On-ground validation of a CNN-based monocular pose estimation system for uncooperative spacecraft

Bridging domain shift in rendezvous scenarios

Pasqualetto Cassinis, Lorenzo; Menicucci, Alessandra; Gill, Eberhard; Ahrns, Ingo; Sanchez-Gestido, Manuel

DOI
10.1016/j.actaastro.2022.04.002

Publication date
2022

Document Version
Final published version

Published in
Acta Astronautica

Acta Astronautica 196 (2022) 123–138

Contents lists available at ScienceDirect

Acta Astronautica

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

On-ground validation of a CNN-based monocular pose estimation system for uncooperative spacecraft: Bridging domain shift in rendezvous scenarios

Lorenzo Pasqualetto Cassinis [a,*], Alessandra Menicucci [a], Eberhard Gill [a], Ingo Ahrns [b], Manuel Sanchez-Gestido [c]

[a] Delft University of Technology, Kluyverweg 1 2629 HS, Delft, The Netherlands
[b] Airbus DS GmbH, Airbusallee 1, 28199, Bremen, Germany
[c] ESTEC, Keplerlaan 1, 2201 AZ, Noordwijk, The Netherlands

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Relative pose estimation
Active Debris Removal
In-orbit servicing
On-ground validation
Convolutional Neural Networks
Domain adaptation

ABSTRACT

The estimation of the relative pose of an inactive spacecraft by an active servicer spacecraft is a critical task for close-proximity operations, such as In-Orbit Servicing and Active Debris Removal. Among all the challenges, the lack of available space images of the inactive satellite makes the on-ground validation of current monocular camera-based navigation systems a challenging task, mostly due to the fact that standard Image Processing (IP) algorithms, which are usually tested on synthetic images, tend to fail when implemented in orbit. In response to this need to guarantee a reliable validation of pose estimation systems, this paper presents the most recent advances of ESA's GNC Rendezvous, Approach and Landing Simulator (GRALS) testbed for close-proximity operations around uncooperative spacecraft. The proposed testbed is used to validate a Convolutional Neural Network (CNN)-based monocular pose estimation system on representative rendezvous scenarios with special focus on solving the domain shift problem which characterizes CNNs trained on synthetic datasets when tested on more realistic imagery. The validation of the proposed system is ensured by the introduction of a calibration framework, which returns an accurate reference relative pose between the target spacecraft and the camera for each lab-generated image, allowing a comparative assessment at a pose estimation level. The VICON Tracker System is used together with two KUKA robotic arms to respectively track and control the trajectory of the monocular camera around a scaled 1:25 mockup of the Envisat spacecraft. After an overview of the facility, this work describes a novel data augmentation technique focused on texture randomization, aimed at improving the CNN robustness against previously unseen target textures. Despite the feature detection challenges under extreme brightness and illumination conditions, the results on the high exposure scenario show that the proposed system is capable of bridging the domain shift from synthetic to lab-generated images, returning accurate pose estimates for more than 50% of the rendezvous trajectory images despite the large domain gaps in target textures and illumination conditions.

1. Introduction

Nowadays, the safety and operations of satellites in orbit have become paramount for key Earth-based applications, such as remote sensing, navigation, and telecommunication. In this context, advancements in the field of Guidance, Navigation, and Control (GNC) were made in the past years to cope with the challenges involved in In-Orbit Servicing (IOS) and Active Debris Removal (ADR) missions [1,2]. For such scenarios, the estimation of the relative pose (position and attitude) of an uncooperative target object by an active servicer spacecraft represents a critical task. Compared to cooperative close-proximity missions, the pose estimation problem is complicated by the fact that the target object is not functional and/or not able to aid the relative navigation. Hence, optical sensors on the servicer spacecraft shall be preferred over Radio Frequency (RF) sensors to cope with a lack of navigation devices such as Global Navigation Satellite Systems (GNSS) sensors and/or antennas on the target.

From a high-level perspective, optical sensors can be divided into active and passive devices, depending on whether they require power to function, i.e. LIght Detection And Ranging (LIDAR) sensors and Time-Of-Flight (TOF) cameras, or if they passively acquire light, i.e. monocular and stereo cameras. Spacecraft relative navigation usually exploits

* Corresponding author.
E-mail addresses: L.PasqualettoCassinis@tudelft.nl (L. Pasqualetto Cassinis), A.Menicucci@tudelft.nl (A. Menicucci), E.K.A.Gill@tudelft.nl (E. Gill), ingo.ahrns@airbus.com (I. Ahrns), Manuel.Sanchez.Gestido@esa.int (M. Sanchez-Gestido).

https://doi.org/10.1016/j.actaastro.2022.04.002
Received 20 December 2021; Received in revised form 16 February 2022; Accepted 4 April 2022
Available online 10 April 2022

Electro-Optical (EO) sensors such as stereo cameras [3,4] and/or a LIDAR sensor [5] in combination with one or more monocular cameras, in order to overcome the partial observability that results from the lack of range information in monocular-based systems [6]. In this framework, pose estimation systems based solely on a monocular camera are recently becoming an attractive alternative to systems based on active sensors or stereo cameras, due to their reduced mass, power consumption and system complexity [7,8]. However, a significant effort is still required to comply with most of the demanding requirements for a robust and accurate monocular-based relative navigation system. Notably, the aforementioned navigation system cannot rely on known visual markers, as they are typically not available on an uncooperative target. Since the extraction of visual features is an essential step in the pose estimation process, advanced Image Processing (IP) techniques are required to extract keypoints (or interest points), corners, and/or edges on the target body. In model-based methods, the detected features are then matched with pre-defined features on an offline wireframe 3D model of the target to solve for the relative pose. This is usually achieved by solving the Perspective-n-Points (PnP) problem [9]. In other words, a reliable detection of key features is critical to guarantee safe operations around an uncooperative target, e.g. under adverse orbital conditions.

Unfortunately, standard IP algorithms usually lack of feature detection robustness when applied to space images [10], undermining the overall navigation system and, in turn, the close-proximity operations around the uncooperative target. From a pose initialization standpoint, the extraction of target features can in fact be jeopardized by external factors, such as adverse illumination conditions, low Signal-to-Noise ratio (SNR) and Earth in the background, as well as by target-specific factors, such as the presence of complex textures and features on the target body. Moreover, most of the IP methods are based on the image gradient, detecting textured-rich features or highly visible parts of the target silhouette. As such, the detected features are image-specific and can vary in number and typology depending on the image histogram. This means that most of these techniques cannot accommodate an offline feature selection step, which necessitates a computationally expensive image-to-model correspondence process to ensure that each detected 2D feature is matched with its 3D counterpart on the available wireframe model of the target object.

In recent years, Convolutional Neural Networks (CNNs) are emerging as a valid and robust alternative to more traditional monocular-based pose estimation systems, with two main CNN-based architectures currently being investigated. Initially, *end-to-end* architectures in which a single CNN replaced the entire pose estimation pipeline were more adopted [11–14]. However, since the pose accuracies of these systems proved to be lower than the accuracies returned by standard PnP solvers, especially in the estimation of the relative attitude [11], keypoints-based architectures stood out as the preferred option. Specifically, average orientation errors of $1.31° \pm 2.24°$ were achieved by keypoints-based methods as opposed to the average orientation errors of $9.76° \pm 18.51°$ achieved by end-to-end methods. These averages were computed across test images of the TANGO spacecraft as part of the Spacecraft Pose Estimation Dataset (SPEED) challenge [15,16]. In keypoints-based CNN systems, a CNN is used only at a feature detection level to replace standard IP algorithms, and the output features are fed to a PnP solver together with their body coordinates, which are made available through the wireframe 3D model of the target body. Due to the fact that the trainable features can be selected offline prior to the training, the matching of the extracted feature points with the features of the wireframe model can be performed without the need of a large search space for the image-model correspondences, which usually characterizes most of the edges/corners-based methods [10]. However, due to a lack of availability of representative space images, these CNN systems often need to be trained with synthetic renderings of the available target model. As a result, their feature detection robustness on more realistic images is usually unknown and difficult to predict. In other words, the synthetic datasets used to train the CNNs tend to fail in representing the textures of the target mockup as well as the external illuminations, resulting in inaccurate detections and low pose estimation accuracies [15,17]. In this context, two desirable aspects stand out: First, a proper on-ground validation framework shall be sought to test the CNNs robustness against representative images of the target spacecraft, generated in a laboratory environment which recreates space-like conditions. Notably, a calibration framework shall be established which returns an accurate reference for the relative pose between the monocular camera and the target mockup for each generated image (pose labels), in order to be able to quantify the CNN performance at both keypoints detection and pose estimation levels. Second, novel techniques shall be investigated to improve the performance of CNNs trained using synthetic images on actual space imagery. This aspect is referred to as the *domain shift problem* [18].

