AstroVision: Towards Autonomous Feature Detection and Description for Missions to Small Bodies Using Deep Learning

Travis Driver\textsuperscript{a,}\textsuperscript{*}, Katherine A. Skinner\textsuperscript{b}, Mehregan Dor\textsuperscript{a}, Panagiotis Tsiotras\textsuperscript{a}

\textsuperscript{a}Georgia Institute of Technology, Atlanta, Georgia, USA
\textsuperscript{b}University of Michigan, Ann Arbor, Michigan, USA

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.

Keywords: Keypoint Detection, Feature Description, Feature Tracking, Deep Learning, Computer Vision, Spacecraft Navigation, Small Bodies

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 \cite{1} and resource utilization \cite{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 \cite{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 \cite{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) \cite{6}. DTM construction typically involves extensive human-in-the-loop verification and carefully designed image acquisition plans to achieve optimal results \cite{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 \cite{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 \cite{10}, e.g., during the OSIRIS-REx Touch-And-Go (TAG) sample collection event \cite{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 \cite{11, 12, 13}.

While automated feature tracking methods have been investigated to reduce reliance on current human-in-the-loop practices for missions to small bodies \cite{14, 15}, these works have focused exclusively on traditional handcrafted features (e.g., SIFT \cite{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 \cite{17, 18, 19, 20}. However, transferring recent advances in deep learning to small body science applications is challenging due to the unavailability of relevant, annotated datasets.

\textsuperscript{*}Corresponding author

Email address: travisdriver@gatech.edu (Travis Driver)

Preprint submitted to Elsevier

3 Aug 2022
data [21]. To the best our knowledge, there exists no large-scale, annotated dataset comprised entirely of real small body images. Indeed, previous work has relied entirely on simulated data [22, 23, 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 undergoes multiple iterations through testing and development of the DTMs [9]. 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 [28, 5]. This process is illustrated in Figure 1a. 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 \( I : \Omega \rightarrow \mathbb{R} \) and \( I' : \Omega' \rightarrow \mathbb{R} \) with pixel domains \( \Omega \subset \mathbb{R}^2 \) and \( \Omega' \subset \mathbb{R}^2 \), respectively. Keypoints \( p_k \in \Omega \) (\( p'_k \in \Omega' \)) localize salient regions in the image, which are typically extracted from a saliency map \( S : \Omega \rightarrow \mathbb{R} \). 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 \( d_k \in \mathbb{R}^d \) referred
to as the descriptor associated with the keypoint \( p_k \). Consider, for instance, corresponding keypoints \( \{p_k\}_{k \in K} \) and \( \{p_{k'}\}_{k' \in K'} \) 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

\[
    d(d_i, d_{i'}) < \min \left( \min_{k \neq l} d(d_i, d_{l'}), \min_{k' \neq l'} d(d_i, d_{l'}) \right)
\]

(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):

\[
    \mathcal{M} := \left\{(l, l') \mid d(d_i, d_{i'}) < \min_{k \neq l} d(d_i, d_{l'}) \right\}
\]

\[
\bigcap \left\{(l, l') \mid d(d_i, d_{i'}) < \min_{k' \neq l'} d(d_i, d_{l'}) \right\} .
\]

(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 Figure 1b. Exploiting recently developed matching approaches based on deep learning \[30\] will be the subject of future work.

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 \( \mathcal{B} \), and let \( \mathcal{C}_i \) denote the camera frame at time index \( i \) with origin \( \mathcal{C}_i \).

Moreover, let \( \ell_i^B \) be the 4th surface landmark expressed in \( \mathcal{B} \), let \( q_i^{\mathcal{C}_i} = \left[q_{x,k}^{\mathcal{C}_i} q_{y,k}^{\mathcal{C}_i} q_{z,k}^{\mathcal{C}_i}\right]^\top \in \mathbb{R}^3 \) denote the vector from \( \mathcal{C}_i \) to the \( k \)th landmark expressed in \( \mathcal{C}_i \), and let \( p_i^{(i)} = \left[u_i^{(i)} v_i^{(i)}\right]^\top \in \mathbb{R}^2 \) denote the 2D image coordinates of the \( k \)th landmark observed by camera \( \mathcal{C}_i \), i.e., the keypoint.

