mathcal{B}} \\ &= \frac{1}{d_k^{\mathcal{C}_i}} K \mathbf{q}_k^{\mathcal{C}_i} \end{aligned} \tag{3}$$

where $d_k^{\mathcal{C}_i} = q_{z,k}^{\mathcal{C}_i}$ is the landmark depth in $\mathcal{C}_i$, $\underline{\boldsymbol{\ell}}_k^{\mathcal{B}} = \begin{bmatrix} \left(\boldsymbol{\ell}_k^{\mathcal{B}}\right)^\top & 1 \end{bmatrix}^\top \in \mathbb{P}^3$ and $\underline{\mathbf{p}}_k^{(i)} = \begin{bmatrix} \left(\mathbf{p}_k^{(i)}\right)^\top & 1 \end{bmatrix}^\top \in \mathbb{P}^2$ denote the homogeneous coordinates of $\boldsymbol{\ell}_k^{\mathcal{B}}$ and $\mathbf{p}_k^{(i)}$, respectively, $T_{\mathcal{C}_i\mathcal{B}} \in \mathrm{SE}(3)$ denotes the relative pose of $\mathcal{B}$ with respect to $\mathcal{C}_i$:

$$T_{\mathcal{C}_i\mathcal{B}} = \begin{bmatrix} R_{\mathcal{C}_i\mathcal{B}} & \mathbf{r}_{\mathrm{BC}_i}^{\mathcal{C}_i} \\ \mathbf{0}^{1\times 3} & 1 \end{bmatrix}, \tag{4}$$

and K is the camera calibration matrix:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

where f_x and f_y are the *focal lengths* in the x- and y-directions of the camera frame, and (c_x, c_y) is the *principal point* of the camera. The geometry of the pinhole camera model is illustrated in Fig. 2. Conversely, a 2D keypoint may be *backward-projected* into 3D coordinates via

$$\begin{aligned} \underline{\boldsymbol{\ell}}_k^{\mathcal{B}} = \Pi^{-1}\left(\mathbf{p}_k^{(i)}, d_k^{\mathcal{C}_i}, T_{\mathcal{C}_i\mathcal{B}}; K\right) &= T_{\mathcal{C}_i\mathcal{B}}^{-1} \begin{bmatrix} d_k^{\mathcal{C}_i} K^{-1} \underline{\mathbf{p}}_k^{(i)} \\ 1 \end{bmatrix} \\ &= T_{\mathcal{B}\mathcal{C}_i} \underline{\mathbf{q}}_k^{\mathcal{C}_i}. \end{aligned} \tag{6}$$

Then, given corresponding keypoints $\mathbf{p}_k^{(i)}$ and $\mathbf{p}_k^{(j)}$ observed by cameras $\mathcal{C}_i$ and $\mathcal{C}_j$, respectively, the *essential matrix* $E := [\mathbf{r}_{\mathrm{C}_i\mathrm{C}_j}^{\mathcal{C}_j}]_\times R_{\mathcal{C}_j\mathcal{C}_i}$ satisfies

$$\left(\underline{\mathbf{p}}_k^{(j)}\right)^\top K^{-\top} E K^{-1} \underline{\mathbf{p}}_k^{(i)} = 0, \tag{7}$$

where we have assumed a shared camera matrix K for simplicity, and $[\cdot]_\times$ denotes the skew-symmetric cross-product matrix, defined for any $\mathbf{r} \in \mathbb{R}^3$, such that

$$[\mathbf{r}]_\times = \begin{bmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{bmatrix}. \tag{8}$$

The well-known five-point algorithm [31] can be used to solve for E given five or more correspondences. Finally, $R_{\mathcal{C}_j\mathcal{C}_i}$ and $\mathbf{r}_{\mathrm{C}_i\mathrm{C}_j}^{\mathcal{C}_j}$ (up to some unknown scale) can be estimated using singular value decomposition (SVD) of the associated essential matrix and by imposing the *Cheirality constraint*, i.e., triangulating the landmark associated with keypoints $\mathbf{p}_k^{(i)}$, $\mathbf{p}_k^{(j)}$ and enforcing that the associated landmark lies in front of the cameras [32].

2.3. Structure-from-motion

In the structure-from-motion (SfM) or simultaneous localization and mapping (SLAM) setting, we are interested in *simultaneously* estimating a collection of camera poses $\mathcal{T} := \left\{T_{\mathcal{C}_i\mathcal{B}} \in \mathrm{SE}(3) \mid i = 1, \ldots, m\right\}$ and a network of landmarks (the *map*) $\mathcal{L} := \{\boldsymbol{\ell}_k^{\mathcal{B}} \in \mathbb{R}^3 \mid k = 1, \ldots, n\}$. Note that the SfM solution is innately expressed in some arbitrary body-fixed frame since most SfM techniques assume operation in a static scene, typically referred to as the "world" frame [33]. SfM seeks the maximum *a-posteriori* (MAP) estimate of the poses $\mathcal{T}$ and landmarks $\mathcal{L}$, given the (independent) keypoint *measurements* $\mathcal{P} := \{\hat{\mathbf{p}}_k^{(i)} \in \mathbb{R}^2 \mid i = 1, \ldots, m, k = 1, \ldots, n\}$:

$$\mathcal{T}^*, \mathcal{L}^* = \arg\max_{\mathcal{T},\mathcal{L}} p(\mathcal{T}, \mathcal{L} \mid \mathcal{P}) \tag{9}$$

$$\propto \arg\max_{\mathcal{T},\mathcal{L}} p(\mathcal{T}, \mathcal{L})\, p(\mathcal{P} \mid \mathcal{T}, \mathcal{L}) \tag{10}$$

$$= p(\mathcal{T}, \mathcal{L}) \prod_i \prod_k p\left(\hat{\mathbf{p}}_k^{(i)} \mid T_{\mathcal{C}_i\mathcal{B}}, \boldsymbol{\ell}_k^{\mathcal{B}}\right). \tag{11}$$

By assuming measurements $\hat{\mathbf{p}}_k^{(i)}$ are corrupted by zero-mean Gaussian noise, i.e., $\hat{\mathbf{p}}_k^{(i)} = \mathbf{p}_k^{(i)} + \boldsymbol{\eta}_k^{(i)}$ where $\boldsymbol{\eta}_k^{(i)} \sim \mathcal{N}(\mathbf{0}, \Sigma_k^{(i)})$, we get

$$p\left(\hat{\mathbf{p}}_k^{(i)} \mid T_{\mathcal{C}_i\mathcal{B}}, \boldsymbol{\ell}_k^{\mathcal{B}}\right) \propto \exp\left\{\|\underline{\hat{\mathbf{p}}}_k^{(i)} - \Pi\left(\boldsymbol{\ell}_k^{\mathcal{B}}, T_{\mathcal{C}_i\mathcal{B}}; K\right)\|_{\Sigma_k^{(i)}}^2\right\}, \tag{12}$$

where $\|\mathbf{e}\|_\Sigma^2 := \mathbf{e}^\top \Sigma^{-1} \mathbf{e}$. The MAP estimate can be formulated as the solution to a nonlinear least-squares problem by taking the negative logarithm of (11):

$$\mathcal{T}^*, \mathcal{L}^* = \arg\min_{\mathcal{T},\mathcal{L}} \sum_{i,k} \|\underline{\hat{\mathbf{p}}}_k^{(i)} - \Pi\left(\boldsymbol{\ell}_k^{\mathcal{B}}, T_{\mathcal{C}_i\mathcal{B}}; K\right)\|_{\Sigma_k^{(i)}}^2, \tag{13}$$

where we have omitted the priors $p(\mathcal{T}, \mathcal{L})$ for conciseness and generality, which can be ignored if no prior information is assumed (i.e., $p(\mathcal{T}, \mathcal{L}) = const.$) or can encode relative pose constraints via known dynamical models [34]. This process is commonly referred to as *Bundle Adjustment* (BA). Note that the optimization process of SPC decouples estimation of the poses and landmarks, i.e., *a priori* landmark position and camera pose estimates are passed back-and-forth between

the pose determination and DTM construction steps, respectively, until convergence [6].

In this work, we focus on two-view pose estimation by estimating the essential matrix using the five-point algorithm. Future work will focus on incorporating our feature detection and description methods into a full SfM pipeline.

3. Related work

In this section, we give an overview of both handcrafted and data-driven feature detection and description methods (Section 3.1), and then discuss existing datasets and benchmarks for vision tasks in the vicinity of a small body (Section 3.2) and data-driven relative navigation techniques (Section 3.3).

3.1. Feature detection and description

Many computer vision algorithms rely on local image features. The seminal work of David Lowe's Scale Invariant Feature Transform (SIFT) [16] laid the foundation for the field, where he outlined a rigorous framework for identifying and describing image features. SIFT follows a *detect-then-describe* paradigm, whereby a series of predetermined (or *handcrafted*) filters are applied to the image for keypoint localization, followed by pooling and normalization of image gradients to form the descriptor. SIFT aims to extract features that are invariant to changes in scale, illumination, and rotation. Keypoints are extracted from local extrema of the saliency map derived by convolving the difference of Gaussians (DoG) kernel with the input image, as the DoG function provides a close approximation to the scale-normalized Laplacian of Gaussian function which has been shown to be scale invariant [35]. This detection scheme generally results in keypoints centered around large gradients in the image (e.g., edges, corners). Descriptors are then computed by pooling gradients in a local window of each keypoint into histograms according to their orientation, where a canonical orientation is assigned to each keypoint according to the dominant gradient orientation to provide robustness to rotation. The oriented histograms are then concatenated and normalized to form the descriptor vector. Speeded-up Robust Features (SURF) built upon the success of SIFT to enable more efficient feature detection and description by leveraging integral images to eliminate the need for computing the DoG [36]. Oriented FAST and Rotated BRIEF (ORB) has become a popular alternative to SIFT, especially for SLAM applications [29]. ORB is based on Features from Accelerated Segment Test (FAST) detectors [37] and Binary Robust Independent Elementary Features (BRIEF) descriptors [38] and outputs binary descriptor vectors, enabling more efficient matching.

More recently, feature detection and description methods that leverage deep *convolutional neural networks* (CNNs) have achieved state-of-the-art performance and have been shown to outperform handcrafted methods, especially in scenarios involving significant illumination, scale, and perspective change [17,19,20,39]. The first data-driven methods focused on individual components of the full image processing pipeline, including keypoint detection [40], orientation estimation [41], and feature description [42]. Yi et al. [43] developed the first complete learning-based pipeline, Learned Invariant Feature Transform (LIFT). LIFT uses a patch-based Siamese training architecture and implements each component of the traditional feature detector and descriptor scheme sequentially using CNNs. The approach relies on an incremental training procedure to pretrain each subnetwork component individually, with a final training phase that optimizes over the entire network end-to-end. LFNet [39] proposed a sequential two-stage approach: the first stage learns keypoint detection and the second stage learns feature description. SuperPoint [17] developed a network composed of separate interest point and descriptor decoders that operate on a spatially reduced representation of the input image from a shared encoder network. Simulated data of simple geometric shapes is used to pretrain the interest point detector, which is then combined with a random homographic warping procedure to train the network end-to-end in a *self-supervised* fashion.

Towards joint detection and description, the seminal work of D2-Net [18] proposed a *detect-and-describe* approach that trains a single deep CNN to detect and describe salient image features. *Reliability* (or *distinctiveness*) of descriptors is enforced through a triplet margin ranking loss term which is weighted according to soft detection scores to jointly enforce *repeatability* of detections. R2D2 [19] leverages the detect-and-describe paradigm to perform simultaneous feature detection and description, but repeatability and reliability are enforced in separate terms in the loss function. Repeatability is enforced through maximization of the cosine similarity of the detection scores of corresponding image patches, while reliability of the descriptors is learned through maximizing a differentiable approximation of the average precision [44] between corresponding patch descriptors. ASLFeat [20] builds upon the success of D2-Net and proposes a multi-level detection scheme to generate detection scores that enable more accurate keypoint localization, and leverages deformable convolutional networks (DCNs) [45] to model local geometric variations in the image and learn more transformation invariant features. ASLFeat is trained using the BlendedMVS [46] and GL3D [47] datasets, which contain 125,623 high-resolution images of 543 different scenes annotated with depth information using scene reconstructions from a dense SfM pipeline. Although the training data is exceptionally comprehensive, we seek to capitalize on the recent success of deep feature detection and description methods by training these models on domain-relevant data to increase feature tracking performance for missions to small bodies.

3.2. Datasets and benchmarks for vision tasks in the vicinity of a small body

Morrell et al. [15] and Dennison et al. [14] conduct an extensive evaluation of handcrafted feature extraction methods on synthetic images of comet 67P and asteroid 433 Eros, respectively, where SIFT demonstrates superior overall performance with respect to the algorithms studied. While the results are promising, the experiments were conducted in a controlled, simulated environment of a single target body, and their benchmarks were not made publicly available. Conversely, in this paper, we benchmark both handcrafted and data-driven feature detection and description methods on real imagery of multiple small bodies with different surface characteristics and under varying illumination, scale, and perspective.

With respect to small body image datasets, we are only aware of the work by Zhou et al. [23,24], which includes images of both mock-up and computer-generated asteroid models. The authors fabricate in-house models to represent arbitrary small bodies as opposed to leveraging available models of asteroids observed from past or current small body missions. The authors in [23,24] do not apply their learned models on real mission imagery. In our work, we train and test our approach on real imagery.