Several laboratory testbeds exist to generate images of a target spacecraft's mockup with a monocular camera [19], e.g. the Space Rendezvous Laboratory (SLAB) at Stanford University [15], the Orbital Robotics & GNC laboratory (ORGL) at the European Space Research and Technology Centre (ESTEC) [20], and the Testbed for Robotic Optical Navigation (TRON) at the German Aerospace Agency (DLR) [21]. However, only a few detailed calibration procedures were recently described which allow the accurate estimation of the reference relative pose between camera and target [22,23]. Moreover, the calibration of the target spacecraft highly depends on the presence (cooperative target) or not (uncooperative target) of visual markers, as well as on the rendezvous trajectory that shall be recreated (static or rotating target). Above all, the challenges in recreating illumination conditions, together with the laboratory constraints on the robot movements, are retained as the main limiting factors in the recreation of realistic rendezvous scenarios. Despite recent efforts aimed at extending the capability to recreate almost any camera-target relative pose on-ground with highly accurate pose labels [23], there is still the need to extend the capabilities of on-ground validation setups to allow the recreation of representative rendezvous trajectories.

In relation to the domain shift problem in CNNs, various works have been carried out in recent years to leverage the domain shift from synthetic training to real test imagery, either via *data augmentation* [18,24,25] or via *domain adaptation* [26,27]. Although domain adaptation techniques are often effective and can produce impressive results by adapting the CNN on a specific target domain post training, they require some images of the new domain to adapt to, and hence they are not domain-agnostic. As such, domain adaptation is not well suited for generalizing the CNN performance to many potential target domains, which could be the case in ADR missions. On the other hand, data augmentation techniques consist of introducing variations in the synthetic training domain without any a-priori knowledge of the target domain. In essence, the idea is to extend the standard data augmentation effects, such as random cropping, zooming, rotation, flipping etc. with texture and complex illumination variations. By doing that, Tobin et al. [18] already showed that a CNN can generalize from synthetic environments to new domains by using an unrealistic but diverse set of random textures. Following this line of reasoning, Jackson et al. [24] and Geirhos et al. [25] further discovered that by randomizing textures during training, CNNs can learn the shape of objects rather than textures, improving their robustness to domain shift.

Despite promising results on terrestrial applications, the domain shift problem is still a complex and unexplored topic in space, mostly due to the challenges in recreating representative space-like scenarios on-ground. Although recent works investigated the impact of simple training augmentation on the CNN performance on the SPEED lab-generated images [28,29], the laboratory domain was tuned to not differ too much from the synthetic domain. Furthermore, the mockup of the target spacecraft used to generate the lab-images did not differ considerably from the CAD model adopted during synthetic rendering, leading to relatively small domain variations. In an attempt to

assess the performance of a CNN-based pose estimation system in more challenging scenarios, the authors [17] further investigated the impact of texture randomization on domain adaptation by recreating more adverse illumination conditions and by allowing deviations between the target mockup and the CAD model used for synthetic rendering. However, only a small number of static images were generated, leading to a very limited validation. Moreover, the challenges of recreating realistic rendezvous trajectories were not tackled, and the adopted calibration procedure was not deemed accurate enough to return a reliable ground truth of the relative pose.

Building on the authors' previous findings [17,30] and inspired by the promising texture randomization results presented in earlier works [28], the main objective of this paper is to investigate the impact of training data augmentation on the CNN performance on representative space imagery generated on-ground. In order to do so, special focus is put on the recreation of a dedicated calibration pipeline to validate the proposed pose estimation system on representative rendezvous scenarios. The main contributions of this work are:

1. To propose a novel CNN training augmentation pipeline focused on texture randomization.
2. To improve the on-ground validation capabilities of the GRALS testbed towards the recreation of representative rendezvous trajectories.
3. To assess the performance of the proposed CNN-based system under challenging domain shifts.

The paper is organized as follows. Section 2 introduces the proposed pose estimation framework. The laboratory setup and the calibration procedure are described in Sections 3 and 4. In Section 5, the CNN training, validation and testing phases are detailed, with special focus to the augmentation and randomization pipeline. Next, the pose estimation results are presented in Section 6. Finally, Section 7 provides the main conclusions and recommendations.

2. Pose estimation framework

This work considers a servicer spacecraft flying relative to a target spacecraft located in a Low Earth Orbit (LEO), with the relative motion being described in a Local Vertical Local Horizontal (LVLH) reference frame co-moving with the servicer (Fig. 1a). Furthermore, it is assumed that the servicer is equipped with a single monocular camera. The relative attitude of the target with respect to the servicer can then be defined as the rotation of the target body-fixed frame B with respect to the servicer camera frame C, where these frames are tied to each spacecraft's body. The vector from the camera origin to the target origin defines their relative position. Together, these two quantities characterize the relative pose. This information can then be transformed from the camera frame to the servicer's center of mass by accounting for the relative pose of the camera with respect to the LVLH frame.

From a high-level perspective, a model-based monocular pose estimation system receives as input a 2D image and matches it with an existing wireframe 3D model of the target spacecraft to estimate the target pose with respect to the servicer camera. As illustrated in Fig. 1b, the pose estimation problem consists in determining the position of the target's center of mass $\mathbf{t}^C$ and its attitude with respect to the camera frame C, represented by the rotation matrix $\mathbf{R}_B^C$. The Perspective-n-Points (PnP) equations,

$$\mathbf{r}^C = \begin{pmatrix} x^C & y^C & z^C \end{pmatrix}^T = \mathbf{R}_B^C \mathbf{r}^B + \mathbf{t}^C \tag{1}$$

$$\mathbf{p} = \left(u_i, v_i\right) = \left(\frac{x^C}{z^C} f_x + C_x, \frac{y^C}{z^C} f_y + C_y\right), \tag{2}$$

relate the unknown pose with a feature point $\mathbf{p}$ in the image plane via the relative position $\mathbf{r}^C$ of the feature with respect to the camera frame. Here, $\mathbf{r}^B$ is the point location in the 3D model, expressed in the body-frame coordinate system B, whereas f_x and f_y denote the focal lengths of the camera and (C_x, C_y) is the principal point of the image.

From these equations, it can already be seen that an important aspect of estimating the pose resides in the capability of the IP system to extract features $\boldsymbol{p}$ from a 2D image of the target spacecraft, which in turn need to be matched with pre-selected features $\boldsymbol{r}^B$ in the wireframe 3D model. Notably, such wireframe model of the target needs to be made available prior to the estimation. Notice also that the problem is not well defined for $n < 3$ feature points, and can have up to four positive solutions for $n = 3$ [32]. Generally, more features are required in presence of large noise and/or symmetric objects.