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

\[
    \mathcal{E}_k^{(i)} = \Pi \left( \ell_i^B, \mathcal{C}_i, \mathcal{B} \right) = \frac{1}{d_{k}^C} K q_{k}^{\mathcal{C}_i},
\]

(3)

where \( d_{k}^C \) is the landmark depth in \( \mathcal{C}_i \), \( \mathcal{E}_k^{(i)} = \left[\mathcal{E}_k^x \mathcal{E}_k^y \mathcal{E}_k^z\right]^\top \in \mathbb{R}^3 \) and \( \mathcal{C}_i, \mathcal{B} \in \text{SE}(3) \) denotes the relative pose of \( \mathcal{B} \) with respect to \( \mathcal{C}_i \):

\[
    T_{\mathcal{C}_i, \mathcal{B}} = \begin{bmatrix} R_{\mathcal{C}_i, \mathcal{B}} & r_{\mathcal{C}_i, \mathcal{B}} \ 0_{1 \times 3} & 1 \end{bmatrix}.
\]

(4)

and \( K \) is the camera calibration matrix:

\[
    K = \begin{bmatrix} f_x & 0 & c_x \ 0 & f_y & c_y \ 0 & 0 & 1 \end{bmatrix}
\]

(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 Figure 2. Conversely, a 2D keypoint may be backward-projected into 3D coordinates via

\[
    \mathcal{E}_k^{B} = \Pi^{-1} \left(p_i^{(i)}, \mathcal{C}_i, \mathcal{B}; K\right) = T_{\mathcal{C}_i, \mathcal{B}}^{-1} \left[d_{k}^C K^{-1} p_i^{(i)}\right]
\]

\[
= T_{\mathcal{B}, \mathcal{C}_i} q_{k}^{\mathcal{C}_i}.
\]

(6)
measurements estimate of the poses $T_{\text{frame}}$ [33]. SfM seeks the maximum localization and mapping (SLAM) setting, we are interested in simultaneously estimating a collection of camera poses $T := \{T_{\mathcal{C},\mathcal{B}} \in \text{SE}(3) \mid i = 1, \ldots, m\}$ and a network of landmarks (the map) $\mathcal{L} := \{L_k \in \mathbb{R}^3 \mid k = 1, \ldots, n\}$. Note that the SfM solution is naturally 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 $T$ and landmarks $\mathcal{L}$, given the (independent) keypoint measurements $\mathcal{P} := \{p_k^{(i)} \in \mathbb{R}^2 \mid i = 1, \ldots, m, k = 1, \ldots, n\}:

\begin{equation}
T^*, \mathcal{L}^* = \arg \max_{T, \mathcal{L}} p(T, \mathcal{L} \mid \mathcal{P}) \propto \arg \max_{T, \mathcal{L}} p(T, \mathcal{L}) p(\mathcal{P} \mid T, \mathcal{L}) \tag{9}
\end{equation}

\begin{equation}
= p(T, \mathcal{L}) \prod_{i} \prod_{k} p(p_k^{(i)} \mid T_{\mathcal{C},\mathcal{B}}, \mathcal{L}_k^B). \tag{10}
\end{equation}

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

\begin{equation}
p(p_k^{(i)} \mid T_{\mathcal{C},\mathcal{B}}, \mathcal{L}_k^B) \propto \exp\left\{\|e_k^{(i)} - \Pi(L_k^B, T_{\mathcal{C},\mathcal{B}}; K)\|_2^2 / \Sigma_n^{(i)}\right\}, \tag{12}
\end{equation}