3.3. Data-driven relative navigation

Fuchs et al. [25] train a random forest classifier on patches extracted from 119 images of the comets Hartley 2 and Tempel 1. However, significant performance degradation is observed when applied to unseen bodies, demonstrating the necessity to train models on data from a diverse set of small body instances. Pugliatti et al. [22] employ a custom U-Net for segmentation of small body images into a constrained set of classes (i.e., terminator, boulders, craters, surface, background) using synthetic images of 7 different small bodies (e.g., 101955 Bennu, 21 Lutetia). However, the performance suffers when applied to real images.

Data-driven crater detection has also received much attention, especially for lunar applications. Wang et al. [48] leverage a lightweight

Table 1
Dataset information.

Mission	Target	Type	# Images	Shape model Ref.
Dawn [53]	1 Ceres	Asteroid (G-type)	38540	Park et al. [54]
	4 Vesta	Asteroid (V-type)	17504	Gaskell et al. [55]
Cassini [56]	Dione (Saturn IV)	Icy Moon	1381	Gaskell [57]
	Epimetheus (Saturn XI)	Icy Moon	133	Daly et al. [58]
	Janus (Saturn X)	Icy Moon	184	Daly et al. [58]
	Mimas (Saturn I)	Icy Moon	307	Gaskell [59]
	Phoebe (Saturn IX)	Icy Moon	96	Daly et al. [58]
	Rhea (Saturn V)	Icy Moon	665	Daly et al. [60]
	Tethys (Saturn III)	Icy Moon	751	Daly et al. [60]
Hayabusa [61]	25143 Itokawa	Asteroid (S-type)	603	Park et al. [54]
Hayabusa2 [62]	162173 Ryugu	Asteroid (C-type)	788	Gaskell et al. [63]
Mars Express [64]	Phobos (Mars I)	Moon	890	Gaskell [65]
NEAR [66]	433 Eros	Asteroid (S-type)	11156	Gaskell [67]
OSIRIS-REx [68]	101955 Bennu	Asteroid (B-type)	16618	Barnouin et al. [69]
Rosetta [70,71]	67P/C-G	Comet	26314	Gaskell et al. [72]
	21 Lutetia	Asteroid (M-type)	40	Jorda et al. [73]
TOTALS: 8 missions	16 bodies		115,970 images	

CNN architecture pretrained on Martian crater samples to extract feature maps, which are then fed into a fully convolutional architecture to perform crater detection. Detected craters are then matched against an *a priori* database to produce a navigation solution. Silvestrini et al. [49] train a MobileNetV2 [50] architecture to detect craters in synthetic images of the lunar surface. Silburt et al. [51] implement a custom U-Net architecture to detect and identify craters from digital elevation maps (DEMs). Downes et al. [52] leverage the architecture of [51], but instead perform crater detection on orthorectified images of the lunar surface. Lee et al. [26] employ a CNN-based object detector to discriminate between a catalog of handpicked lunar surface landmarks, while also predicting landmark detection probabilities as a function of the Sun's relative azimuth and elevation. The reliance on a catalog of known landmarks for navigation and the specification of craters as the most salient features limit the range of applications of these technologies. Instead of explicitly specifying the features-of-interest beforehand, we allow the network to learn the most salient features for a wide variety of surface characteristics.

4. The AstroVision dataset

In this section, we present our novel small body image dataset, referred to as AstroVision, for training and evaluation of keypoint detection and feature description methods. AstroVision features over 110,000 real images of 16 small bodies from 8 missions, as shown in Fig. 3. We describe the full data generation pipeline of AstroVision in the following subsections. Next, we develop a novel benchmarking suite (Section 5) and train a deep feature detection and description network (Section 6) using our dataset.

4.1. Image and ancillary data extraction

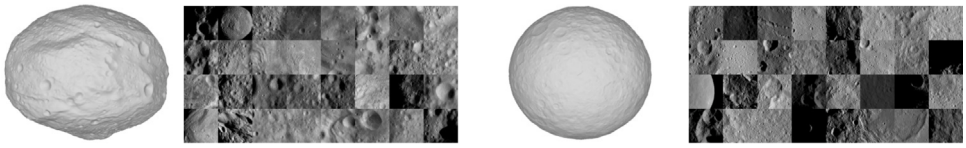
AstroVision leverages publicly available images and ancillary data (i.e., camera pose, camera calibration, shape models) from both legacy and active small body science missions provided through NASA's Planetary Data System (PDS) [74] and maintained by NASA's Navigation and Ancillary Information Facility (NAIF). High-fidelity shape models (i.e., watertight, 3D triangular surface meshes) are developed as part of the relative navigation pipeline of small body missions, as they are critical for characterization of the body and relative navigation in subsequent phases. Specifically, shape models for these missions are typically developed using SPC [6]. SPC leverages feature correspondences between images captured during an extended characterization phase procured by human operators on the ground. A network of landmarks is estimated using stereophotogrammetry and subsequently densified using photometric stereo techniques via *a priori* camera pose and Sun pointing estimates and a reflectance model. The process yields high-quality shape models that are precisely registered to the images and provide the foundation for our small body image dataset. For more details about the shape reconstruction and state estimation process, we refer the reader to [6], [10], and [7]. Moreover, information and references for the various missions, images, and shape models used in this work are provided in Table 1.

Images provided by PDS are commonly stored using the Flexible Image Transport System (FITS), the standard data format used in astronomy, with pixel intensity values in units of either radiance ($\mathrm{W\,s^{-1}\,m^{-2}}$) or reflectance (unitless). We linearly scale pixel intensities to $[0, 1]$ before converting to a grayscale Portable Network Graphics (PNG) image. Photometrically calibrated (e.g., flat field and dark current correction) images were utilized when available. Moreover, we provide undistorted images to ensure alignment with the depth maps by leveraging geometric distortion estimates derived during a meticulous calibration procedure conducted both on the ground and during flight by mission scientists. See Appendix A for specific calibration details for each mission.

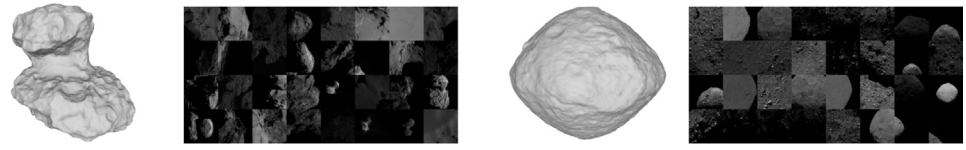
4.2. Data generation

The suite of AstroVision data products includes a landmark map, a depth map, and a mask for each image as shown in Fig. 4. The *landmark map* provides a consistent, discrete set of reference points for sparse correspondence computation and is derived by forward-projecting vertices from a medium-resolution (i.e., ~800k facets) shape model onto the image plane. We classify visible landmarks by tracing rays[1] from the landmarks toward the camera origin and recording landmarks whose line-of-sight ray does not intersect the 3D model. The *depth map* provides a dense representation of the imaged surface and is computed by backward-projecting rays at each pixel in the image and recording the depth of the intersection between the ray and a high-resolution (i.e., ~3.2 million facets) shape model. Finally, the

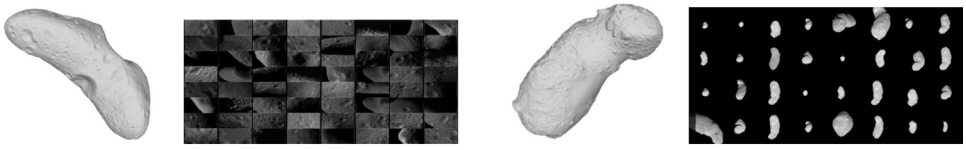
[1] Ray tracing uses the Trimesh library: https://trimsh.org/



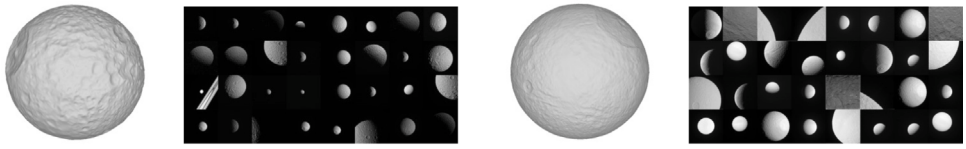
(a) Dawn @ 4 Vesta (b) Dawn @ 1 Ceres



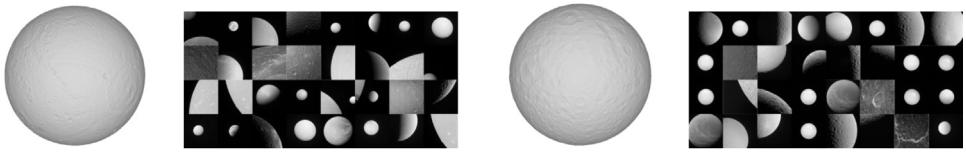
(c) Rosetta @ 67P/Churyumov-Gerasimenko (d) OSIRIS-REx @ 101955 Bennu



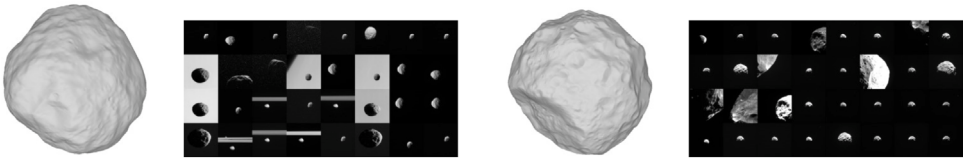
(e) NEAR @ 433 Eros (f) Hayabusa @ 25143 Itokawa



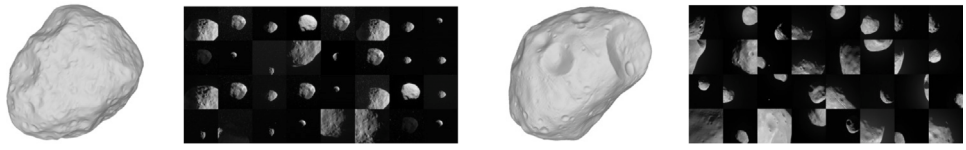
(g) Cassini @ Mimas (Saturn I) (h) Cassini @ Tethys (Saturn III)



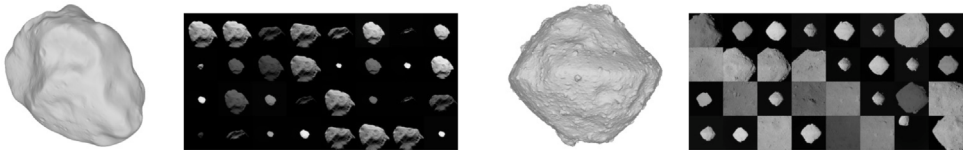
(i) Cassini @ Dione (Saturn IV) (j) Cassini @ Rhea (Saturn V)



(k) Cassini @ Janus (Saturn X) (l) Cassini @ Phoebe (Saturn IX)

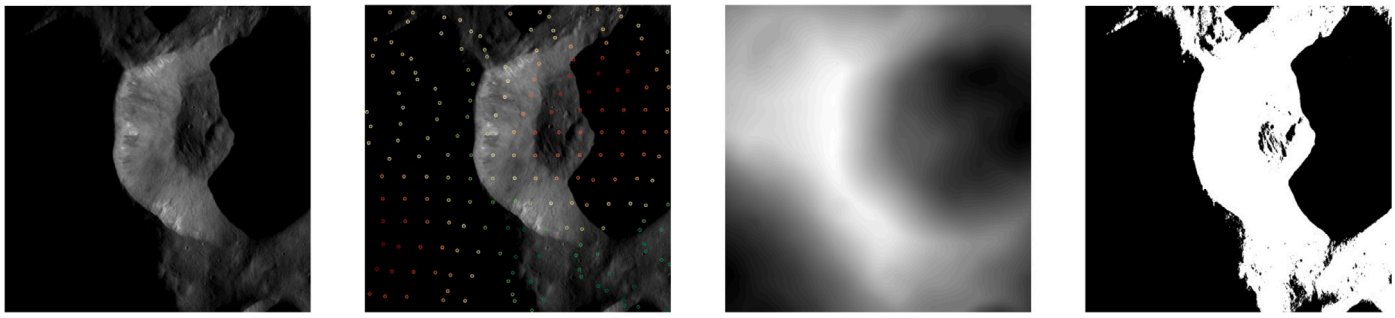


(m) Cassini @ Epimetheus (Saturn XI) (n) Mars Express @ Phobos (Mars I)



(o) Rosetta @ 21 Lutetia (p) Hayabusa2 @ 162173 Ryugu

Fig. 3. AstroVision image datasets. Shape model references are provided in Table 1.



(a) Image (b) Landmark map (c) Depth map (d) Mask

Fig. 4. Example of AstroVision data products.

mask provides an estimate of the non-occluded portions of the imaged surface.