The on-ground validation pipeline of the proposed pose estimation system is shown in Fig. 2 and consists of the following main stages:

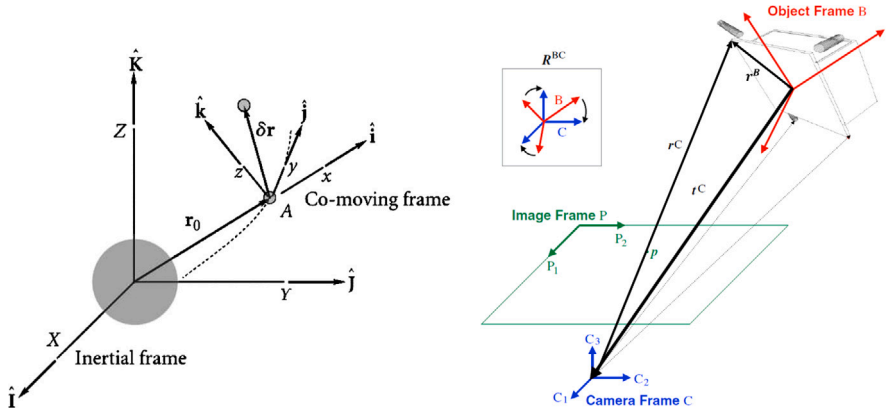
1. **Calibration procedure and Image Acquisition**: laboratory images of a scaled 1:25 mockup model of the Envisat spacecraft are generated by mounting the camera on a robotic arm which performs a rendezvous trajectory around the mockup. Besides, the camera is calibrated with respect to the Envisat mockup in order to associate reference labels of the relative pose between the adopted monocular camera and the mockup for each generated image.
2. **Dataset Generation and CNN Training**: a keypoints-based CNN is trained and validated on augmented datasets. The augmentation is performed by introducing image noise, artificial lights, random background and random textures into synthetically-generated images of the Envisat rendering model.
3. **Online Inference**: the keypoints-based CNN is tested on both synthetic and lab-generated images. The relative pose is estimated by feeding a PnP solver with the detected keypoints as well as with the intrinsic camera parameters and the 3D model of Envisat.
4. **Post-Processing and Validation of Pose Estimation Results**: the CNN-based pose estimation results on the lab-generated images are validated against the reference pose labels, derived from the calibrated setup.

2.1. Pose estimation solver

Following the promising pose estimation results achieved in ADR scenarios in recent studies [9,15,33,34], the Efficient Perspective-n-Points (EPnP) method followed by Gauss–Newton refinement [35] is selected to estimate the relative pose from a set of detected features. This method solves the PnP problem in Eqs. (1)–(2) in closed-form with the EPnP algorithm, and uses the estimated pose as initial guess for an iterative pose refinement. The fundamental equation of the EPnP algorithm consists of rewriting the PnP problem as a function of a 12-dimensional vector $\boldsymbol{y}$ containing the so-called *control point* coordinates in the camera reference system,

$$\boldsymbol{M}\boldsymbol{y} = \boldsymbol{0}, \tag{3}$$

where $\boldsymbol{M}$ is a known $2n \times 12$ matrix. It can be proven [35] that the pose solution belongs to the kernel of $\boldsymbol{M}$, and therefore that it can be expressed as a linear combination of the columns of the right-singular vectors of $\boldsymbol{M}$ corresponding to the null singular values of $\boldsymbol{M}$. As a result, an iterative refinement based on the Gauss–Newton method can be performed with little additional computational cost whilst improving the pose estimate. Note that the EPnP algorithm cannot return an estimate if less than four features are provided as input. As such, no pose estimate can be expected when a large amount of detected keypoints falls below the set threshold for the detection accuracy.



(a) Co-moving LVLH frame [31] (b) PnP problem (Figure adapted from [7])

Fig. 1. Representation of the relative motion framework [31] (a) and schematic of the pose estimation problem using a monocular image (b).

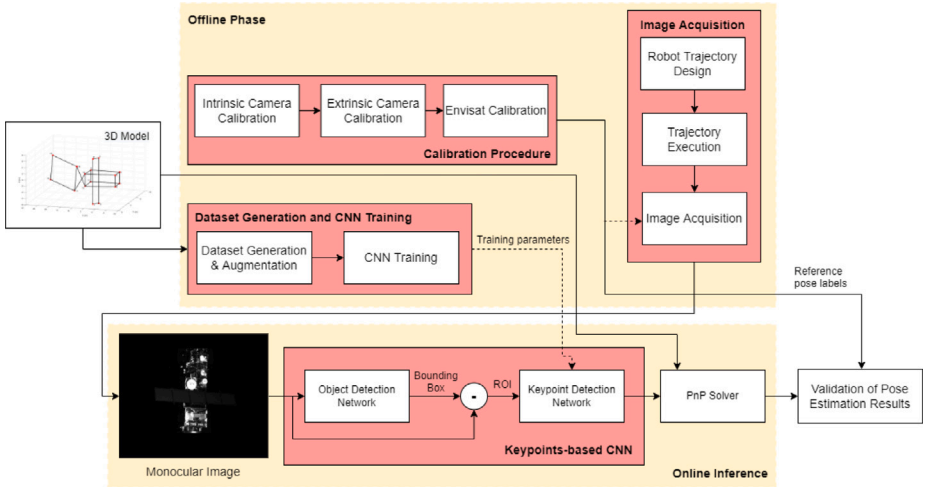


Fig. 2. Illustration of the proposed on-ground validation of the CNN-based pose estimation system.

3. The ORGL testbed

The adopted laboratory setup is illustrated in Fig. 3 and makes use of the GNC Rendezvous, Approach and Landing Simulator (GRALS) testbed of the ORGL facility at ESTEC. The setup consists of the following elements: (a) a 1:25 scaled mockup of the Envisat spacecraft; (b) a Prosilica GC2450 monocular camera; (c) a wall KUKA robotic arm, used to move the Envisat mockup; (d) a ceiling KUKA robotic arm, used to move the camera; (e) the VICON Tracker System (VTS), used to track objects with retro-reflective markers and to provide estimates of their pose with respect to a user-defined reference frame; (f) a lamp mounted on a UR-5 robot, used to recreate the Sun illumination; (g) an external computer providing the software interface between the monocular camera, the VTS and the KUKA robotic arms.

3.1. VICON tracking system

The VTS is a highly accurate motion capture system capable of tracking dynamic objects with millimeter accuracy [36]. The system includes a set of 44 calibrated IR cameras, some retro-reflecting spherical markers which can be detected and tracked by the cameras, and a software interface to stream telemetry to the external computer. In the current setup, a subset of 10 cameras is selected such that the total field of view covers the operating volume in which the image acquisition is carried out.

3.2. KUKA software and hardware elements

The KUKA robotic arms are controlled from the external computer via a Robot Software Interface (RSI) connection. The arms can translate along both ceiling/wall rails and rotate around their six joints, thus guaranteeing a total of 14 degrees of freedom. By default, the command to the robotic arms is represented in terms of end effector pose with respect to a pre-defined KUKA base frame. However, the KUKA software allows user-defined *base* and *tool* reference frames, such that any command can be expressed in terms of a selected tool frame pose with respect to a selected base frame.

3.3. GRALS illumination conditions

In order to recreate a realistic space environment from an illumination standpoint, a movable lamp is mounted on a UR-5 robot and directed towards the target mockup during image acquisition. The lamp is a dimmable, uniform and collimated light source with a spectral response close to 6000 K and exclusive optical lens which provide high uniformity (±5%), shadow-free backlight illumination.[1] Besides, black curtains are placed around the robots' work zone in order to mask most of the background noise, such as unwanted reflections from the robots' rails.

[1] https://www.metaphase-tech.com/backlights/collimated-backlights/.