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

\begin{equation}
T^*, \mathcal{L}^* = \arg \min_{T, \mathcal{L}} \sum_{i,k} \|e_k^{(i)} - \Pi(L_k^B, T_{\mathcal{C},\mathcal{B}}; K)\|_2^2 / \Sigma_n^{(i)}. \tag{13}
\end{equation}

where we have omitted the priors $p(T, \mathcal{L})$ for completeness and generality, which can be ignored if no prior information is assumed (i.e., $p(T, \mathcal{L}) = \text{const}$.) 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 pre-determined (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. 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 \cite{29}. ORB is based on Features from Accelerated Segment Test (FAST) detectors \cite{37} and Binary Robust Independent Elementary Features (BRIEF) descriptors \cite{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 \cite{30} \cite{17} \cite{19} \cite{50}. The first data-driven methods focused on individual components of the full image processing pipeline, including keypoint detection \cite{40}, orientation estimation \cite{41}, and feature description \cite{42}. Yi et al. \cite{13} 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 \cite{39} proposed a sequential two-stage approach: the first stage learns keypoint detection and the second stage learns feature description. SuperPoint \cite{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 pre-train 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 \cite{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 \cite{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 \cite{44} between corresponding patch descriptors. ASLFeat \cite{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) \cite{45} to model local geometric variations in the image and learn more transformation invariant features. ASLFeat is trained using the BlendedMVS \cite{16} and GL3D \cite{17} 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 \cite{15} and Dennison et. al \cite{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, 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 \cite{23} \cite{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 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 \cite{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 \cite{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 received much attention, especially for lunar applications. Wang et. al \cite{18} leverages a lightweight CNN architecture pre-trained 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 for localization. Silburt et. al \cite{19} implement a custom U-Net architecture to detect and identify craters from digital elevation maps (DEMs). Lee et al. \cite{20} 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 this technology. 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 Figure 4.

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 novel 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) [21] and maintained by NASA’s Navigation and Ancillary Information Facility. 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 [16]. 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 [1], [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 (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 Figure 4. The landmark map provides a consistent, discrete set of tie-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 rays¹ 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 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 used. 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 [72] 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

¹Ray tracing uses the Trimesh library: https://trimsh.org/
Figure 3: **AstroVision image datasets.** Shape model references are provided in Table 1.
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):

\[
\text{precision} = \frac{\# \text{ correct matches}}{\# \text{ putative matches}}.
\]

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

\[
\text{recall} = \frac{\# \text{ correct matches}}{\# \text{ ground truth matches}}.
\]

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

\[
\text{accuracy} = \frac{\# \text{ correct matches & nonmatches}}{\# \text{ features}}.
\]

We classify correct nonmatches as keypoints which 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. 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 \( q_{C_jC_i} \) and ground truth \( q_{C_jC_i} \) relative orientation quaternions

\[
\epsilon_q := \cos^{-1}((q_{C_jC_i} q^{C_iC_j})^2 - 1)
\]

is used as a metric for the orientation error, and

\[
\epsilon_t := \cos^{-1}\left(\frac{r_{C_jC_i} \cdot r^{C_iC_j}}{\|r_{C_jC_i}\| \|r^{C_iC_j}\|}\right)
\]

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 \( \epsilon \) 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. Three state-of-the-art data-driven features were selected that leverage different learning approaches (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 detect 5,000 keypoints and descriptors.

Given a set of keypoints and descriptors, putative matches are computed using MNM. Matches are verified by first backward-projecting (via Equation (6)) each key-