In order to generate the visibility masks, both global and dynamic intensity thresholding was used. For the more recent missions (i.e., Dawn, Hayabusa2, OSIRIS-REx, Rosetta), global thresholding was leveraged. For some of the legacy missions (i.e., Cassini, Hayabusa, NEAR, Mars Express), variable vignetting was observed, primarily influenced by exposure time. Therefore, Otsu's method [75] was employed to compute a dynamic threshold for these instances. While illuminated pixels could have been computed by tracing the Sun's incident light ray, estimating the mask independently of the ground truth scene geometry proved to be a useful tool for algorithmic outlier rejection, in addition to an extensive manual cleaning process. Specifically, we compute the ratio of the intersection area between the intensity mask and depth map and the total area of the mask as an alignment measure between the shape model and image, where a nominal value of 0.97 was empirically chosen. Moreover, we found that utilizing these intensity masks during training led to significant performance increases, which will be discussed further in Section 6.5.

5. Small body feature benchmarks

In this section, we conduct a comprehensive evaluation of existing feature detection and description methods using the proposed AstroVision dataset. First, we detail our suite of performance metrics and verification procedures. Then, we present and discuss the benchmarking results.

5.1. Performance metrics

We evaluate the matching performance on a per image pair basis using the standard metrics precision, recall, and accuracy. First, *precision* defines the inlier ratio of the putative matches (as determined by our verification process described in the following section):

$$\text{precision} = \frac{\text{\# correct matches}}{\text{\# putative matches}}. \tag{14}$$

Second, *recall* describes the number of identified ground truth matches:

$$\text{recall} = \frac{\text{\# correct matches}}{\text{\# ground truth matches}}. \tag{15}$$

Third, *accuracy* measures the matching performance with respect to the total number of computed features:

$$\text{accuracy} = \frac{\text{\# correct matches \& nonmatches}}{\text{\# features}}. \tag{16}$$

We classify correct nonmatches as keypoints that were not included in the set of putative or ground truth matches, where we take the minimum of the number of such keypoints in each image in the pair [14].

Finally, we compute the maximum of the angular error between the estimated and ground truth pose orientation and (unit) translation in degrees. Specifically, the angle of rotation between the estimated $\tilde{\mathbf{q}}_{C_jC_i}$ and ground truth $\mathbf{q}_{C_jC_i}$ relative orientation quaternions

$$\epsilon_q := \cos^{-1}\left(2\langle\tilde{\mathbf{q}}_{C_jC_i}, \mathbf{q}_{C_jC_i}\rangle^2 - 1\right) \tag{17}$$

is used as a metric [76] for the orientation error, and

$$\epsilon_t := \cos^{-1}\left(\tilde{\mathbf{r}}^{C_j}_{C_iC_j} \cdot \mathbf{r}^{C_j}_{C_iC_j} \Big/ \|\tilde{\mathbf{r}}^{C_j}_{C_iC_j}\|\|\mathbf{r}^{C_j}_{C_iC_j}\|\right) \tag{18}$$

provides a measure of the translation error. The final pose error metric is taken to be $\epsilon := \max(\epsilon_q, \epsilon_t)$. The normalized cumulative error curve for ϵ is computed for each test sequence and the area under the curve (AUC) is reported for thresholds of 5°, 10° and 20°. We compute AUC using the explicit integration procedure of [30] rather than coarse histograms.

5.2. Implementation

We evaluated the performance of ORB [29] and SIFT [16] as two representatives of *handcrafted* features, since these methods are widely used and have been leveraged as the handcrafted baselines in previous works [17,19,20,30]. Three state-of-the-art *data-driven* features were selected that leverage different learning approaches and architectures (previously detailed in Section 2): SuperPoint [17], R2D2 [19], and ASLFeat [20]. We use the OpenCV implementations of ORB and SIFT and the open-source implementations and pretrained models of the learned features made available by the respective authors. Each feature is limited to detecting 5,000 keypoints and descriptors.

Given a set of keypoints and descriptors, putative matches are computed using MNN (described in Section 2.1). Matches are verified by first backward-projecting (via Eq. (6)) each keypoint in the first image into 3D world coordinates using the ground truth calibration and depth map. The 3D points are then forward-projected (via Eq. (3)) into the second image, and matches are verified by checking that the projected image coordinates are within some distance γ to the keypoint of its matched feature, where we empirically chose a value of $\gamma = 5$ pixels (see Appendix B). Ground truth matches are estimated in a similar way for computing recall, where a ground truth match is registered if there exists a keypoint within $\gamma = 5$ pixels of the projected image coordinate.

Finally, poses are computed from the putative matches by first estimating the essential matrix using the five-point method [31], implemented in OpenCV's `findEssentialMat` function, and RANSAC with an inlier threshold of 1 pixel, followed by SVD of the essential matrix to determine the relative pose, implemented in OpenCV's `recoverPose` function. Evaluation is conducted for $2N$ randomly generated image pairs with at least 20% overlap with respect to the landmark map, where N is the number of images in the respective test dataset rounded up to the nearest multiple of 100, and metrics are averaged over all the image pairs.

5.3. Results & discussion

We evaluated both handcrafted (i.e., ORB and SIFT) and data-driven (i.e., SuperPoint, R2D2, and ASLFeat) feature detection and description

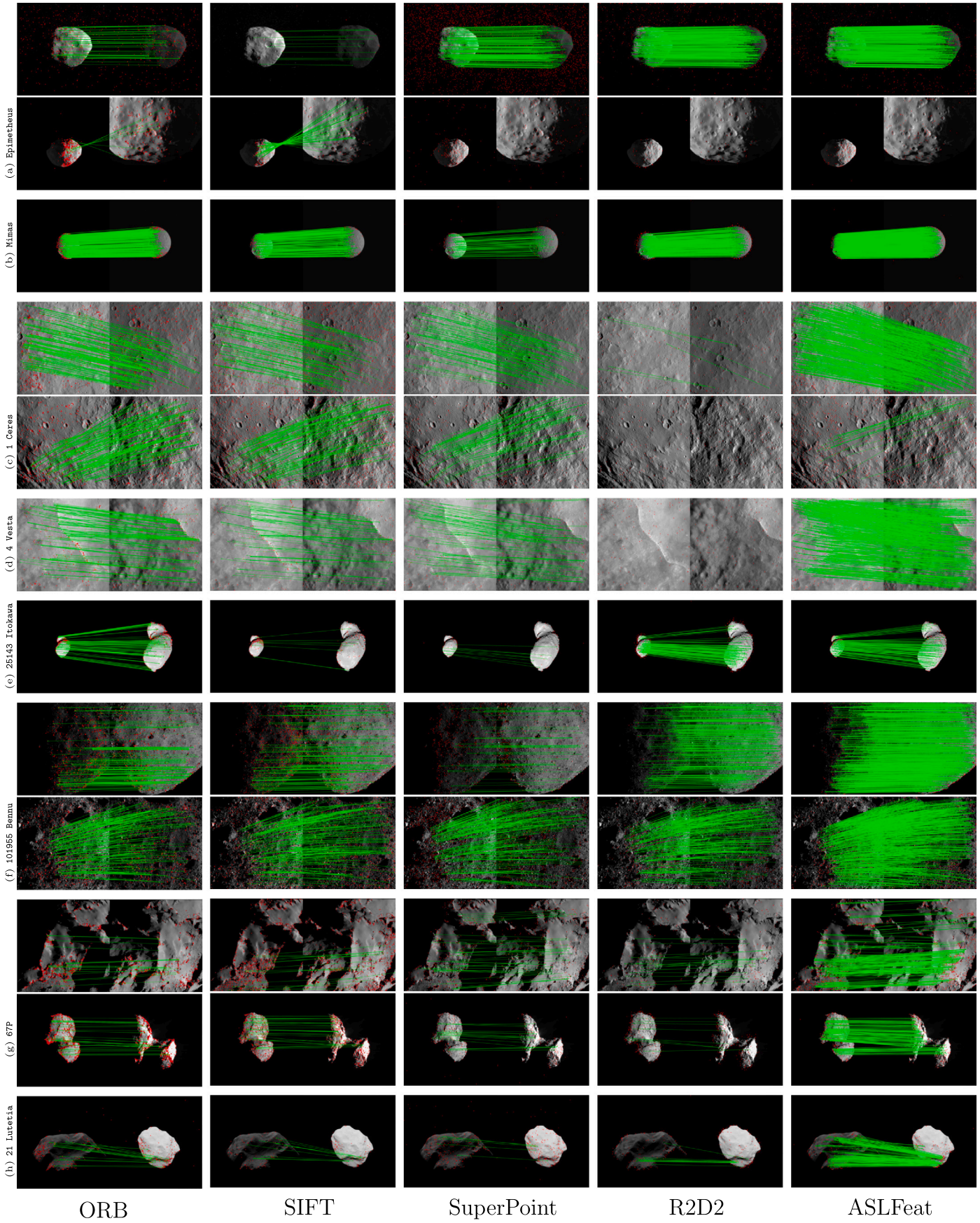


Fig. 5. Qualitative comparison of feature matching. Correct matches are drawn in green, and the keypoints of incorrect matches are drawn in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
AstroVision feature benchmarks. Feature performance with respect to precision (P), recall (R), accuracy (A), and pose AUC in percentages. **First** and second best results are bolded and underlined, respectively. See Section 5.1 for metric definitions.

Dataset							AUC		
(Mean GSD, Median GSD)	# Images	Feature	# Matches	P	R	A	@5°	@10°	@20°
Cassini @ Epimetheus (Saturn XI)	133	ORB	**789**	25.1	26.0	67.1	**2.9**	**10.0**	**16.1**
(326.7 m/pixel, 255.4 m/pixel)		SIFT	204	**32.5**	**36.6**	54.7	2.7	9.5	15.0
		SuperPoint	396	13.6	26.1	59.2	2.6	7.5	12.8
		R2D2	423	25.3	26.1	**77.1**	**2.9**	9.1	14.7
		ASLFeat	386	27.4	29.0	74.7	2.7	8.2	13.7
Cassini @ Mimas (Saturn I)	307	ORB	**746**	10.3	9.2	59.3	0.0	0.1	0.1
(1,176.2 m/pixel, 943.8 m/pixel)		SIFT	340	14.3	15.1	41.1	**0.2**	**0.2**	**0.4**
		SuperPoint	121	8.6	10.4	50.5	0.0	0.0	0.0
		R2D2	209	13.8	8.8	**75.5**	0.1	0.1	0.1
		ASLFeat	372	**21.8**	**15.7**	65.3	**0.2**	**0.2**	0.3
Dawn @ 1 Ceres	3624	ORB	1377	43.0	63.7	74.2	7.8	18.0	30.1
(122.0 m/pixel, 35.4 m/pixel)		SIFT	**1656**	42.3	72.2	69.4	**28.8**	**44.3**	**56.6**
		SuperPoint	442	42.9	**75.7**	70.1	13.1	28.3	43.5
		R2D2	954	**50.0**	52.8	**85.8**	8.9	20.0	32.4
		ASLFeat	1535	48.4	67.8	80.2	12.9	27.1	42.4
Dawn @ 4 Vesta	2006	ORB	1348	32.7	46.2	70.8	5.9	12.4	21.0
(63.3 m/pixel, 21.2 m/pixel)		SIFT	1350	37.1	52.3	64.0	**17.9**	28.7	38.8
		SuperPoint	506	38.7	55.0	65.8	11.3	21.3	32.7
		R2D2	926	55.9	46.7	**86.9**	11.4	22.3	34.1
		ASLFeat	**1524**	**59.0**	**66.1**	84.3	17.5	**31.9**	**46.0**
Hayabusa @ 25143 Itokawa	603	ORB	**625**	5.1	4.7	51.3	1.4	2.3	3.8
(95.5 cm/pixel, 78.7 cm/pixel)		SIFT	217	4.8	5.0	35.8	1.9	3.3	4.8
		SuperPoint	79	7.3	**12.7**	42.3	1.7	3.1	5.4
		R2D2	339	10.7	9.4	**67.0**	**2.6**	**4.6**	**8.0**
		ASLFeat	338	**13.5**	11.3	47.5	2.2	4.2	7.6
OSIRIS-REx @ 101955 Bennu	1789	ORB	**1496**	12.2	13.4	60.1	1.5	3.2	5.6
(21.9 cm/pixel, 9.9 cm/pixel)		SIFT	1317	13.7	15.3	55.2	5.6	8.8	11.8
		SuperPoint	747	18.1	20.3	55.4	3.8	7.3	11.1
		R2D2	502	29.3	18.3	**84.7**	4.2	8.6	13.8
		ASLFeat	1378	**33.1**	**30.9**	68.7	**8.0**	**14.4**	**20.9**
Rosetta @ 67P	3039	ORB	**1303**	15.3	14.9	59.5	0.7	1.5	3.3
(5.5 m/pixel, 2.4 m/pixel)		SIFT	1168	15.7	16.6	44.7	2.4	4.8	7.7
		SuperPoint	485	17.6	20.7	49.9	1.6	3.6	6.4
		R2D2	634	20.2	16.5	**79.3**	1.9	3.9	7.1
		ASLFeat	1147	**25.0**	**24.0**	62.8	**3.4**	**6.4**	**10.6**
Rosetta @ 21 Lutetia	40	ORB	430	21.3	21.0	57.0	2.3	4.2	8.6
(230.5 m/pixel, 228.1 m/pixel)		SIFT	283	23.7	31.7	46.6	5.9	9.8	15.9
		SuperPoint	381	26.7	30.7	55.5	4.2	8.0	16.2
		R2D2	588	33.2	25.6	**74.7**	3.1	6.0	13.3
		ASLFeat	**970**	**42.9**	**35.0**	71.9	**6.0**	**12.1**	**23.8**

algorithms. These results are summarized in Table 2, and qualitative comparisons are provided in Fig. 5. We also list the mean and median ground sample distance (GSD) for each dataset, i.e., the distance on the surface of the body covered by each pixel, which is a function of the distance to the surface of the body when the image was taken, as well as the camera intrinsics. SIFT demonstrates competitive performance on the Dawn and Cassini datasets, outperforming many of the data-driven methods, but suffers when applied to datasets with harsher illumination (i.e., Rosetta @ 67P, OSIRIS-REx @ 101955 Bennu). The efficacy of the orientation encoding of SIFT in certain scenarios can be seen in Fig. 5(a), although this behavior does not seem to be typical (see Fig. 10). Superpoint achieves high recall but low precision and generally underperforms with respect to all other methods except ORB. Although R2D2 demonstrates high precision and accuracy, we found that the feature matches generally result in poor pose estimates. Finally, ASLFeat exhibits high precision, recall, and accuracy, which translates into generally superior relative pose estimates as indicated by the AUC score, and consistently ranks among the top-performing methods with respect to all datasets. Therefore, we selected the ASLFeat network for end-to-end training using the AstroVision data products. This is detailed in the next section.