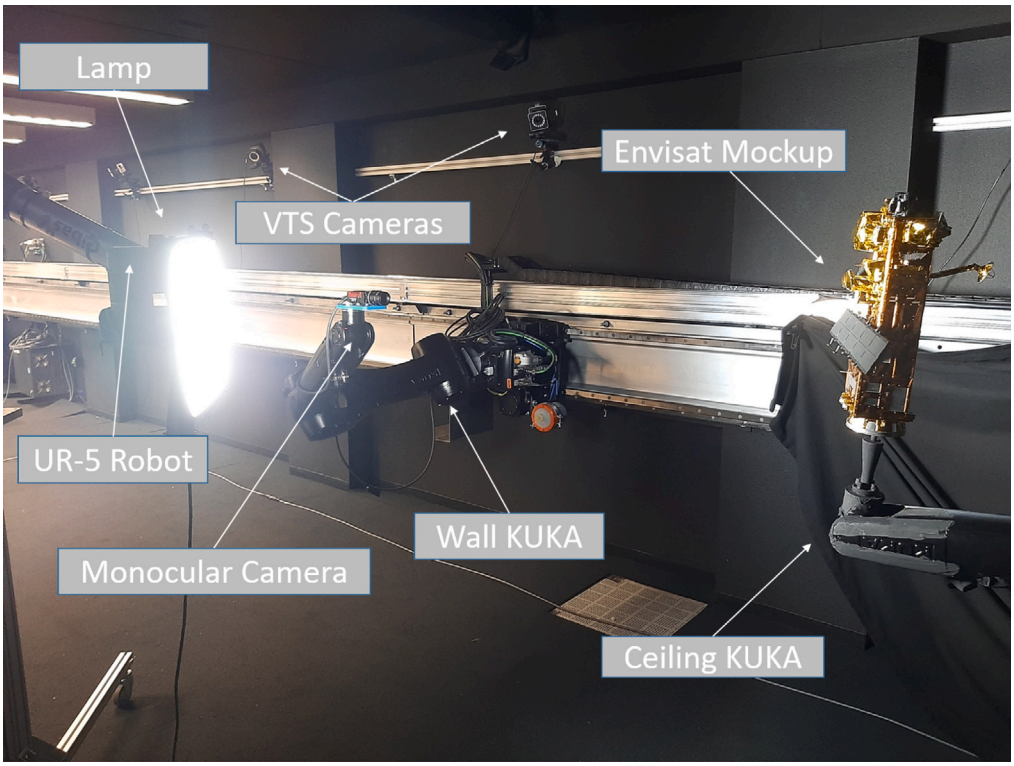


Fig. 3. GRALS testbed with the scaled 1:25 Envisat mockup mounted on the wall KUKA, the VTS and the monocular camera mounted on the ceiling KUKA. Two of the VTS cameras and the Sun lamp are also shown.

4. Calibration framework

The calibration setup consists of the elements described in Section 3 and is inspired by the calibration procedure reported in [22]. The objective is to estimate the relative pose between the monocular camera and the Envisat mockup for each generated image.

4.1. Reference frames definition

Referring to Fig. 4, the following reference frames are defined:

- *LVLH Reference Frame O*: this is the reference frame in which the rendezvous trajectory is expressed (Fig. 1). Its origin is located on the center of mass of the Envisat mockup, and its axes are parallel to the radial, along-track and cross-track directions.
- *VTS Reference Frame V*: this is the reference frame in which all the objects tracked by VTS are expressed. The frame is defined by a calibration tool consisting of a set of 5 IR markers (Wand calibration object). Notably, the origin and orientation of this frame can be arbitrarily set by the user prior to calibration by placing the Wand object at the desired location.
- *Camera Frame C*: this frame is defined such that the third axis is perpendicular to the image plane and aligned with the optical axis of the camera, with the other two axes planar to the focal plane of the camera.
- *Envisat Body Frame B*: this is a rigid frame oriented with its axis parallel to the sides of the Envisat mockup and centered on the Envisat geometrical center.
- *Ceiling/Wall KUKA end effector frames CE/WE*: these frames are centered on the ceiling/wall KUKA end effectors, with their third axis perpendicular to the end effector plate.
- *Markers Object Frame M*: this frame is built from retro-reflective VTS markers attached to a planar surface (not shown in Fig. 4).

The transformation between each of these frames can be expressed by a roto-translation matrix $\boldsymbol{T}$, which incorporates the relative rotation matrix $\boldsymbol{R}$ and the relative position vector $\boldsymbol{t}$,

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}_{1\times 3} & 1 \end{pmatrix}. \tag{4}$$

4.2. Calibration procedure

The purpose of the calibration procedure is to estimate the relative roto-translation matrix $\boldsymbol{T}_B^C$ between the camera frame C and the Envisat body frame B for every monocular image acquired during the desired trajectory. Referring to Fig. 5, the procedure consists of the following steps:

1. Camera Intrinsic Calibration — Estimation of the Camera Intrinsic Parameters.
2. Determination of the roto-translation matrices $\boldsymbol{T}_V^{WE}$, $\boldsymbol{T}_V^{CE}$ — Calibration of the VTS frame with respect to the Ceiling and Wall KUKA end effector frames and definition of LVLH frame O in both KUKA robots.
3. Estimation of the roto-translation matrix $\boldsymbol{T}_C^V$ - Camera Extrinsic Calibration with respect to the VTS frame.
4. Camera Calibration with respect to the LVLH frame $\boldsymbol{T}_O^C$ - Definition of Camera tool frame C in the wall KUKA.
5. Mockup Calibration with respect to the LVLH frame $\boldsymbol{T}_O^B$ - Definition of Mockup Body tool frame B in the ceiling KUKA.
6. Estimation of the roto-translation matrix $\boldsymbol{T}_B^C$ - Mockup-to-Camera Calibration.

Since the purpose of each calibration step is to define both camera and target tools in their respective endeffector frames, this calibration procedure does not have to be repeated each time a different trajectory is recreated, provided that the same mounting configurations are kept. Also, notice that by calibrating each object with respect to their KUKA

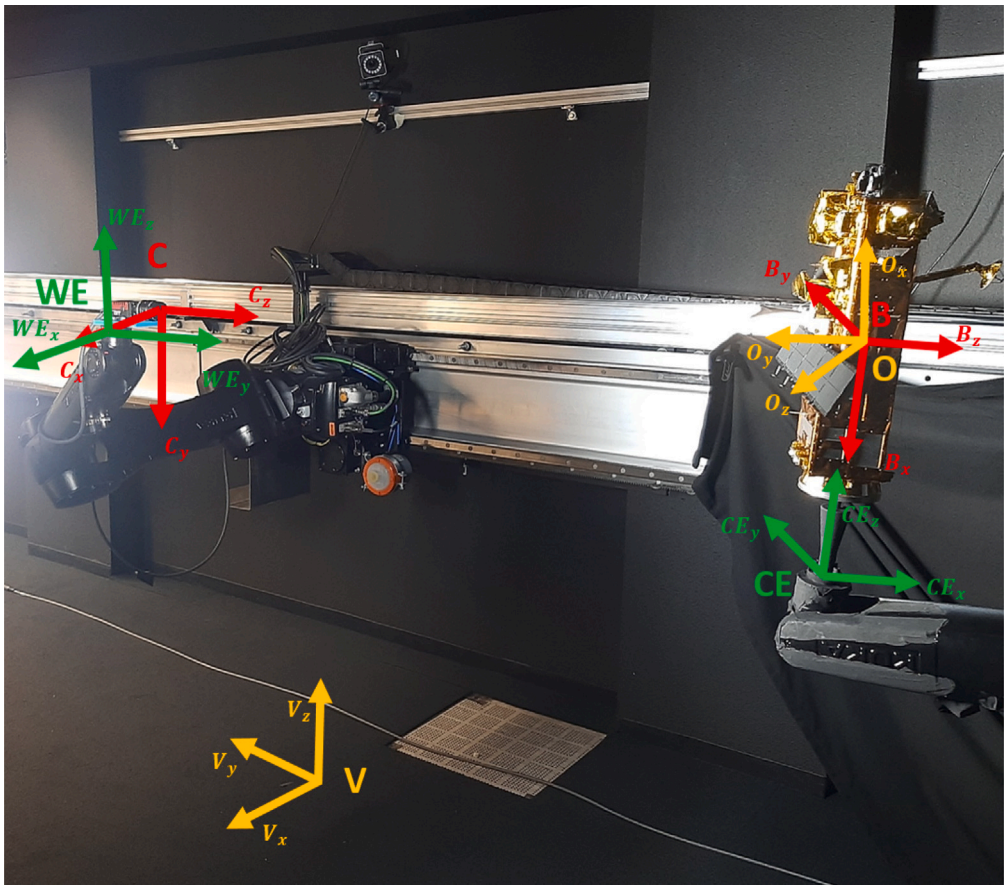


Fig. 4. Illustration of the reference frames adopted during the calibration procedure. The Wall end effector (*WE*) and Ceiling end effector (*CE*) frames are known a-priori, whereas the Camera frame *C* and the Envisat Body frame *B* are unknown and need to be estimated during calibration. The VTS frame *V*, in which the VTS measurements are expressed, can be arbitrarily set by the user. Similarly, the location of the LVLH frame *O* can be chosen based on the constraints on the robot movements for the given laboratory setup. Both the VTS frame *V* and the LVLH frame *O* can be set as tool/base frames in the KUKA environment.