point in the first image into 3D world coordinates using the ground truth calibration and depth map. The 3D points are then forward-projected (via Equation (3)) into the second image, and matches are verified by checking that the projected image coordinates are within some distance \( \gamma \) to the keypoint of its matched feature, where we empirically chose a value of \( \gamma = 5 \) pixels. Ground truth matches are
| Dataset | Feature | AUC | @5° | @10° | @20° |
|---------|---------|-----|-----|------|------|
| Cassini @ Epimetheus (Saturn XI) | ORB | 133 | 295 | 9.2 | 14.9 |
| (326.7 m/pixel, 255.4 m/pixel) | SIFT | 204 | 32.5 | 15.0 |
| | SuperPoint | 396 | 13.6 | 12.8 |
| | R2D2 | 223 | 25.3 | 14.7 |
| | ASLFeat | 386 | 27.4 | 13.7 |
| Cassini @ Mimas (Saturn I) | ORB | 307 | 843 | 0.0 | 0.1 |
| (1,176.2 m/pixel, 943.8 m/pixel) | SIFT | 340 | 14.3 | 0.4 |
| | SuperPoint | 121 | 8.6 | 0.0 |
| | R2D2 | 209 | 13.8 | 0.1 |
| | ASLFeat | 372 | 21.8 | 0.3 |
| Dawn @ 1 Ceres | ORB | 3624 | 1437 | 3.7 | 15.9 |
| (122.0 m/pixel, 35.4 m/pixel) | SIFT | 1656 | 42.8 | 56.6 |
| | SuperPoint | 442 | 13.1 | 43.5 |
| | R2D2 | 954 | 50.0 | 32.4 |
| | ASLFeat | 1555 | 48.4 | 42.4 |
| Dawn @ 4 Vesta | ORB | 2006 | 1465 | 2.6 | 11.1 |
| (63.3 m/pixel, 21.2 m/pixel) | SIFT | 1350 | 17.9 | 38.8 |
| | SuperPoint | 506 | 11.3 | 34.1 |
| | R2D2 | 926 | 55.9 | 46.7 |
| | ASLFeat | 1526 | 59.1 | 46.0 |
| Hayabusa @ 25143 Itokawa | ORB | 603 | 767 | 0.9 | 2.9 |
| (95.5 cm/pixel, 78.7 cm/pixel) | SIFT | 217 | 1.9 | 4.8 |
| | SuperPoint | 79 | 1.7 | 5.4 |
| | R2D2 | 339 | 10.7 | 8.0 |
| | ASLFeat | 338 | 13.5 | 7.6 |
| OSIRIS-REx @ 101955 Bennu | ORB | 1789 | 1581 | 0.3 | 1.6 |
| (21.9 cm/pixel, 9.9 cm/pixel) | SIFT | 1317 | 4.9 | 11.8 |
| | SuperPoint | 747 | 5.6 | 11.1 |
| | R2D2 | 502 | 29.3 | 13.8 |
| | ASLFeat | 1378 | 33.1 | 20.9 |
| Rosetta @ 67P | ORB | 3039 | 1426 | 0.2 | 1.6 |
| (5.5 m/pixel, 2.4 m/pixel) | SIFT | 1168 | 15.7 | 7.7 |
| | SuperPoint | 485 | 1.6 | 6.4 |
| | R2D2 | 634 | 20.2 | 7.1 |
| | ASLFeat | 1147 | 25.0 | 10.6 |
| Rosetta @ 21 Lutetia | ORB | 40 | 486 | 1.3 | 4.3 |
| (230.5 m/pixel, 228.1 m/pixel) | SIFT | 238 | 12.1 | 15.9 |
| | SuperPoint | 381 | 26.7 | 16.2 |
| | R2D2 | 588 | 33.2 | 13.3 |
| | ASLFeat | 970 | 42.9 | 23.8 |
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 \([21]\), 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 2\( N \) 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 algorithms. These results are summarized in Table 2 and qualitative comparisons are provided in Figure 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. 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 Figure 5a although this behavior does not seem to be typical (see Figure 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 ASLFeat network for end-to-end training using the AstroVision data products.

This is detailed in the next section.

We recognize the very 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 estimates given even small amounts of measurement noise. Indeed, the Cassini @ Mimas images have a mean GSD of 1,176.2 m/pixel, 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 \([74]\), special surfaces which yield multiple essential matrix estimates that satisfy Equation (19) Detection (e.g., via the iterative method presented in Figure 4) 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. 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 \( D \in \mathbb{R}^{h \times w \times d} \).