We recognize the low AUC values for all methods on the Cassini @ Mimas dataset. The relatively symmetric and homogeneous surface topology of Mimas generally resulted in low matching precision, and image pairs with high inlier ratios usually corresponded to pairs with relatively low baseline with respect to the radial imaging depth (e.g., Fig. 5(b)) resulting in spurious relative translation estimates given even small amounts of measurement noise. Indeed, the Cassini @ Mimas images have a mean GSD of 1,176.2 m/pixel, due in part to the approximately 190,000 km average radial distance to the body, almost four times that of the next highest value. We also observed correspondence configurations that resulted in ambiguous essential matrix estimates. This suggests that the points may lie close to a so-called *critical surface* [77], special surfaces which yield multiple essential

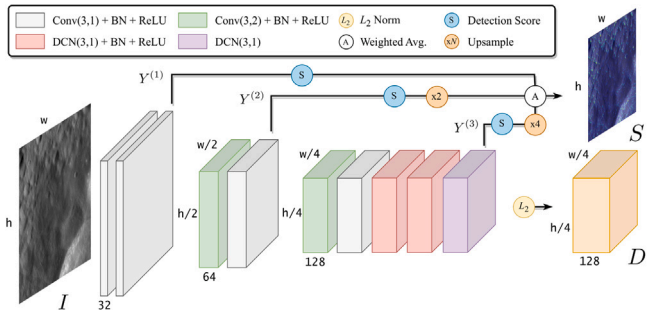


Fig. 6. ASLFeat architecture. Conv(a, b) (and DCN(a, b)) denotes convolution with a kernel size of a and stride b.

matrix estimates that satisfy Eq. (7). Detection (e.g., via the iterative method presented in [78]) and rectification (e.g., by considering more views in a full SfM solution) of these degenerate configurations will be the subject of future work.

6. Learning features from small body imagery

In this section, we leverage the AstroVision dataset to train a deep feature detection and description network.

6.1. Network architecture

Predicated on our evaluation benchmarks, we leverage the ASLFeat [20] network architecture, shown in Fig. 6. Given an image $I \in \mathbb{R}^{h\times w\times c}$, ASLFeat uses a single deep CNN to generate both a detection score (saliency) map $S \in \mathbb{R}^{h\times w}$ and a dense descriptor volume $D \in \mathbb{R}^{h/4\times w/4\times d}$.

The score map S is computed through aggregation of elements in intermediate feature maps $Y^{(\ell)} \in \mathbb{R}^{h_\ell\times w_\ell\times b_\ell}$, $\ell = 1,2,3$. Specifically, local peakiness over the channels $Y_c^{(\ell)}$, $c = 1,\dots,b_\ell$, of the descriptor volume is used to compute channel-wise detection scores (dropping the ℓ subscript and superscript for conciseness):

$$\beta_{ij}^c = \texttt{softplus}\left(y_{ij}^c - \frac{1}{b}\sum_t y_{ij}^t\right), \tag{19}$$

where y_{ij}^c is the element at pixel $(i,j) \in \{1,\dots,h\}\times\{1,\dots,w\}$ in Y_c and $\texttt{softplus}(x) = \log(1+\exp(x))$. Next, the local detection score is defined as

$$\alpha_{ij}^c = \texttt{softplus}\left(y_{ij}^c - \frac{1}{|\mathcal{N}(i,j)|}\sum_{(i',j')\in\mathcal{N}(i,j)} y_{i'j'}^c\right), \tag{20}$$

where $\mathcal{N}(i,j)$ is the set of 9 neighbors of the pixel (i,j) (including itself). The elements of the ℓ^{th} score map $S^{(\ell)}$ are computed as $s_{ij}^{(\ell)} = \max_c(\alpha_{ij}^c, \beta_{ij}^c)$. Finally, each score map is bilinearly upsampled to the spatial resolution of the input image, and the elements in the final score map S are computed via a weighted average

$$S = \frac{1}{\sum_\ell w_\ell}\sum_\ell w_\ell S^{(\ell)}, \tag{21}$$

where the weights w_1, w_2, w_3 have been empirically set to $1,2,3$, respectively.

Given correspondences $\mathcal{M} := \{(k,\tau(k)) \mid \tau : K \leftrightarrow K'\}$ between keypoints $\{\mathbf{p}_k\}_{k\in K}$ and $\{\mathbf{p}'_{k'}\}_{k'\in K'}$ extracted from images I and I', respectively, the total loss is formulated as

$$L(D,D',S,S';\mathcal{M}) = \frac{1}{|\mathcal{M}|}\sum_{(l,l')\in\mathcal{M}} \frac{s_l s'_{l'}}{\sum_{(k,k')\in\mathcal{M}} s_k s'_{k'}} m(\mathbf{d}_l, \mathbf{d}'_{l'}). \tag{22}$$

where s_k ($s'_{k'}$) is the detection score and $\mathbf{d}_k$ ($\mathbf{d}'_{k'}$) is the descriptor at keypoint $\mathbf{p}_k$ ($\mathbf{p}'_{k'}$), and $m(\cdot,\cdot)$ is the descriptor reliability loss. Note that descriptors and detection scores at *subpixel* locations can be computed through (e.g., bilinear) interpolation of the score map S (S') and descriptor volume D (D'). ASLFeat leverages a hardest-contrastive margin ranking loss [79] to enforce descriptor reliability:

$$m(\mathbf{d}_l, \mathbf{d}'_{l'}) = \max\left(\|\mathbf{d}_l - \mathbf{d}'_{l'}\| - M_p, 0\right) + \max\left(M_n - \min\left(\min_{k'\neq l'}\|\mathbf{d}_l - \mathbf{d}'_{k'}\|, \min_{k\neq l}\|\mathbf{d}_k - \mathbf{d}'_{l'}\|\right), 0\right), \tag{23}$$

where M_p and M_n are the margins for positive and negative pairs, respectively.

The formulated loss L in Eq. (22) produces a weighted average of the margin terms m over all matches based on their detection scores. Thus, in order for the loss to be minimized, the most distinctive correspondences (with a lower margin term) will get higher relative detection scores and vice versa.

Table 3
Train/test split.

Dataset	# Images
Train	
Cassini @ Dione (Saturn IV) (**D**)	1381
Cassini @ Janus (Saturn X) (**J**)	184
Cassini @ Phoebe (Saturn IX) (**P**)	96
Cassini @ Rhea (Saturn V) (**R**)	665
Cassini @ Tethys (Saturn III) (**T**)	751
Dawn @ 1 Ceres (**C**)	34916
Dawn @ 4 Vesta (**V**)	15498
Hayabusa2 @ 162173 Ryugu (**U**)	788
Mars Express @ Phobos (**M**)	890
NEAR @ 433 Eros (**E**)	11156
OSIRIS-REx @ 101955 Bennu (**B**)	14829
Rosetta @ 67P (**G**)	23275
TOTAL	104429
Test	
Cassini @ Epimetheus (Saturn XI)	133
Cassini @ Mimas (Saturn I)	307
Dawn @ 1 Ceres	3624
Dawn @ 4 Vesta	2006
Hayabusa @ 25143 Itokawa	603
OSIRIS-REx @ 101955 Bennu	1789
Rosetta @ 21 Lutetia	40
Rosetta @ 67P	3039
TOTAL	11541

6.2. Implementation details

We train ASLFeat using a procedure similar to the original implementation [20]. The train/test split is shown in Table 3, where we use an approximate 90/10 split.

Training. The model is trained from scratch with ground truth cameras and depths from our AstroVision dataset. The training consumes ~900k image pairs resized to 480 × 480 using a batch size of 2. The relative perspective change between an image pair is limited during training, where the angle of rotation between the orientation quaternions of the respective images with respect to the body-fixed frame, as defined by Eq. (17), is used as a metric for the relative perspective change between two images. We ignore image pairs with a value greater than $\epsilon_q(\mathbf{q}_{\mathcal{C}_i\mathcal{B}}, \mathbf{q}_{\mathcal{C}_j\mathcal{B}}) = 60°$. We also ignore image pairs with large differences in radial distance to the body, i.e., $\min(\|\mathbf{r}_{\mathcal{C}_i\mathcal{B}}\|, \|\mathbf{r}_{\mathcal{C}_j\mathcal{B}}\|)/\max(\|\mathbf{r}_{\mathcal{C}_i\mathcal{B}}\|, \|\mathbf{r}_{\mathcal{C}_j\mathcal{B}}\|) < 0.5$. Training is supervised by computing dense putative keypoints and ground truth correspondences for each image pair, i.e., $\{\mathbf{p}_k\}_{k\in K}$, $\{\mathbf{p}'_{k'}\}_{k'\in K'}$, and $\mathcal{M}$ in Eq. (22). First, putative training image pairs are computed by querying the landmark map for sparse correspondences and only keeping

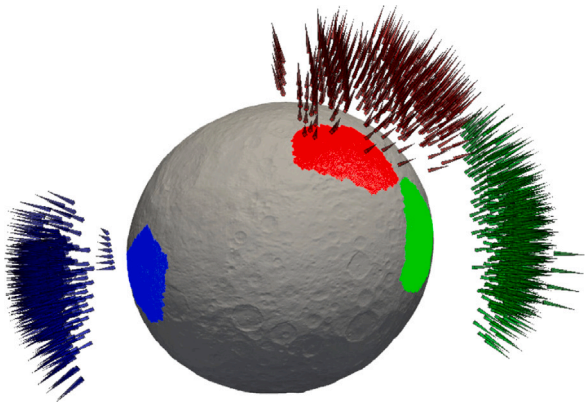


Fig. 7. Example image clusters. Three representative clusters from the Dawn @ 1 Ceres dataset where the camera frustums and observed surface area of each cluster are color coded. Only cameras below an altitude of 1000 km are drawn. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pairs with at least 128 shared landmarks. Next, points sampled from a uniform grid of coordinates in the first image are taken to be the set of putative keypoints $\{\mathbf{p}_k\}_{k\in K}$, and the dense matches $\mathcal{M}$ are derived by projecting these keypoints into the second image using the ground truth depth and camera calibration and pose labels. Note that uniformly sampling putative keypoints across the entire image as opposed to restricting learning to a specific class of features (e.g., craters) allows the network to learn the most salient features for a wide variety of surface characteristics. Additionally, the visibility masks are used to remove matches that have keypoints in occluded regions of either image (see Section 6.5). Learning gradients are computed for image pairs that have at least 128 matches, while a maximum of 512 randomly selected matches are used for back-propagating gradients.

Each input image is standardized to have zero mean and unit standard deviation. The SGD optimizer is used with momentum of 0.9, and an exponentially decaying learning rate is used with an initial value of 0.1. We use a two-stage training procedure as suggested by [20]. Specifically, all regular convolutions are trained for 400k iterations in the first stage of training. In the second stage, the DCNs are trained with the initial learning rate of 0.01 for another 400k iterations.

Testing. Non-maximum suppression is applied (sized 3) to remove detections that are spatially too close. The position of the detected keypoints is improved using a local refinement and edge-elimination procedure over the detection score map following the approach used in SIFT [16]. The descriptors are then bilinearly interpolated at the refined (subpixel) positions. We select the top-k keypoints (nominally $k = 5000$) with respect to their detection scores and empirically discard those whose scores are lower than 0.5.

6.3. Experiments

We withheld data corresponding to 4 different small bodies with variable surface characteristics from training, i.e., Cassini @ Epimetheus, Cassini @ Mimas, Hayabusa @ 25143 Itokawa, and Rosetta @ 21 Lutetia. In doing so, we test the network's ability to reliably compute features upon arrival at a previously unexplored small body. The network was also tested on held-out images of small bodies it saw during training. This emulates a scenario in which images obtained during earlier stages of a mission could be used to train the network for feature extraction in later phases of the mission. In order to minimize overlap between the train and test sets, we cluster images within each dataset according to the backward-projected 3D coordinates of the principle point in each image using k-means [80] with a value of $k = 64$. Seven of these clusters are held out for testing while the remaining are used during training. A visualization of a subset of the clusters for the Dawn @ 1 Ceres dataset is shown in Fig. 7. Matching and verification are conducted using the procedure described in Section 5.2.