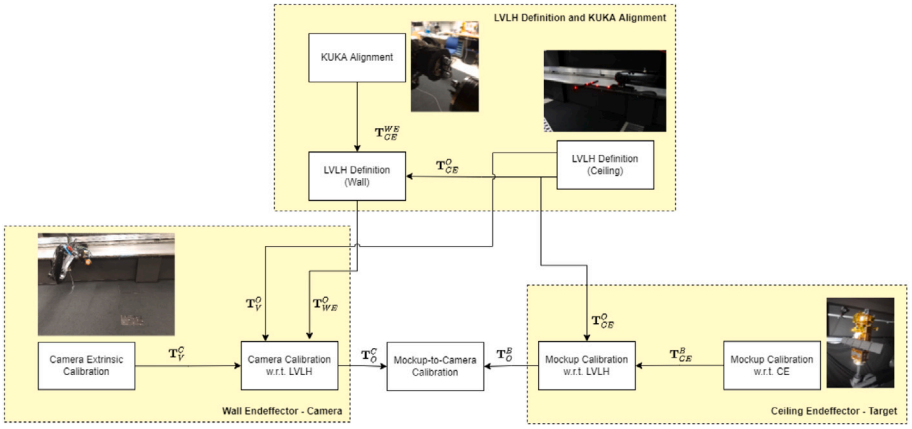
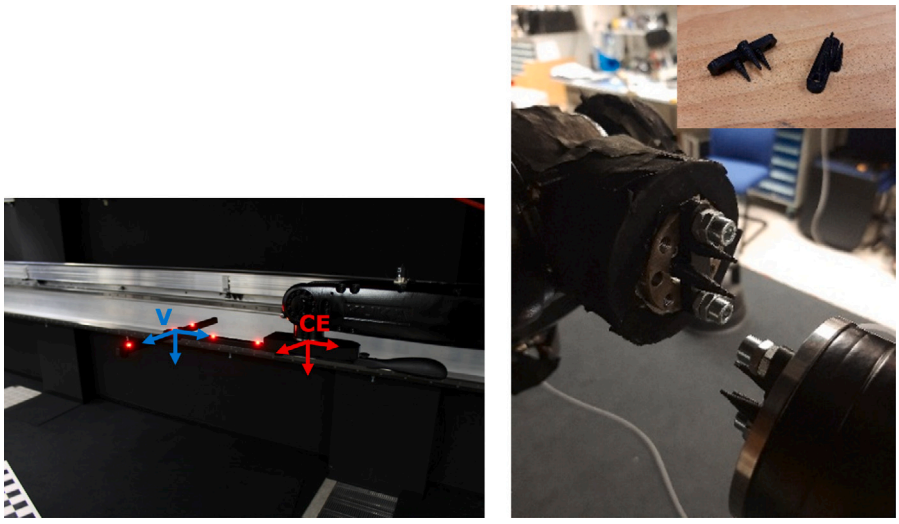


Fig. 5. Description of the transformations required to compute the final mockup-to-camera relative pose. The LVLH definition in both the wall and ceiling KUKA is done in order to define the LVLH frame as base frame in both robots.

end effectors, it is possible to express them with respect to the common LVLH frame *O*. This is accomplished in order to (i) retrieve representative ground truth relative camera-mockup pose labels for each generated monocular image of the Envisat mockup, and (ii) represent the commanded motion of the robotic arms in terms of camera and Envisat pose with respect to the LVLH frame *O*. This latter aspect is very functional since it is desirable to command the translational and rotational motion in the same reference frame of the intended trajectory.

4.3. VTS frame calibration and definition of LVLH frame O

The calibration of the VTS is an essential step towards the overall calibration of the GRALS setup. This calibration is performed (i) to define the LVLH frame *O* in which the rendezvous trajectory will be represented, and (ii) to express the camera frame *C* in the Wall end effector frame *WE* after the camera extrinsic calibration. To do so, the first step consists in defining the VTS frame *V* in the Ceiling end effector frame *CE* (Fig. 6a). This is done by mounting the VTS's Wand calibration object onto the CE and exploiting the knowledge of the



(a) VTS's Wand mounting on ceiling's end effector (b) Alignment of the two KUKA end effectors with the use of alignment pins

Fig. 6. VTS frame (V) definition with respect to the Ceiling end effector frame (CE) and alignment of the two KUKA end effector with the use of alignment pins.

geometry of the mount. At this stage, the LVLH frame O is constructed by matching it to the V frame ($O = V$), in order to ease frames transformations. Once the VTS frame is expressed in the CE frame, the Wand object is removed from the CE, and alignment pins are mounted on both the CE and the WE to align the two KUKAs with sub-millimeter accuracy (Fig. 6b). As a result, the V frame can be expressed in both end effector frames interchangeably.

4.3.1. Camera intrinsic calibration

The first step of the calibration procedure consists of the estimation of the camera intrinsic parameters needed in Eqs. (1)–(2) to solve the PnP problem. This is accomplished by taking multiple images of a chessboard with different camera views and using the *estimateCameraParameters*[2] Matlab built-in function. This function estimates the intrinsic parameters $[f_x, f_y, C_x, C_y]$ and the distortion coefficients of a monocular camera, whilst also returning the images used to estimate the camera parameters and the standard estimation errors for the single camera calibration.

4.3.2. Camera extrinsic calibration

A high-level schematic of the extrinsic camera calibration procedure is illustrated in Fig. 7. The first task is to recreate a planar object M by placing some retro-reflective markers onto a planar surface. Based on similar setups [22], 10 markers were used to recreate the object M.

Next, the planar object M is moved in order to generate pictures of the retro-reflective markers under different camera views. The pixel location of each marker is then extracted by using the Matlab built-in Circular Hough Transform (CHT) algorithm. A manual 2D-3D point correspondence is performed in order to associate each detected marker with its three-dimensional location in the M frame. At this stage, the EPnP algorithm is used to solve the PnP problem and obtain an estimate of the roto-translation between the camera frame C and the VTS frame V, exploiting the knowledge of the relative pose of the markers object M with respect to the VTS frame.

Subsequently, 15 images of the object M are taken with different camera views, and the CHT is applied to each of them to extract the pixel location of the retro-reflective markers. For each frame, the 2D-3D point correspondence can be made by using the initial estimate of $\boldsymbol{T}_C^V$.

The PnP problem can then be solved by means of a non-linear least squares solver, by minimizing the following sum of squares [22]:

$$\sigma_1(\boldsymbol{x}) = \sum_{k=1}^{N_p} \sum_{j=1}^{N_m} \left\| \boldsymbol{p}_{f,i}(k) - \boldsymbol{\pi}\left(\boldsymbol{m}_{f,i}^V(k), \boldsymbol{T}_C^V\right) \right\| \tag{5}$$

$$\boldsymbol{\pi}\left(\boldsymbol{m}_{f,i}^V(k), \boldsymbol{T}_C^V\right) = \left(\frac{x_{f,i}^C}{z_{f,i}^C} f_x + c_x, \frac{y_{f,i}^C}{z_{f,i}^C} f_y + c_y \right) \tag{6}$$

$$\boldsymbol{m}_{f,i}^C = \begin{pmatrix} x_{f,i}^C & y_{f,i}^C & z_{f,i}^C \end{pmatrix}^T = \boldsymbol{R}_V^C \boldsymbol{M}_{f,i}^V + \boldsymbol{t}_C^V \tag{7}$$

where N_m is the number of fiducial markers, N_p is the number of frames and $m_{f,i}^M$ represents the location of the ith marker in the VTS frame V. The output of the minimization is a refined estimate of $\boldsymbol{T}_C^V$, which is used to reproject the 3D retroreflective markers on the image plane and compute the deviation from the correct 2D location of each marker. Fig. 8 shows the reprojection error across the whole set of images of the markers object M. The pixel error can be represented by a distribution with $\boldsymbol{\mu} = [0.14, -0.15]$ px and $\boldsymbol{\sigma} = [1.6, 2]$ px. Overall, the pixel error deviation does not exceed 0.08% of the image size. This compares well with the extrinsic calibration results obtained by Valmorbida et al. [22].