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

\[
\beta_{ij}^{c} = \text{softplus} \left( y_{ij}^{c} - \frac{1}{b} \sum_{\ell} y_{ij}^{\ell} \right)
\]

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

\[
\alpha_{ij}^{c} = \text{softplus} \left( y_{ij}^{c} - \frac{1}{|N(i, j)|} \sum_{(i', j') \in N(i, j)} y_{i'j'}^{c} \right)
\]

where \( N(i, j) \) is the set of 9 neighbors of the pixel \((i, j) \) (including itself). The elements of the \( \ell \)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

\[
S = \frac{1}{\sum_{\ell} w_{\ell}} \sum_{\ell} w_{\ell} S^{(\ell)},
\]

where the weights \( w_{1}, w_{2}, w_{3} \) have been empirically set to 1, 2, 3, respectively.

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

\[
L(D, D', S; \gamma, \rho, \alpha, \beta; M) = \frac{1}{|M|} \sum_{(I, I') \in M} \sum_{(k, k') \in M} s_{k} s'_{k'} \|d_{k} - d'_{k'}\|
\]

where \( s_{k} (s'_{k'}) \) is the detection score and \( d_{k} (d'_{k'}) \) is the descriptor at keypoint \( p_{k} (p'_{k'}) \), and \( m(\cdot, \cdot) \) is the descriptor...
Figure 5: Qualitative comparison of feature matching. Correct matches are drawn in green, and the keypoints of incorrect matches are drawn in red.
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 [76] to enforce descriptor reliability:

$$m(d_i, d'_j) = \max(||d_i - d'_j|| - M_p, 0) + \max(M_n - \min_{k \neq j} ||d_i - d'_k||, \min_{k \neq j} ||d_k - d'_j||),$$

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

The formulated loss $L$ in Equation (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.

### 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 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 Equation (17), is used as a metric for the relative perspective change between two images. We ignore image pairs with a value greater than $\epsilon_q(q_{C,G}, q_{C',G}) = 60^\circ$. The training consumes $\sim$800k image pairs resized to $480 \times 480$ using a batch size of 2. Ground truth matches for training are computed by first querying the landmark map for sparse correspondences. Dense matching is performed on image pairs with at least 128 shared landmarks by projecting a uniform grid of coordinates in the first image into the second image using the ground truth depth and calibration. Additionally, the visibility masks are used to remove matches that have keypoints in occluded regions of either image. 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 using a two-stage training procedure as suggested by [20].

**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 [10]. The descriptors are then bilinearly interpolated at the 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 [76] to enforce descriptor reliability:

$$m(d_i, d'_j) = \max(||d_i - d'_j|| - M_p, 0) + \max(M_n - \min_{k \neq j} ||d_i - d'_k||, \min_{k \neq j} ||d_k - d'_j||),$$

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

The formulated loss $L$ in Equation (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 |
deed, 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 Figure 8g). 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 Figure 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 Figure 8g.

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 Figures 10 and 11. We leverage Equation (17) as a measure for perspective change, and

$$\epsilon_s := \cos^{-1}(\mathbf{s}_i^C \cdot \mathbf{s}_j^C)$$

(24)

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

| Dataset | # Images | Feature | # Matches | P   | R   | A   |
|---------|----------|---------|-----------|-----|-----|-----|
|         |          |         |           |     |     | 0°  | 10° | 20° |
| Cassini @ Epimetheus (Saturn XI)† | 133 | ASLFeat | 386 | 27.4 | 29.0 | 74.7 | 2.7 | 8.2 | 13.7 |
|         |          |         |           |     |     |     | 2.7 | 8.6 | 14.0 |
|         |          | ASLFeat-CVGBEDTRPJMU | 396 | 28.9 | 27.5 | 74.1 |     |     |     |
| Cassini @ Mimas (Saturn I)† | 307 | ASLFeat | 372 | 21.8 | 15.7 | 65.3 | 0.2 | 0.2 | 0.3 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | 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 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | 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 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | ASLFeat-CVGBEDTRPJMU | 1412 | 70.3 | 69.7 | 87.4 | 17.5 | 33.0 | 48.7 |
| Hayabusa @ 25143 Itokawa† | 603 | ASLFeat | 338 | 13.5 | 11.3 | 47.5 | 2.2 | 4.2 | 7.6 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | 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 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | 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 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | ASLFeat-CVGBEDTRPJMU | 837 | 30.4 | 23.9 | 69.8 | 4.2 | 7.9 | 13.4 |
| Rosetta @ 21 Lutetia† | 561 | ASLFeat | 748 | 43.1 | 31.1 | 76.3 | 8.4 | 13.2 | 22.3 |
|         |          |         |           |     |     | 0°  | 10° | 20° |
|         |          | ASLFeat-CVGBEDTRPJMU | 778 | 41.3 | 31.1 | 76.3 | 8.4 | 13.2 | 22.3 |