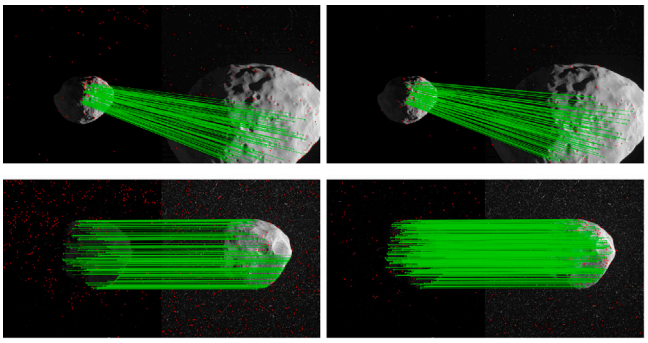
6.4. Results & discussion

The ASLFeat model trained on AstroVision data, i.e., ASLFeat-CVGBEDTRPJMU, is compared against the pretrained model. These results are shown in Table 4 and qualitative comparisons are shown in Fig. 8. The model trained on AstroVision consistently outperforms the pretrained model with respect to precision, recall, accuracy, and AUC. Importantly, the AstroVision-trained model achieves increased matching performance on many of the novel testing instances, i.e., Cassini

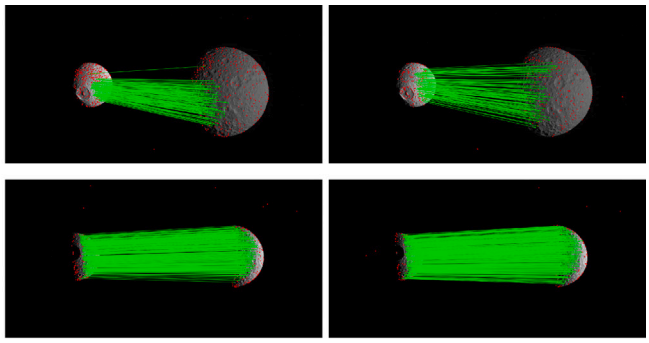
Table 4
AstroVision-trained model compared to pretrained. Performance of the AstroVision-trained ASLFeat model compared to pretrained with respect to precision (P), recall (R), accuracy (A), and pose AUC in percentages. See Section 5.1 for metric definitions.

							AUC		
Dataset	# Images	Feature	# Matches	P	R	A	@5°	@10°	@20°
Cassini @ Epimetheus (Saturn XI)[a]	133	ASLFeat	386	27.4	**29.0**	**74.7**	**2.7**	8.2	13.7
		ASLFeat-CVGBEDTRPJMU	**396**	**28.9**	27.5	74.1	**2.7**	**8.6**	**14.0**
Cassini @ Mimas (Saturn I)[a]	307	ASLFeat	**372**	21.8	**15.7**	65.3	**0.2**	**0.2**	**0.3**
		ASLFeat-CVGBEDTRPJMU	328	**23.6**	14.9	**67.1**	0.0	0.1	0.2
Dawn @ 1 Ceres	3624	ASLFeat	**1535**	48.4	67.8	80.2	12.9	27.1	42.4
		ASLFeat-CVGBEDTRPJMU	1514	**52.8**	**71.5**	**82.1**	**15.9**	**31.6**	**46.9**
Dawn @ 4 Vesta	2006	ASLFeat	**1524**	59.0	66.1	84.3	**17.5**	31.9	46.0
		ASLFeat-CVGBEDTRPJMU	1412	**70.3**	**69.7**	**87.4**	**17.5**	**33.0**	**48.7**
Hayabusa @ 25143 Itokawa[a]	603	ASLFeat	338	13.5	**11.3**	47.5	2.2	4.2	7.6
		ASLFeat-CVGBEDTRPJMU	**363**	**15.2**	11.0	**53.7**	**2.9**	**5.0**	**8.8**
OSIRIS-REx @ 101955 Bennu	1789	ASLFeat	**1378**	33.1	**30.9**	68.7	**8.0**	**14.4**	**20.9**
		ASLFeat-CVGBEDTRPJMU	858	**34.2**	28.4	**79.5**	6.7	12.6	19.3
Rosetta @ 67P	3039	ASLFeat	**1147**	25.0	**24.0**	62.8	3.4	6.4	10.6
		ASLFeat-CVGBEDTRPJMU	837	**30.4**	23.9	**69.8**	**4.2**	**7.9**	**13.4**
Rosetta @ 21 Lutetia[a]	40	ASLFeat	**970**	**42.9**	**35.0**	71.9	6.0	12.1	**23.8**
		ASLFeat-CVGBEDTRPJMU	778	41.3	31.1	**76.3**	**8.4**	**13.2**	22.3

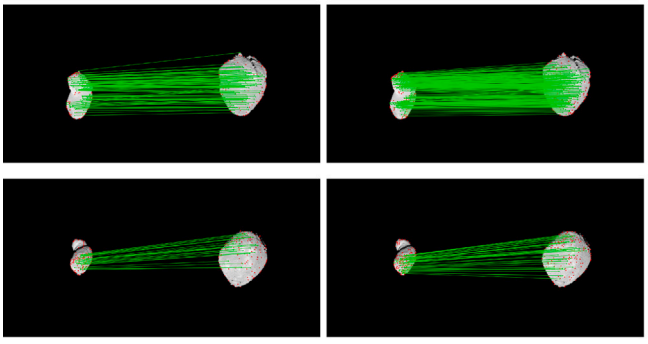
[a]No images of this body were included in the training set.



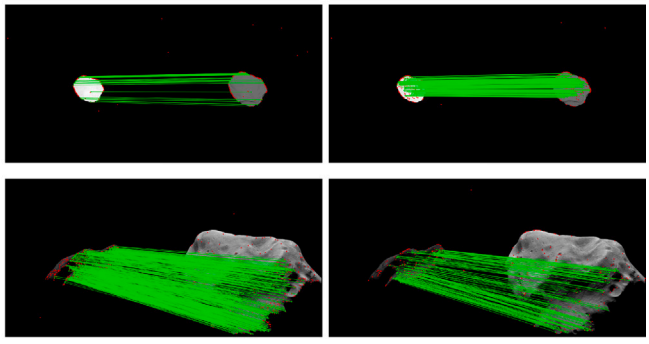
(a) Cassini @ Epimetheus (Saturn XI)



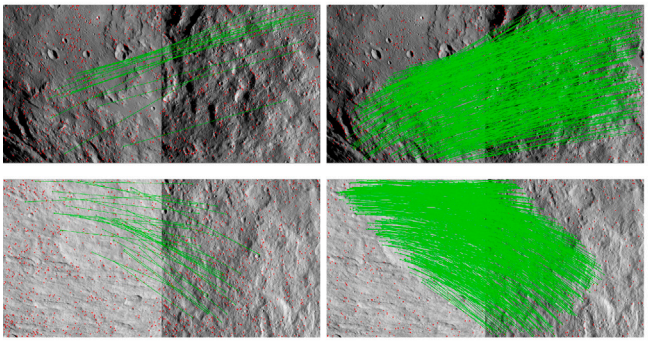
(b) Cassini @ Mimas (Saturn I)



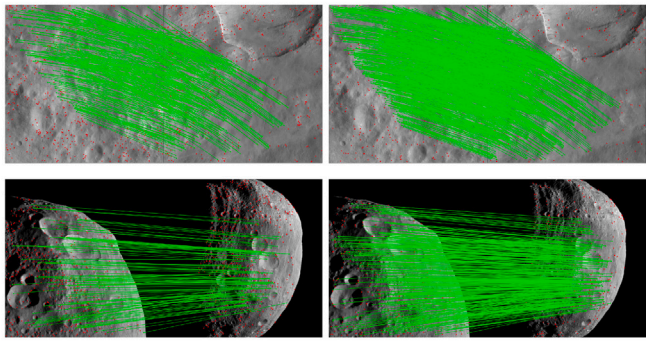
(c) Hayabusa @ 25143 Itokawa



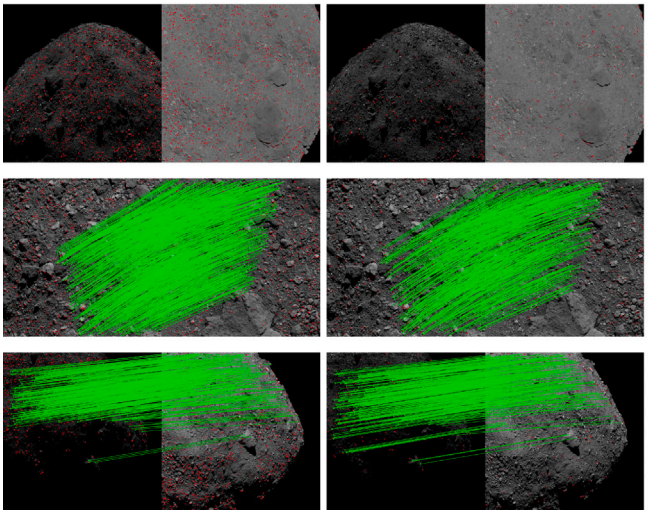
(d) Rosetta @ 21 Lutetia



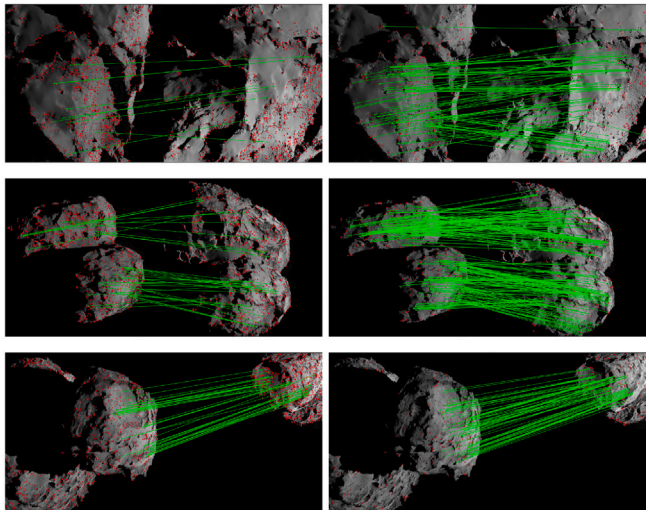
(e) Dawn @ 1 Ceres



(f) Dawn @ 4 Vesta



(g) OSIRIS-REx @ 101955 Bennu



(h) Rosetta @ 67P

Fig. 8. Qualitative comparison between pretrained (left) and AstroVision-trained (right) model feature matches. Correct matches are drawn in green, and the keypoints of incorrect matches are drawn in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

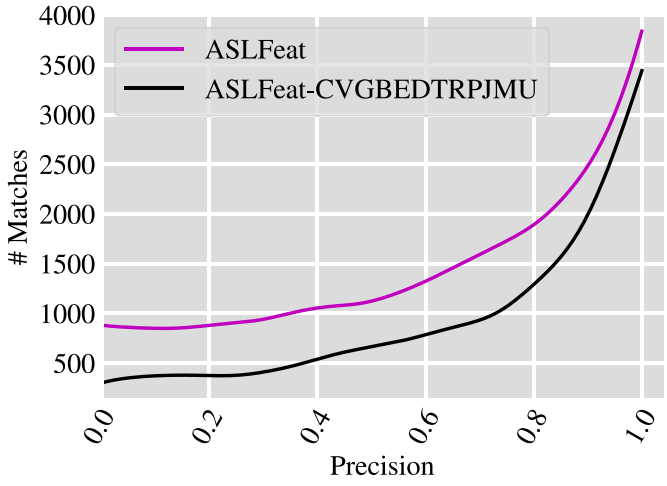


Fig. 9. Precision versus number of matches on the OSIRIS-REx @ 101955 Bennu test set.

@ Epimetheus, Cassini @ Mimas, and Hayabusa @ 25143 Itokawa. Indeed, very little is known about the surface characteristics of a small body prior to arrival. Our model obtains higher precision and accuracy on all novel test instances with the exception of Rosetta @ 21 Lutetia. Despite the lower precision and recall on Rosetta @ 21 Lutetia, we are able to achieve significantly better pose estimates as indicated by the pose AUC metric. This is most likely due to the more uniform distribution of matches on the surface of the body, whereas the pretrained network primarily computes matches on the boundary of the body (see Fig. 8(d)). Our model generally exhibits slightly lower recall, but achieves higher AUC on all novel test instances excluding Cassini @ Mimas.

Moreover, ASLFeat-CVGBEDTRPJMU demonstrates impressive performance on the held-out images of the small bodies it saw during training. Our model demonstrates considerably higher performance with respect to all metrics on the Dawn @ 1 Ceres and Dawn @ 4 Vesta test sets. Intuitively, training on AstroVision data results in more conservative feature matching on the difficult OSIRIS-REx @ 101955 Bennu and Rosetta @ 67P test sets, as indicated by the higher precision and accuracy and lower recall and number of matches, which exhibit hard and rapidly changing illumination, significant perspective changes, and repetitive surface characteristics. We achieve slightly lower pose AUC as compared to the pretrained model for the OSIRIS-REx@ 101955 Bennu test set despite having higher precision and significantly higher accuracy. This is most likely due to the reduced number of matches, although this is primarily restricted to low-precision image pairs as shown in Fig. 9. Indeed, for difficult image pairs with precision close to zero, ASLFeat-CVGBEDTRPJMU features typically result in an order of magnitude fewer incorrect matches compared to the pretrained model. An example of this is provided in Fig. 8(g).