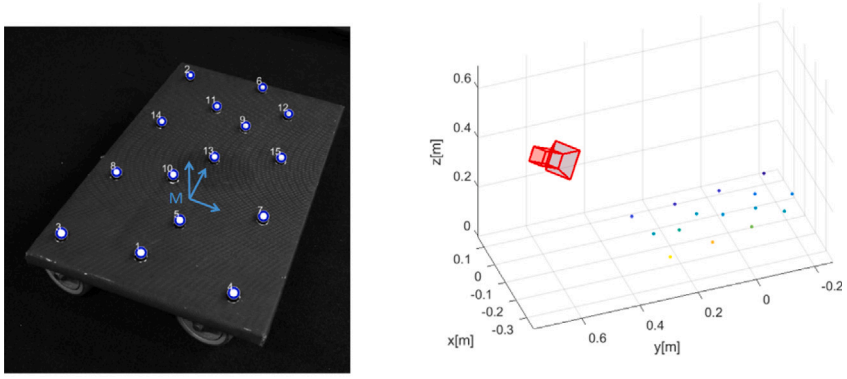
4.4. Mockup calibration

The calibration of the Envisat mockup consists in estimating the relative pose of the Envisat body frame B with respect to the Ceiling end effector frame CE (Fig. 9). Thanks to the adopted design for the mount, which guarantees a unique fixation of the mockup onto the CE, this transformation can be derived directly from the CAD geometry of the mount and the location of the B frame with respect to the mockup mounting interface. Although in principle a dedicated mockup calibration via retro-reflective VICON markers should return more accurate results, the challenges in reconstructing the transformation from the markers frame to the body frame B is currently considered a limiting factor. Specifically, the large number of instruments located on the target and the uneven surface of the Multi-Layer insulation (MLI) prevent from accurately mounting the markers.

4.5. Global calibration error analysis

Overall, the calibration steps described in the previous sections are essential to estimate the desired Mockup-to-Camera transformation $\boldsymbol{T}_B^C$.

[2] mathworks.com/help/vision/ref/estimatecameraparameters.html.



(a) (b)

Fig. 7. Estimation of the roto-translation between the camera frame C and the markers frame M. The markers detection by the CHT algorithm (a) is shown beside the estimated roto-translation of the camera with respect to the M object (b).

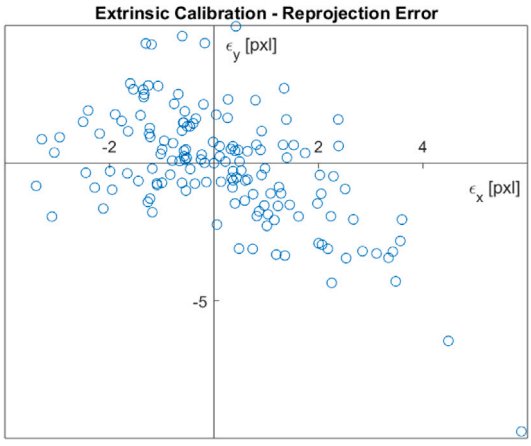


Fig. 8. Reprojection error after camera extrinsic calibration.

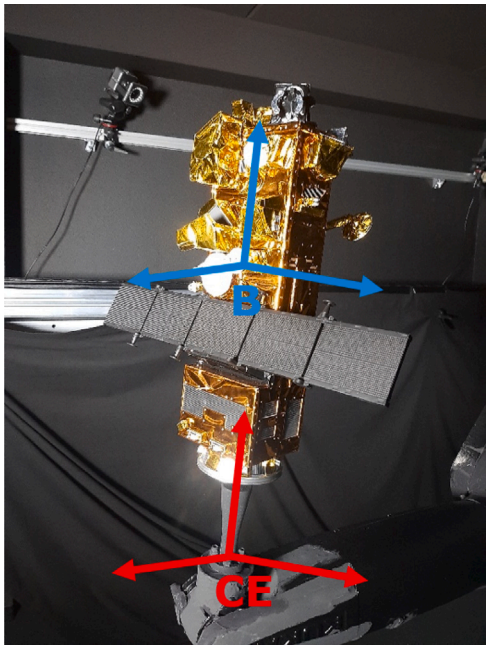


Fig. 9. Illustration of the mounting of the Envisat mockup on the CE.

Each of these steps is characterized by individual calibration inaccuracies that contribute to the global error of the calibration setup $\boldsymbol{E}_{\text{Cal}}$:

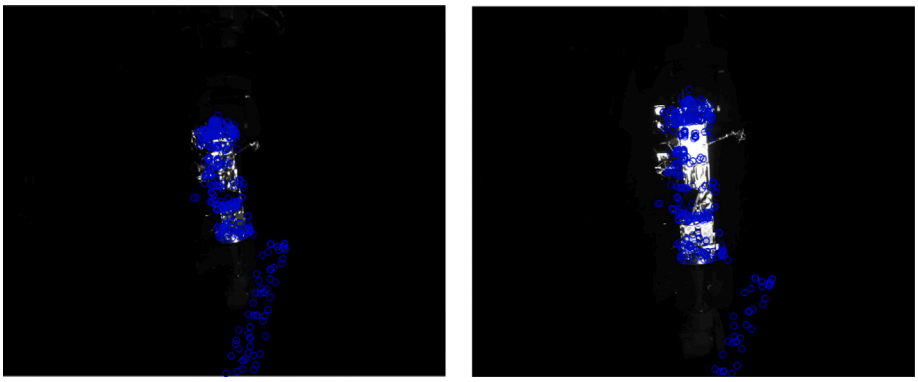
$$\boldsymbol{E}_{\text{Cal}} = \boldsymbol{E}_{\text{Cal}}(\boldsymbol{E}_{\text{Int}}, \boldsymbol{E}_{\text{Ext}}, \boldsymbol{E}_{\text{VTS}}, \boldsymbol{E}_{\text{Al}}, \boldsymbol{E}_{\text{KUKA}}, \boldsymbol{E}_{\text{Env}}) \tag{8}$$

in which:

- $\boldsymbol{E}_{\text{Int}}$, $\boldsymbol{E}_{\text{Ext}}$ represent the reprojection error due to the intrinsic and extrinsic camera calibration with respect to the VTS frame V.
- $\boldsymbol{E}_{\text{VTS}}$ represents the VTS pose error due to inaccuracies in the detection of the retro-reflective markers by the VTS cameras.
- $\boldsymbol{E}_{\text{Al}}$, $\boldsymbol{E}_{\text{KUKA}}$ represent the pose error due to the robots alignment step and due to the intrinsic KUKA inaccuracies, respectively.
- $\boldsymbol{E}_{\text{Env}}$ represents the reprojection error due to inaccuracies in the Envisat-to-CE mount.

Notably, deriving a quantitative global calibration accuracy for each calibration term is complicated by the fact that some of these accuracies are expressed in terms of reprojection error onto the image plane, whilst others are expressed in terms of pose error. To cope with this limitation, the impact of the global calibration error on the estimation error of the transformation $\boldsymbol{T}_B^C$ is assessed by monitoring the reprojection error ϵ of the mockup's corners on a small subset of five of the generated monocular images of the target spacecraft. This reprojection can be obtained by manually selecting the visible corners in each image of the subset and by comparing their pixel location with the reprojection based on each 3D point from the estimated camera intrinsic parameters and the calibrated relative pose $\boldsymbol{T}_B^C$ (Eqs. (1)–(2)). Table 1 lists the error contribution of each calibration step together with the final mean reprojection error on the selected subset. Notice that the Envisat mounting error $\boldsymbol{E}_{\text{Env}}$ could not be quantified due to the unknown relation between the mounting inaccuracies and the resulting mockup pose inaccuracy. Besides, the entire point cloud, obtained from the CAD model of the mockup, can be reprojected onto each generated image to get a qualitative measure of the calibration accuracy. Fig. 10 illustrates two representative examples of the point cloud reprojection in different relative ranges, together with the mean reprojection error derived from the visible corners. Overall, the same order of magnitude of the reprojection error is observed for the remaining images of the subset, sampled across the trajectory to cover different camera-target ranges.