† No images of this body were included in the training set

Table 5: **ASLFeat-B Benchmark performance.** Performance of ASLFeat-B, i.e., ASLFeat trained on OSIRIS-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.

| Dataset | # Matches | P   | R   | A   | AUC |
|---------|-----------|-----|-----|-----|-----|
|         |           |     |     | 0°  | 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 | 1031 | 49.8 | 61.7 | 76.3 | 12.2 | 26.4 | 41.6 |
| Dawn @ 4 Vesta | 1430 | 48.2 | 54.9 | 79.2 | 11.7 | 23.0 | 34.9 |
| Hayabusa @ 25143 Itokawa | 337 | 9.6 | 6.9 | 43.1 | 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 | 67.1 | 1.0 | 3.3 | 9.2 |

Figure 9: Precision versus number of matches on the OSIRIS-REx @ 101955 Bennu test set.
Figure 10: **Perspective change versus precision.** Perspective change is measured by $\epsilon(q_{E_iB}, q_{E_jB})$ as defined in Equation (17), i.e., the minimum rotation angle between the respective cameras.

Figure 11: **Illumination change versus precision.** Illumination change is measured by $\epsilon_s$ as defined in Equation (24), i.e., the angle between the Sun vectors in the respective cameras.

Figure 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 was 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 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
Figure 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.

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 [78, 79, 80, 81], 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.
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 \[\text{S1}\] for metric definitions.

| Dataset | Feature | masked | P | R | A  |
|---------|---------|--------|---|---|----|
| Cassini @ Epimetheus (Semi-3D) | masking | 0.11 | 0.11 | 0.10 |
| | w/o masking | 0.11 | 0.11 | 0.10 |
| Cassini @ Minus (Semi-3D) | masking | 0.12 | 0.12 | 0.10 |
| | w/o masking | 0.12 | 0.12 | 0.10 |
| Dawn @ 1 Ceres | masking | 0.13 | 0.13 | 0.12 |
| | w/o masking | 0.12 | 0.12 | 0.10 |
| Dawn @ 1 Vesta | masking | 0.12 | 0.12 | 0.10 |
| | w/o masking | 0.12 | 0.12 | 0.10 |
| Hayabusa @ 25143 Itokawa† | masking | 0.12 | 0.12 | 0.10 |
| | w/o masking | 0.12 | 0.12 | 0.10 |
| OSIRIS-REx @ 101955 Bennu masking | masking | 0.12 | 0.12 | 0.10 |
| | w/o masking | 0.12 | 0.12 | 0.10 |
| Rosetta @ 47P | masking | 0.12 | 0.12 | 0.10 |
| | w/o masking | 0.12 | 0.12 | 0.10 |
| Rosetta @ 21 Lutetia† | masking | 0.12 | 0.12 | 0.10 |
| | w/o masking | 0.12 | 0.12 | 0.10 |

† No images of this body were included in the training set.