We experimented with training the network on OSIRIS-REx @ 101955 Bennu data only, referred to as ASLFeat-B, as we suspected the network may be prioritizing discrimination of other feature classes more relevant to the other training instances due to the unique and challenging surface features of Bennu and the lower number of training images relative to some of the other missions (e.g., Dawn @ 1 Ceres, Rosetta @ 67P). Benchmarking results for this experiment are presented in Table 5. ASLFeat-B achieves increased performance with respect to all metrics compared to the pretrained model on the OSIRIS-REx @ 101955 Bennu dataset. We postulate that adding more small body instances with similar surface characteristics will increase performance.

We also compared matching precision against perspective and illumination changes in Figs. 10 and 11. We leverage Eq. (17) as a measure for perspective change, and

$$\epsilon_s := \cos^{-1}\left(\hat{\mathbf{s}}^{\mathcal{C}_i} \cdot \hat{\mathbf{s}}^{\mathcal{C}_j}\right) \tag{24}$$

as a measure of illumination change, where $\hat{\mathbf{s}}^{\mathcal{C}_i}$ and $\hat{\mathbf{s}}^{\mathcal{C}_j}$ denote the (unit) Sun vector in $\mathcal{C}_i$ and $\mathcal{C}_j$, respectively. Our model exhibits superior invariance to both perspective and illumination changes for all test sets.

The detection score maps S for the respective models, as described in Section 6.1, are visualized in Fig. 12. It can be seen that the pretrained model repeatably places high confidence to edges formed from hard shadowing and to features on the boundary between the body and deep space. Features in these regions are known to be relatively unreliable and not repeatable, as the appearance of these features can change dramatically due to the deformation of the shadows, or become completely occluded as the body rotates about its axis [15]. However, the model trained on AstroVision learns to assign low confidence to these regions and gives higher confidence to features corresponding to salient topographic structures such as rocky outcroppings and crater rims.

6.5. Ablation study for masking

We found that utilizing the visibility masks during training led to faster convergence and greatly increased overall performance. Specifically, a grid of image coordinates from the first image is projected into the second image using the ground truth calibrations of each camera and the depth map to generate a collection of ground truth matches during training. We mask keypoints according to the visibility masks of each image and ignore matches with keypoints in occluded regions during training. The results in Table 6 demonstrate how utilizing the visibility masks during training greatly improves the overall performance. It can be seen that there is a slight degradation in precision for

Table 5
ASLFeat-B Benchmark performance. Performance of ASLFeat-B, i.e., ASLFeat trained on OSRIS-REx @ 101955 Bennu data only, with respect to precision (P), recall (R), accuracy (A), and pose AUC in percentages. See Section 5.1 for metric definitions.

					AUC	
Dataset	# Matches	P	R	A	@5°	@10°	@20°
Cassini @ Epimetheus (Saturn XI)	475	28.6	26.7	60.6	2.4	7.2	12.1
Cassini @ Mimas (Saturn I)	306	12.1	8.7	58.1	0.0	0.0	0.1
Dawn @ 1 Ceres	1631	48.8	61.7	76.3	12.2	26.4	41.6
Dawn @ 4 Vesta	1430	48.2	54.9	78.2	11.7	23.0	34.9
Hayabusa @ 25143 Itokawa	337	9.6	6.9	43.3	1.6	2.8	4.9
OSIRIS-REx @ 101955 Bennu	1400	35.4	34.7	70.7	7.9	14.9	21.9
Rosetta @ 67P	1075	22.2	20.9	58.6	2.4	4.8	8.2
Rosetta @ 21 Lutetia	561	25.8	16.3	67.3	1.0	3.3	9.2

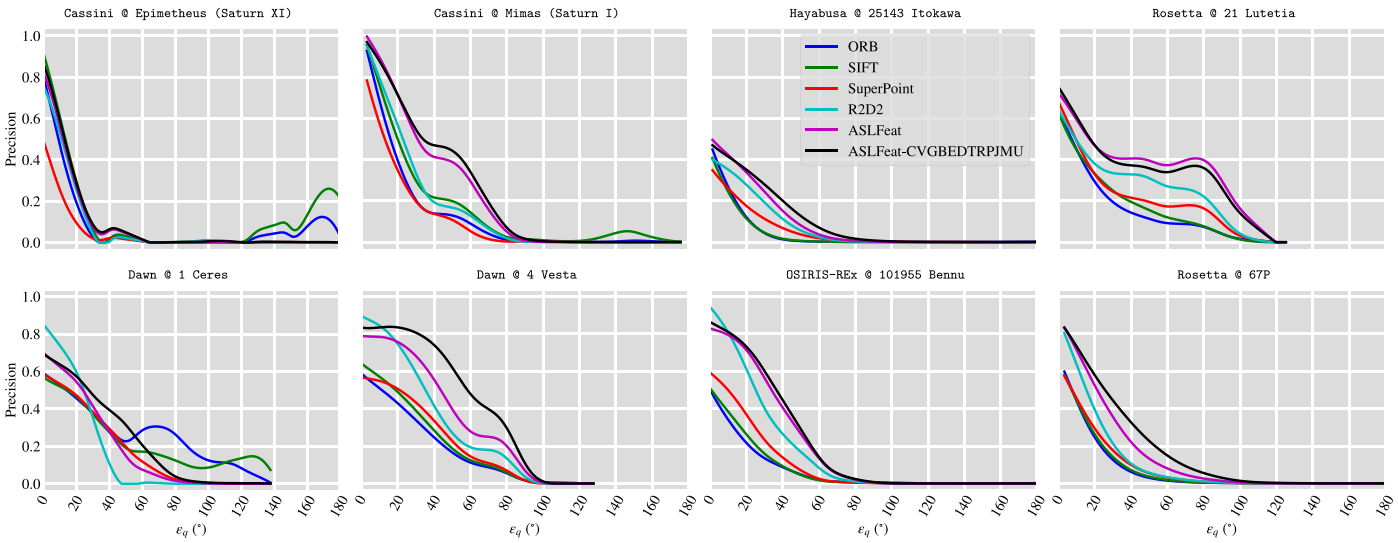


Fig. 10. Perspective change versus precision. Perspective change is measured by $\epsilon_q(\mathbf{q}_{C_iB}, \mathbf{q}_{C_jB})$ as defined in Eq. (17), i.e., the minimum rotation angle between the respective cameras.

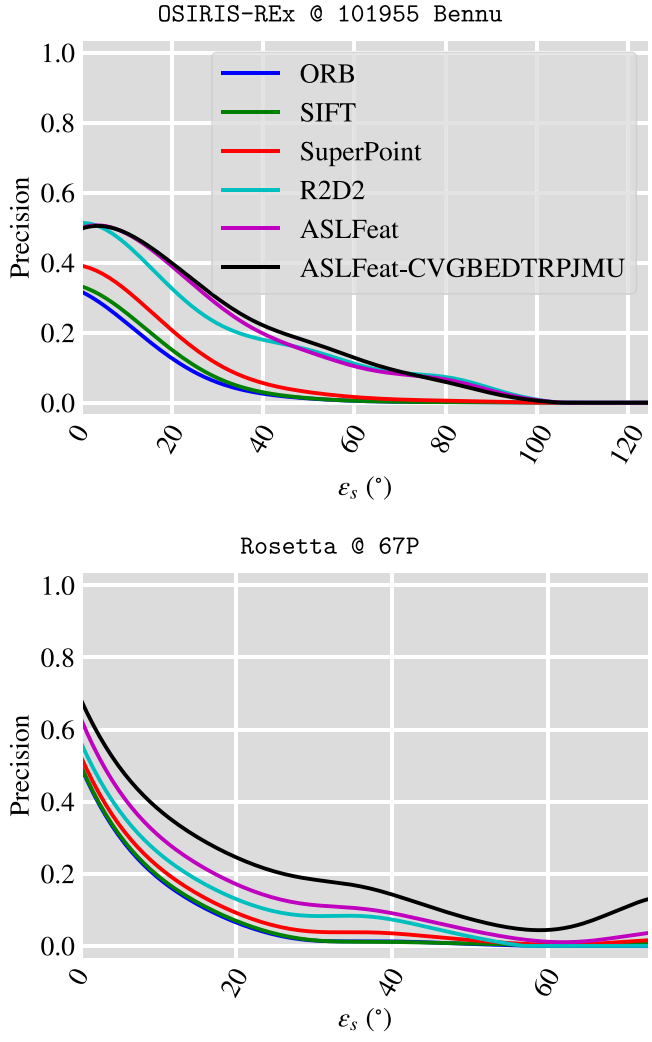


Fig. 11. Illumination change versus precision. Illumination change is measured by ϵ_s as defined in Equation Eq. (24), i.e., the angle between the Sun vectors in the respective cameras.

the Dawn @ 1 Ceres dataset, which has significantly fewer shadowing occlusions as compared to the other datasets. This could indicate that exposing the network to training instances in occluded regions could benefit matching performance. Investigating training strategies that allow the network to effectively learn from occluded matches will be the subject of future work.

7. Conclusion

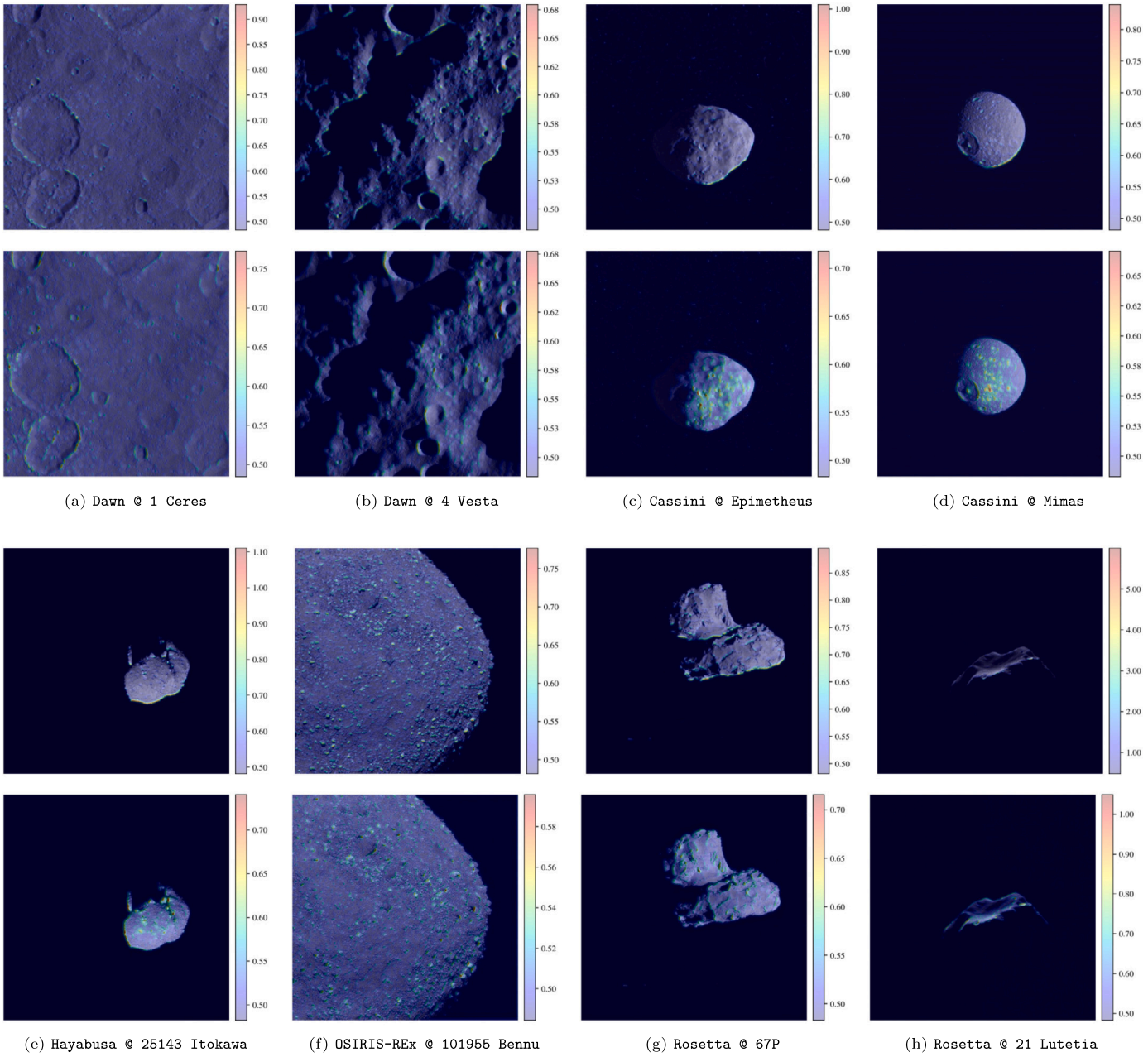
In this paper we presented a first-of-a-kind dataset composed of densely annotated images of small celestial bodies acquired during past and ongoing missions. The AstroVision dataset was leveraged to develop a novel benchmark suite for evaluation of feature detection and description methods on *real* remote imagery of small bodies. Moreover, we showed that leveraging the Astrovision data for training a deep feature detection and description network increases matching and pose estimation performance on small bodies with a wide variety of surface characteristics, including on bodies completely unseen during training. We believe that feature extraction based on deep learning is a promising alternative to current human-in-the-loop practices used in state-of-the-practice small body 3D shape reconstruction methods, e.g., SPC [6]. Furthermore, pending ongoing advancements in space-grade multi-core processors [81–84], deep learning approaches to feature extraction could feasibly be implemented for autonomous relative navigation onboard future spacecraft. Finally, we postulate that the use of AstroVision will extend beyond feature detection and description and enable the deployment of a variety of new deep learning methods for deep space applications, ultimately leading to a significant increase in small body science mission capabilities. The code, data, and trained models will be made available to the public at https://github.com/astrovision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by a NASA Space Technology Graduate Research Opportunity. The authors would like to thank Kenneth Getzandanner and Andrew Liounis from NASA Goddard Space Flight Center for several helpful discussions, and Robert Gaskell from the Planetary Science Institute for providing detailed SPC shape models.