The proposed calibration procedure exhibits a larger total reprojection error, when compared to the sub-pixel results obtained by Park et al. [23] with a more dedicated Robot/World Hand/Eye (RWHE) calibration [37] of Stanford's TRON facility. However, the calibration error of TRON was minimized and computed on a subset of very close-range poses of the target spacecraft. A such, an increase in the reprojection



(a) $\boldsymbol{t}_B^C = 2.8$ m - $\epsilon = 19.5$ px (b) $\boldsymbol{t}_B^C = 2$ m - $\epsilon = 18$ px

Fig. 10. Reprojection of the Envisat point cloud onto the image plane for two representative Mockup-to-Camera relative poses. The mean reprojection error ϵ is computed by manually selecting the visible corners of the mockup and by comparing their pixel location with the reprojected values derived from calibration.

Table 1
Global calibration error analysis. Note that the Envisat mounting error $\boldsymbol{E}_{\text{Env}}$ could not be quantified due to the unknown relation between the mounting inaccuracies and the resulting mockup pose inaccuracy. As described in the text, the total reprojection error is computed by comparing the reprojected corners of Envisat with the visible corners in a subset of 5 representative images.

Calibration error	Value	Description
$\boldsymbol{E}_{\text{Int}}$	0.22 px	Reprojection error
$\boldsymbol{E}_{\text{Ext}}$	(0.14,−0.15) px ± (1.6,2) px	Reprojection error
$\boldsymbol{E}_{\text{VTS}}$	<1 mm	Markers detection error
$\boldsymbol{E}_{\text{AI}}$	0.1 mm–0.02 deg	KUKA accuracy
$\boldsymbol{E}_{\text{KUKA}}$	0.1 mm–0.02 deg	KUKA accuracy
$\boldsymbol{E}_{\text{Env}}$	–	Mounting error
$\boldsymbol{E}_{\text{Cal}}$	**18.7 px**	**Total reprojection error**

error for larger relative ranges is expected. Yet, the uncertainty in the Envisat mounting suggests that $\boldsymbol{E}_{\text{Env}}$ is the main contributor to the larger calibration error observed in the proposed GRALS setup. Nevertheless, it is expected that the relative pose errors resulting from the reprojection offsets can still guarantee a representative ground truth for the intended on-ground validation.

4.6. Rendezvous trajectory generation

Once the GRALS testbed is fully calibrated, any relative trajectory between the monocular camera and the target mockup can be recreated by commanding the two KUKA robotic arms in terms of camera/mockup tool frames with respect to the LVLH frame *O*. To comply with the physical constraint of the robotic arms, a rectilinear approach in-line with the flight path of the servicer spacecraft towards the target spacecraft (so called *V-bar* approach) is considered, as this typically occurs during the final stages of close-proximity operations in rendezvous and docking missions [1,2]. This assumption is justified by the fact that the CNN performance on lab-generated images shall be first validated on simplified relative trajectories, before assessing its performance under more complex relative geometries. Following the same line of reasoning, the relative attitude of the target is simplified by recreating a spinning rotation of around 3.5 deg/s along the main longitudinal body axis, superimposed with precession. The magnitude of the Envisat rotation complies with past optical observation data [38], whereas the direction of rotation is chosen based on the constraints in the robotic arm movements. Moreover, relative distances of 4 m down to 1 m are recreated in the lab which correspond to relative distances in the range of 100 m–25 m for the full-scale target spacecraft.

Lastly, the UR-5 robot is used next to the two KUKA robotic arms to control the lamp at 40°, 60°, and 90° Azimuth angles with respect to the Envisat mockup. The location of the lamp is kept fixed throughout the trajectory, but it is changed at the end of each execution in order to execute the same V-bar approach under different illumination conditions. Fig. 11 shows a sample relative pose under varying Azimuth illumination angles. To guarantee consistency throughout all these illuminations, a LUX-meter is used to ensure that the same light irradiance of 1366 W/m^2 (typical of LEO orbits) is kept while changing the Azimuth angle. Although the exact distribution of the irradiation across different wavelength is not monitored, this ensures that the light irradiance on the target surface does not change for different illumination angles. Notably, the use of a lamp as opposed to a more diffusive illumination guarantees worst-case reflections on the target satellite. As a result, the CNN performance can be stress-tested on worst-case illumination scenarios which differ from the synthetic renderings.

It is important to mention that, although a realistic close-proximity approach would undergo varying illumination angles over time due to the motion of the Sun with respect to the LVLH frame, the current assumption of a fixed light source during the approach is justified by the relatively short duration of the relative trajectory. At the same time, the selected Azimuth range is considered representative of the currently planned ADR missions, since close-proximity approaches at larger illumination angles are typically avoided with careful mission design.

5. Convolutional neural network

The main reason for an emerging interest in CNNs for features extraction lies in the capability of their convolutional layers to extract high-level features of objects with improved robustness against image noise and illumination conditions as compared to standard IP algorithms [30]. As shown in Fig. 12, the first essential step of keypoints-based CNN systems is represented by an Object Detection Network (ODN), e.g. Faster R-CNN [39], R-FCN [40] or MobileNet [41], placed before the main CNN. The ODN regresses the coordinates of a bounding box around the target object, in order to crop a Region Of Interest (ROI) and to increase robustness to scale, variation, and background textures. The cropped ROI is then fed into a Keypoint Detection Network, which convolves with the input image and outputs a set of feature maps. These so-called *heatmaps* are detected around pre-selected features on the target object, such as corners or interest points. The 2D pixel coordinates of the heatmap's peak intensity characterize the predicted feature location, with the intensity and the shape indicating the confidence of locating the corresponding keypoint at this position [42]. As such, wrong detections can be discarded based on a user-defined threshold on the detection accuracy returned by the CNN. Notably, the selection of the CNN will drive the achievable keypoints detection accuracy and robustness. Some architectures, such as the stacked Hourglass [43] and the U-Net [44], perform a downsampling of the input followed

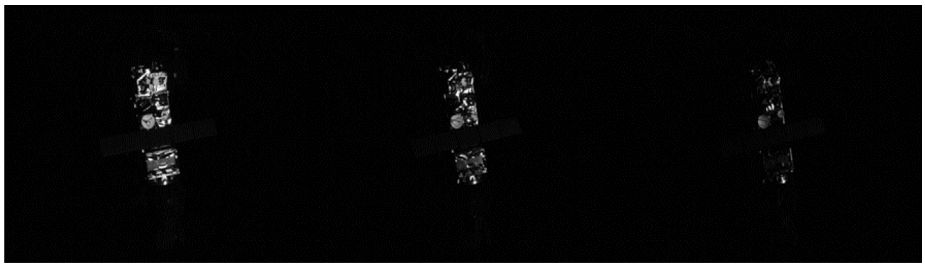


Fig. 11. Example of different illumination conditions for a sample relative pose. Azimuth illumination angles of 40° (left), 60° (center), and 90° (right) are shown.