Acknowledgements

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

Appendix A. Photometric Calibration Details

Table \[\text{A.1}\] 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 readout smear correction have been applied to the images. Radiometric indicates radometric calibration was conducted to convert the raw sensor measurements to units of radiance or reflectance. Deblurring refers to applying a deblurring filter to the radiometrically calibrated images. More details can be found in the technical reports for the respective instrument: Cassini Imaging Science Subsystem (ISS) \[\text{S2}\], Dawn Framing Camera \[\text{S3}\], NEAR Multispectral Imager (MSI) \[\text{S4}\], OSIRIS-REx Camera Suite (OCAMS) \[\text{S5}\], Rosetta NavCam \[\text{S6}\], and Mars Express High Resolution Stereo Camera (HRSC) \[\text{S7}\].

References

[1] A. F. Cheng, A. S. Rivkin, P. Michel, J. Atchison, O. Barnoun, L. Benner, N. L. Chabot, C. Ernst, E. G. Fahnstock, M. Kueppers, P. Pravec, E. Rainey, D. C. Richardson, A. M. Stickel, C. Thomas, AIDA DART asteroid deflection test: Planetary defense objectives, Planetary and Space Science 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 Astronautica 117 (2015) 163–171.

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

[4] M. A. Barucci, E. Dotto, A. C. Lesuisse-Rougourd, Space missions to small bodies: asteroids and cometary nuclei, Astronomy and Astrophysics Review 19 (48) (2011) 1–29.

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

[6] R. W. Gaskell, O. S. Barnoun-jha, D. J. Schoeers, A. S. Konopliv, T. Mukai, S. Ake, 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, Meteoritics & Planetary Science 43 (6) (2008) 1049–1061.

[7] O. Barnoun, M. Daly, E. Palmer, C. Johnson, R. Gaskell, M. Al Asad, E. Bierhaus, K. Craft, C. Ernst, R. Espirito, H. Nair, G. Neumann, L. Nguyen, M. Nolan, E. Mazarico, M. Perry, L. Philipott, J. Roberts, R. Steele, J. Seabrook, H. Susorsey, J. Weirich, D. Laurentta, Digital terrain mapping by the OSIRIS-REx mission, Planetary and Space Science 180 (2020) 104764.

[8] E. E. Palmer, R. Gaskell, M. G. Daly, O. S. Barnoun, C. D. Adam, D. S. Laurentta, Practical stereophotoluminometry for modeling shape and topography on planetary missions, Planetary Science 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, N. Sandi, S. Broschart, M. Wallace, L. A. Canghuala, C. Olson, Small body on-board autonomous navigation, J. of the Astronautical Sciences 58 (3) (2011) 1365–1378.

[11] M. Quadrelli, L. Wood, J. Riedel, M. McHenry, M. Aung, L. Canghuala, R. Volpe, P. Beauchamp, J. Cutts, Guidance, navigation, and control technology assessment for future planetary science missions, J. of Guidance, Control, and Dynamics 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, Frontiers in Robotics and AI 8 (650885) (2021) 1–26.

[13] K. M. Getzandanner, P. G. Antreasian, M. C. Moreau, J. M. Leonard, C. D. Adam, D. Wilben, K. Berry, D. Highsmith, D. Laurentta, 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. of Computer Vision (IJCV) 60 (2) (2004) 91–110.

[17] D. DeTone, T. Malisiewicz, A. Rabinovich, SuperPoint: Self-supervised interest point detection and description, in: Computer Vision and Pattern Recognition Workshop (CVPRW), 2018, pp. 337–349.

[18] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, T. Satller, 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,
| Calibration type                  | Cassini | Dawn | Hayabusa | Mars Express | NEAR | OSIRIS-REx | Rosetta |
|----------------------------------|---------|------|----------|--------------|------|------------|---------|
| Bias + Dark + Smear              | ✓       | ✓    | ✓        | ✓            | ✓    | ✓          | ✓       |
| Radiometric                     | ✓       | ✓    | ✓        | ✓            | ✓    | ✓          | ✓       |
| Deblurred                       | ✓       | ✓    | ✓        | ✓            | ✓    | ✓          | ✓       |

Table A.7: Photometric calibration specifications.