(a) Dawn @ 1 Ceres (b) Dawn @ 4 Vesta (c) Cassini @ Epimetheus (d) Cassini @ Mimas

(e) Hayabusa @ 25143 Itokawa (f) OSIRIS-REx @ 101955 Bennu (g) Rosetta @ 67P (h) Rosetta @ 21 Lutetia

Fig. 12. Detection score maps. Qualitative comparison of detection score maps for the pretrained (top) and ASLFeat-CVGBEDTRPJMU (bottom) models. The color bar indicates the models confidence in the feature corresponding to that pixel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix A. Photometric calibration details

Table A.7 defines the different types of photometric calibration applied to each of the datasets. **Bias** + **Dark** + **Smear** indicates that sensor bias subtraction, dark current (warm pixel) removal, and read-out smear correction have been applied to the images. **Radiometric** indicates radiometric calibration was conducted to convert the raw sensor measurements to units of radiance or reflectance. **Deblurred** refers to applying a deblurring filter to the radiometrically calibrated images. More details can be found in the technical reports for the respective instrumentation: Cassini Imaging Science Subsystem (ISS) [85], Dawn Framing Camera [86], NEAR Multispectral Imager (MSI) [87], OSIRIS-REx Camera Suite (OCAMS) [88], Rosetta NavCam [89], and Mars Express High Resolution Stereo Camera (HRSC) [90].

Appendix B. Matching verification threshold

Inlier and ground truth matches are computed from the putative feature matches of the respective feature detection and description methods via a geometric test that compares projected keypoints of matched features with respect to a distance threshold γ, as described in Section 5.2. We empirically chose a value of $\gamma = 5$ pixels for our experiments, as it was found to be sufficiently strict. However, Fig. B.13

Table 6
Masking ablation study results. Performance of ASLFeat-CVGBEDTRPJMU trained with and without masking with respect to precision (P), recall (R), and accuracy (A). See Section 5.1 for metric definitions.

						AUC		
Dataset	Feature	# Matches	P	R	A	@5°	@10°	@20°
Cassini @ Epimetheus (Saturn XI)[a]	masking	**396**	**28.9**	**27.5**	**74.1**	2.7	8.6	14.0
	w/o masking	391	28.1	25.8	63.1	**2.8**	**8.8**	**14.7**
Cassini @ Mimas (Saturn I)[a]	masking	**328**	**23.6**	**14.9**	**67.1**	0.0	**0.1**	**0.2**
	w/o masking	330	15.0	11.3	57.6	0.0	0.0	0.0
Dawn @ 1 Ceres	masking	1514	52.8	**71.5**	**82.1**	**15.9**	**31.6**	**46.9**
	w/o masking	**1683**	**56.9**	71.3	80.2	13.5	29.0	45.3
Dawn @ 4 Vesta	masking	1412	**70.3**	69.7	**87.4**	**17.5**	**33.0**	**48.7**
	w/o masking	**1494**	63.7	**70.9**	84.9	14.2	28.0	43.2
Hayabusa @ 25143 Itokawa[a]	masking	363	**15.2**	11.0	**53.7**	**2.9**	**5.0**	**8.8**
	w/o masking	**552**	13.8	**11.6**	39.8	2.4	4.5	7.8
OSIRIS-REx @ 101955 Bennu	masking	858	**34.2**	28.4	**79.5**	**6.7**	**12.6**	**19.3**
	w/o masking	**1076**	32.9	**30.1**	75.9	6.0	11.9	18.5
Rosetta @ 67P	masking	837	**30.4**	**23.9**	**69.8**	**4.2**	**7.9**	**13.4**
	w/o masking	**952**	26.4	23.8	61.1	3.4	6.3	10.6
Rosetta @ 21 Lutetia[a]	masking	778	**41.3**	**31.2**	**76.3**	**8.4**	**13.2**	**22.3**
	w/o masking	**943**	33.7	27.9	70.7	2.5	5.7	13.8

[a]No images of this body were included in the training set.

Table A.7
Photometric calibration specifications.

Calibration type	Cassini	Dawn	Hayabusa	Mars Express	NEAR	OSIRIS-REx	Rosetta
Bias + Dark + Smear	✓	✓		✓	✓	✓	✓
Radiometric	✓	✓		✓	✓	✓	✓
Deblurred					✓		

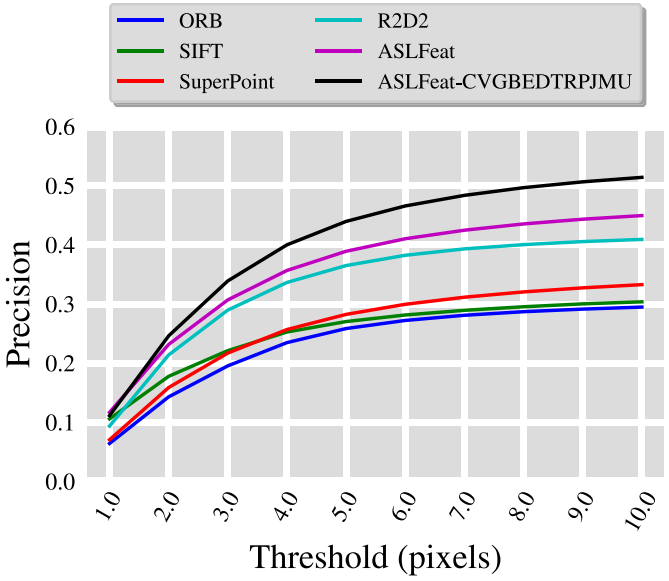


Fig. B.13. Match inlier threshold γ versus precision.

illustrates how the trend in relative precision of the different methods remains relatively unchanged at different inlier thresholds.

References

[1] A.F. Cheng, A.S. Rivkin, P. Michel, J. Atchison, O. Barnouin, L. Benner, N.L. Chabot, C. Ernst, E.G. Fahnestock, M. Kueppers, P. Pravec, E. Rainey, D.C. Richardson, A.M. Stickle, C. Thomas, AIDA DART asteroid deflection test: Planetary defense and science objectives, Planet. Space Sci. 157 (2018) 104–115.

[2] D.D. Mazanek, R.G. Merrill, J.R. Brophy, R.P. Mueller, Asteroid redirect mission concept: A bold approach for utilizing space resources, Acta Astronaut. 117 (2015) 163–171.

[3] A.S. Rivkin, F.E. DeMeo, How many hydrated NEOs are there? J. Geophys. Res. Planets 124 (1) (2019) 128–142.

[4] M.A. Barucci, E. Dotto, A.C. Levasseur-Regourd, Space missions to small bodies: Asteroids and cometary nuclei, Astron. Astrophys. Rev. 19 (48) (2011) 1–29.

[5] C. Norman, C. Miller, R. Olds, C. Mario, E. Palmer, O. Barnouin, M. Daly, J. Weirich, J. Seabrook, C. Bennett, et al., Autonomous navigation performance using natural feature tracking during the OSIRIS-REx touch-and-go sample collection event, Planet. Sci. 3 (101) (2022) 1–21.

[6] R.W. Gaskell, O.S. Barnouin-Jha, D.J. Scheeres, A.S. Konopliv, T. Mukai, S. Abe, J. Saito, M. Ishiguro, T. Kubota, T. Hashimoto, J. Kawaguchi, M. Yoshikawa, K. Shirakawa, T. Kominato, N. Hirata, H. Demura, Characterizing and navigating small bodies with imaging data, Meteorit. Planet. Sci. 43 (6) (2008) 1049–1061.

[7] O. Barnouin, M. Daly, E. Palmer, C. Johnson, R. Gaskell, M. Al Asad, E. Bierhaus, K. Craft, C. Ernst, R. Espiritu, H. Nair, G. Neumann, L. Nguyen, M. Nolan, E. Mazarico, M. Perry, L. Philpott, J. Roberts, R. Steele, J. Seabrook, H. Susorney, J. Weirich, D. Lauretta, Digital terrain mapping by the OSIRIS-REx mission, Planet. Space Sci. 180 (2020) 104764.

[8] E.E. Palmer, R. Gaskell, M.G. Daly, O.S. Barnouin, C.D. Adam, D.S. Lauretta, Practical stereophotoclinometry for modeling shape and topography on planetary missions, Planet. Sci. 3 (102) (2022) 1–16.

[9] P.G. Antreasian, C.D. Adam, K. Berry, J. Geeraert, K.M. Getzandanner, D. Highsmith, J.M. Leonard, E.J. Lessac-Chenen, A.H. Levine, J.V. McAdams, et al., OSIRIS-REx proximity operations and navigation performance at Bennu, in: AIAA SciTech Forum, 2022, pp. 1–34.

[10] S. Bhaskaran, S. Nandi, S. Broschart, M. Wallace, L.A. Cangahuala, C. Olson, Small body landings using autonomous onboard optical navigation, J. Astronaut. Sci. 58 (3) (2011) 1365–1378.

[11] M. Quadrelli, L. Wood, J. Riedel, M. McHenry, M. Aung, L. Cangahuala, R. Volpe, P. Beauchamp, J. Cutts, Guidance, navigation, and control technology assessment for future planetary science missions, J. Guid. Control Dyn. 38 (7) (2015) 1165–1186.

[12] I. Nesnas, B.J. Hockman, S. Bandopadhyay, B.J. Morrell, D.P. Lubey, J. Villa, D.S. Bayard, A. Osmundson, B. Jarvis, M. Bersani, S. Bhaskaran, Autonomous exploration of small bodies toward greater autonomy for deep space missions, Front. Robot. AI 8 (650885) (2021) 1–26.

[13] K.M. Getzandanner, P.G. Antreasian, M.C. Moreau, J.M. Leonard, C.D. Adam, D. Wibben, K. Berry, D. Highsmith, D. Lauretta, Small body proximity operations & TAG: Navigation experiences & lessons learned from the OSIRIS-REx mission, in: AIAA SciTech Forum, 2022, pp. 1–23.

[14] K. Dennison, S. D'Amico, Comparing optical tracking techniques in distributed asteroid orbiter missions using ray-tracing, in: AAS/AIAA Space Flight Mechanics Meeting, 2021, pp. 1–20.

[15] B.J. Morrell, J. Villa, A. Havard, Automatic feature tracking on small bodies for autonomous approach, in: ASCEND, 2020, pp. 1–15.

[16] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. (IJCV) 60 (2) (2004) 91–110.

[17] D. DeTone, T. Malisiewicz, A. Rabinovich, SuperPoint: Self-supervised interest point detection and description, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops, CVPRW, 2018, pp. 337–349.

[18] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, T. Sattler, D2-Net: A trainable CNN for joint description and detection of local features, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 8092–8101.

[19] J. Revaud, C. De Souza, M. Humenberger, P. Weinzaepfel, R2D2: Reliable and repeatable detector and descriptor, in: Advances in Neural Information Processing Systems, NeurIPS, 2019, pp. 1–11.

[20] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, L. Quan, ASLFeat: Learning local features of accurate shape and localization, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 6589–6598.

[21] J. Song, D. Rondao, N. Aouf, Deep learning-based spacecraft relative navigation methods: A survey, Acta Astronaut. 191 (2022) 22–40.

[22] M. Pugliatti, M. Maestrini, P. Di Lizia, F. Topputo, On-board small-body semantic segmentation based on morphological features with U-Net, in: AAS/AIAA Space Flight Mechanics Meeting, 2021, pp. 1–20.

[23] D. Zhou, G. Sun, J. Song, W. Yao, 2D vision-based tracking algorithm for general space non-cooperative objects, Acta Astronaut. 188 (2021) 193–202.

[24] D. Zhou, G. Sun, X. Hong, 3D visual tracking framework with deep learning for asteroid exploration, 2021, ArXiv arXiv:2111.10737.

[25] T. Fuchs, D.R. Thompson, B.D. Bue, J. Castillo-Rogez, S.A. Chien, D. Gharibian, K.L. Wagstaff, Enhanced flyby science with onboard computer vision: Tracking and surface feature detection at small bodies, Earth Space Sci. 2 (10) (2015) 417–434.

[26] H. Lee, H.-L. Choi, D. Jung, S. Choi, Deep neural network-based landmark selection method for optical navigation on Lunar highlands, IEEE Access 8 (2020) 99010–99023.

[27] R. Olds, C. Miller, C. Norman, C. Mario, K. Berry, E. Palmer, O. Barnouin, M. Daly, J. Weirich, J. Seabrook, et al., The use of digital terrain models for natural feature tracking at asteroid Bennu, Planet. Sci. 3 (100) (2022) 1–11.

[28] D.A. Lorenz, R. Olds, A. May, C. Mario, M.E. Perry, E.E. Palmer, M. Daly, Lessons learned from OSIRIS-REx autonomous navigation using natural feature tracking, in: IEEE Aerospace Conf., 2017, pp. 1–12.

[29] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: IEEE Int. Conf. on Computer Vision, ICCV, 2011, pp. 2564–2571.

[30] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, SuperGlue: Learning feature matching with graph neural networks, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 4938–4947.

[31] D. Nistér, An efficient solution to the five-point relative pose problem, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 26 (6) (2004) 756–770.

[32] D. Forsyth, J. Ponce, Computer Vision: A Modern Approach, Second Ed., Prentice Hall, 2011, p. 792, URL https://hal.inria.fr/hal-01063327.

[33] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, J.J. Leonard, Past, present, and future of simultaneous localization and mapping: Toward the Robust-perception age, IEEE Trans. Robot. 32 (6) (2016) 1309–1332.

[34] M. Dor, T. Driver, K. Getzandanner, P. Tsiotras, AstroSLAM: Autonomous monocular navigation in the vicinity of a celestial small body – theory and experiments, 2022, Preprint, arXiv:2202.00350.

[35] T. Lindeberg, Scale-space theory: A basic tool for analyzing structures at different scales, J. Appl. Stat. 21 (1–2) (1994) 225–270.

[36] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-Up Robust Features (SURF), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.

[37] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European Conf. on Computer Vision, ECCV, 2006, pp. 430–443.

[38] M. Calonder, V. Lepetit, C. Strecha, P. Fua, BRIEF: Binary Robust Independent Elementary Features, in: European Conf. on Computer Vision, ECCV, 2010, pp. 778–792.

[39] Y. Ono, E. Trulls, P. Fua, K.M. Yi, LF-Net: Learning local features from images, in: Int. Conf. on Neural Information Processing Systems, NeurIPS, 2018, pp. 6237–6247.

[40] Y. Verdie, K. Yi, P. Fua, V. Lepetit, TILDE: A Temporally Invariant Learned DEtector, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 5279–5288.

[41] K.M. Yi, Y. Verdie, P. Fua, V. Lepetit, Learning to assign orientations to feature points, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 107–116.

[42] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, F. Moreno-Noguer, Discriminative learning of deep convolutional feature point descriptors, in: IEEE Int. Conf. on Computer Vision, ICCV, 2015, pp. 118–126.

[43] K.M. Yi, E. Trulls, V. Lepetit, P. Fua, LIFT: Learned Invariant Feature Transform, in: European Conf. on Computer Vision, ECCV, 2016, pp. 467–483.

[44] K. He, Y. Lu, S. Sclaroff, Local descriptors optimized for average precision, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 596–605.

[45] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets v2: More deformable, better results, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 9308–9316.

[46] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, L. Quan, BlendedMVS: A large-scale dataset for generalized multi-view stereo networks, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 1790–1799.

[47] T. Shen, Z. Luo, L. Zhou, R. Zhang, S. Zhu, T. Fang, L. Quan, Matchable image retrieval by learning from surface reconstruction, in: Asian Conf. on Computer Vision, ACCV, 2018, pp. 415–431.

[48] H. Wang, J. Jiang, G. Zhang, CraterIDNet: An end-to-end fully convolutional neural network for crater detection and identification in remotely sensed planetary images, Remote Sens. 10 (7) (2018) 1067.

[49] S. Silvestrini, M. Piccinin, G. Zanotti, A. Brandonisio, I. Bloise, L. Feruglio, P. Lunghi, M. Lavagna, M. Varile, Optical navigation for Lunar landing based on convolutional neural network crater detector, Aerosp. Sci. Technol. 123 (2022) 107503.

[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobilenetV2: Inverted residuals and linear bottlenecks, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 4510–4520.

[51] A. Silburt, M. Ali-Dib, C. Zhu, A. Jackson, D. Valencia, Y. Kissin, D. Tamayo, K. Menou, Lunar crater identification via deep learning, Icarus 317 (2019) 27–38.

[52] L. Downes, T.J. Steiner, J.P. How, Deep learning crater detection for Lunar terrain relative navigation, in: AIAA SciTech Forum, 2020, pp. 1–12.

[53] C.T. Russell, C.A. Raymond, Dawn mission to Vesta and Ceres, Space Sci. Rev. 163 (2011) 3–23.

[54] R. Park, A. Vaughan, A. Konopliv, A. Ermakov, N. Mastrodemos, J. Castillo-Rogez, S. Joy, A. Nathues, C. Polanskey, M. Rayman, et al., High-resolution shape model of Ceres from stereophotoclinometry using Dawn imaging data, Icarus 319 (2019) 812–827.

[55] R.W. Gaskell, SPC shape and topography of Vesta from Dawn imaging data, in: AAS Division for Planetary Sciences Meeting # 44, 2012, p. 209.03.

[56] M. Dougherty, L. Esposito, S.M. Krimigis, Saturn from Cassini-Huygens, Springer, 2009.

[57] R.W. Gaskell, Gaskell Dione shape model V1.0, NASA Planet. Data Syst. (2020) http://dx.doi.org/10.26033/dh1c-ab91.

[58] R.T. Daly, C.M. Ernst, R.W. Gaskell, O.S. Barnouin, P.C. Thomas, New stereophotoclinometry shape models for irregularly shaped Saturnian satellites, in: Lunar Planet. Sci. Conf., 2018, pp. 1–2.

[59] R.W. Gaskell, Gaskell Mimas shape model V2.0, NASA Planet. Data Syst. (2020) http://dx.doi.org/10.26033/22n9-v553.

[60] R.W. Gaskell, Gaskell Tethys shape model V1.0, NASA Planet. Data Syst. (2020) http://dx.doi.org/10.26033/ebj2-t561.

[61] A. Fujiwara, J. Kawaguchi, D. Yeomans, M. Abe, T. Mukai, T. Okada, J. Saito, H. Yano, M. Yoshikawa, D. Scheeres, et al., The rubble-pile asteroid Itokawa as observed by Hayabusa, Science 312 (5778) (2006) 1330–1334.

[62] Y. Tsuda, T. Saiki, F. Terui, S. Nakazawa, M. Yoshikawa, S.-i. Watanabe, H.P. Team, Hayabusa2 mission status: Landing, roving and cratering on asteroid Ryugu, Acta Astronaut. 171 (2020) 42–54.

[63] R.W. Gaskell, J. Saito, M. Ishiguro, Kubota, T.T. Hashimoto, N. Hirata, S. Abe, O. Barnouin-Jha, Gaskell Itokawa shape model V1.1, NASA Planet. Data Syst. (2021) http://dx.doi.org/10.26033/3b2j-yy57.

[64] J.-P. Bibring, Y. Langevin, J.F. Mustard, F. Poulet, R. Arvidson, A. Gendrin, B. Gondet, N. Mangold, P. Pinet, F. Forget, et al., Global mineralogical and aqueous Mars history derived from OMEGA/Mars Express data, Science 312 (5772) (2006) 400–404.

[65] R.W. Gaskell, Gaskell Phobos shape model V1.0, NASA Planet. Data Syst. (2020) http://dx.doi.org/10.26033/xzv5-bw95.

[66] A.F. Cheng, A. Santo, K. Heeres, J. Landshof, R. Farquhar, R. Gold, S. Lee, Near-earth asteroid rendezvous: Mission overview, J. Geophys. Res. Planets 102 (E10) (1997) 23695–23708.

[67] R.W. Gaskell, Gaskell Eros shape model V1.1, NASA Planet. Data Syst. (2021) http://dx.doi.org/10.26033/d0gq-9427.

[68] D. Lauretta, S. Balram-Knutson, E. Beshore, W. Boynton, C. Drouet d'Aubigny, D. DellaGiustina, H. Enos, D. Golish, C. Hergenrother, E. Howell, et al., OSIRIS-REx: sample return from asteroid (101955) Bennu, Space Sci. Rev. 212 (1) (2017) 925–984.

[69] O. Barnouin, M. Daly, E. Palmer, R. Gaskell, J. Weirich, C. Johnson, M. Al Asad, J. Roberts, M. Perry, H. Susorney, et al., Shape of (101955) Bennu indicative of a rubble pile with internal stiffness, Nat. Geosci. 12 (4) (2019) 247–252.

[70] M. Taylor, N. Altobelli, B. Buratti, M. Choukroun, The Rosetta mission orbiter science overview: The comet phase, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 375 (2097) (2017) 20160262.

[71] R. Schulz, H. Sierks, M. Küppers, A. Accomazzo, Rosetta fly-by at asteroid (21) Lutetia: An overview, Planet. Space Sci. 66 (1) (2012) 2–8.
[72] R.W. Gaskell, L. Jorda, E. Palmer, C. Jackman, C. Capanna, S. Hviid, P. Gutiérrez, Comet 67P/CG: Preliminary shape and topography from SPC, in: AAS/Division for Planetary Sciences Meeting, Vol. 46, 2014, pp. 209–204.
[73] L. Jorda, R.W. Gaskell, M.K.J. Kaasalainen, B. Carry, Rosetta shape model of asteroid Lutetia, NASA Planet. Data Syst. (2013) http://dx.doi.org/10.26007/aajh-r451.
[74] NASA Planetary Data System (PDS), https://pds.nasa.gov/.
[75] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[76] D.Q. Huynh, Metrics for 3D rotations: Comparison and analysis, J. Math. Imaging Vis. 35 (2) (2009) 155–164.
[77] Q.-T. Luong, O.D. Faugeras, The fundamental matrix: Theory, algorithms, and stability analysis, Int. J. Comput. Vis. (IJCV) 17 (1) (1996) 43–75.
[78] P.H. Torr, A. Zisserman, S.J. Maybank, Robust detection of degenerate configurations while estimating the fundamental matrix, Comput. Vis. Image Underst. 71 (3) (1998) 312–333.
[79] C. Choy, J. Park, V. Koltun, Fully convolutional geometric features, in: IEEE Int. Conf. on Computer Vision, ICCV, 2019, pp. 8958–8966.
[80] S. Vassilvitskii, D. Arthur, k-Means++: The advantages of careful seeding, in: ACM-SIAM Sym. on Discrete Algorithms, 2006, pp. 1027–1035.
[81] G. Lentaris, K. Maragos, I. Stratakos, L. Papadopoulos, O. Papanikolaou, D. Soudris, M. Lourakis, X. Zabulis, D. Gonzalez-Arjona, G. Furano, High-performance embedded computing in space: Evaluation of platforms for vision-based navigation, J. Aerospace Inform. Syst. 15 (4) (2018) 178–192.
[82] A.D. George, C.M. Wilson, Onboard processing with hybrid and reconfigurable computing on small satellites, Proc. IEEE 106 (3) (2018) 458–470.
[83] L. Kosmidis, I. Rodriguez, A. Jover, S. Alcaide, J. Lachaize, J. Abella, O. Notebaert, F.J. Cazorla, D. Steenari, GPU4S: Embedded GPUs in space - latest project updates, Microprocess. Microsyst. 77 (103143) (2020) 1–10.
[84] V. Kothari, E. Liberis, N.D. Lane, The final frontier: Deep learning in space, in: Int. Workshop on Mobile Computing Systems and Applications, 2020, pp. 45–49.
[85] B. Knowles, Cassini Imaging Science Subsystem (ISS) Data User's Guide, Tech. rep., Space Science Institute, 2018, URL https://pds-atmospheres.nmsu.edu/data_and_services/atmospheres_data/Cassini/logs/iss_data_user_guide_180916.pdf.
[86] S. Schröder, P. Gutierrez-Marques, Dawn Framing Camera: Calibration Pipeline, Tech. rep. DA-FC-MPAE-RP-272, Planetary Science Institute, 2013, URL https://sbnarchive.psi.edu/pds3/dawn/fc/DWNC7FC2_1B/DOCUMENT/CALIB_PIPELINE/DA-FC-MPAE-RP-272_2_B.PDF.
[87] S. Murchie, M. Robinson, D. Domingue, H. Li, L. Prockter, S. Hawkins, W. Owen, B. Clark, N. Izenberg, Inflight calibration of the NEAR multispectral imager: II. Results from Eros approach and orbit, Icarus 155 (1) (2002) 229–243.
[88] D. Golish, B. Rizk, OSIRIS-REx Camera Suite Calibration Description, Tech. rep., Planetary Science Institute, 2019, URL https://sbnarchive.psi.edu/pds4/orex/orex.ocams/document/ocams_calibration_description_v1.7.pdf.
[89] B. Geiger, M. Barthelemy, R. Andrés, EAICD ROSETTA-NAVCAM, Tech. rep. RO-SGS-IF-0001, European Space Agency, 2020, URL https://pds-smallbodies.astro.umd.edu/holdings/ro-c-navcam-3-prl-mtp003-v1.0/document/ro-sgs-if-0001.pdf.
[90] G. Neukum, R. Jaumann, HRSC: the High Resolution Stereo Camera of Mars Express, Tech. rep. SP-1240, European Space Agency, 2002, URL https://pds-geosciences.wustl.edu/mex/mex-m-hrsc-3-rdr-v3/mexhrs_1001/document/hrsc_esa_sp.pdf.