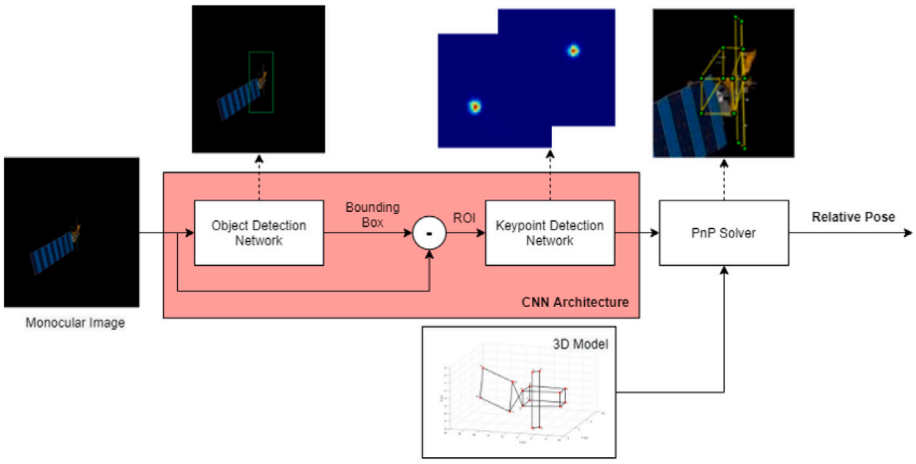


Fig. 12. Proposed CNN architecture and interface with the PnP solver.

Table 2
Parameters of the camera used to generate the synthetic images in Cinema 4D©.

Parameter	Value	Unit
Image resolution	256 × 256	Pixels
Focal length	$3.9{\cdot}10^{-3}$	m
Pixel size	$1.1{\cdot}10^{-5}$	m

in series by an upsampling, in order to detect features at different scales. However, recent advances [33] demonstrated that by using parallel sub-networks across multiple resolutions, rather than multi-resolution serial stages, the CNN can manage to maintain a richer feature representation, facilitating more accurate and precise heatmaps. For this reason, the HRNet [45] architecture currently represents the state-of-the-art in keypoint detection, and is chosen in the proposed pose estimation system.

5.1. Augmentation pipeline

In Fig. 13, the first step of the proposed pipeline for the datasets augmentation and randomization consists in generating ideal synthetic images of the Envisat 3D model. A highly-textured, realistic Envisat model is rendered in the Cinema 4D© software by keeping the virtual camera (parameters in Table 2) fixed and by randomly varying the pose of the rendering model with respect to the camera. Besides, the Azimuth and Elevation of the Sun are randomly varied by ±40 deg around the scenario in which the Sun is exactly behind the camera, in order to recreate favorable as well as more adverse illumination conditions. Next, a randomization pipeline is introduced which adds the following effects to the rendering:

- Texture randomization. This is performed in order to increase the CNN robustness against texture variations between the synthetic and lab models of Envisat. The randomization is achieved in two different ways, by either adding a shader to each material in order to noise the textures, or by directly shuffling the textures of the materials. Besides, the reflectance of each material of the rendering model is also randomized, in order to increase the variability of the material properties across the target body.
- Light randomization. Four additional lights are introduced in random locations, aside from the main Sun illumination, in order to increase the CNN robustness against the illumination conditions recreated in the laboratory setup.
- Background randomization. Random scenes are used as image background in order to increase the CNN robustness against the laboratory environment. Specifically, external disturbance sources in the lab are likely to return non-zero pixel values in the image background, leading to inaccurate CNN detections if the training dataset would lack of background augmentation.

Remarkably, the proposed texture randomization differs from most of the implementations described in Section 1 in that it takes place *before* the actual rendering, and not in post-processing. As a result, the randomization can be performed directly on the actual spacecraft materials and textures without jeopardizing the target body shape. This latter aspect could happen when random texture patterns are superimposed to the target image after rendering. Furthermore, the inclusion of both texture and light augmentation aims at generalizing the training domain to the GRALS testbed's illuminations whilst improving the CNN robustness against previously unseen textures of the target mockup.

Following the Cinema4D© rendering, an additional pipeline is used to further augment the generated images. This is performed by introducing the Earth in the background in some of the images and by corrupting the images with the following noise models:

- Gaussian, shot, impulse and speckle noise.
- Gaussian, defocus, motion and zoom blurs.
- Spatter, color jitter and random erase.

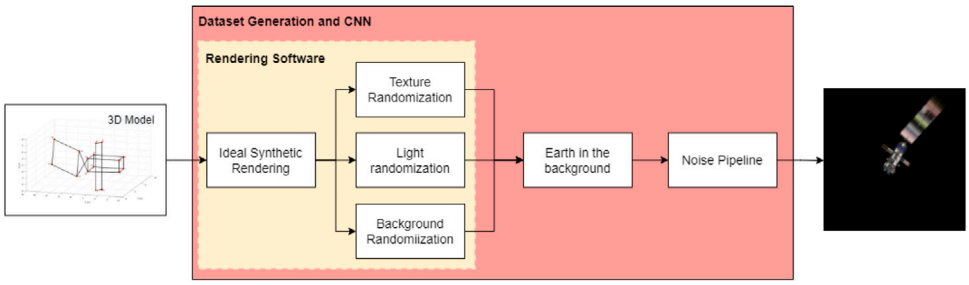
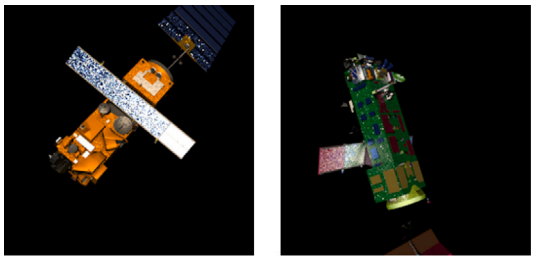


Fig. 13. Dataset augmentation pipeline.

Table 3
Augmentation breakdown. The randomization in the last augmentation step refers to both random lights, textures, and background.

Description	Number of images
No augmentations	1000
Random lights	550
Random lights & textures	2000
Random lights & background	350
All randomizations & Noise & Earth	20,500
Total	**24,400**



(a) Shader effect (b) Randomized textures

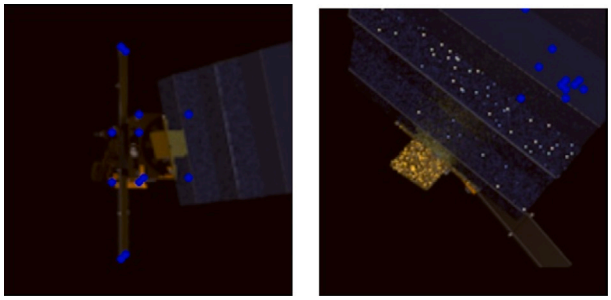
(c) Random background (d) Motion blur

Fig. 14. Output examples of the randomization pipeline.

Table 3 lists all the augmentation techniques together with the number of generated images, whereas Fig. 14 shows four representative examples of the adopted data augmentation techniques. A total of 24,400 images were rendered and further split into training (70%), validation (15%) and test (15%) datasets. These percentages were selected based on other augmentation pipelines [46].

5.2. Training, validation and test

During training, the validation dataset is used beside the training dataset to compute the validation losses and avoid overfitting. The Adam optimizer [47] is used with a cosine decaying learning rate with initial value of 10^{-3} and decaying factor of 0.1. The network is trained for a total of 210 epochs. Finally, the network performance after training is assessed with the synthetic test dataset.



(a) RMSE = 0.47 px (b) RMSE = 93 px

Fig. 15. Example of high (a) and low (b) detection accuracies during poor visibility or occlusion.

The performance is assessed in terms of Root Mean Squared Error (RMSE) between the ground truth (GT) and the x, y coordinates of the extracted features, which is computed as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n_{tot}} \left[(x_{GT,i} - x_i)^2 + (y_{GT,i} - y_i)^2\right]}{n_{tot}}}. \tag{9}$$

The CNN performance on the test dataset shows a mean detection accuracy of 0.97 px, with a RMSE mean $\mu = 2.78$ px and a Mean Absolute Deviation (MAD) of 2.87 px. Overall, this proves that the network is capable of accurately detecting the pre-trained keypoint features in most of the synthetic test images. Notably, wrong detections occur when the solar panel completely hides the main Envisat body. However, the CNN returns good detection accuracies when only parts of Envisat are occluded, demonstrating the capability of learning the relative position between features during partial observability (Fig. 15).

6. Results

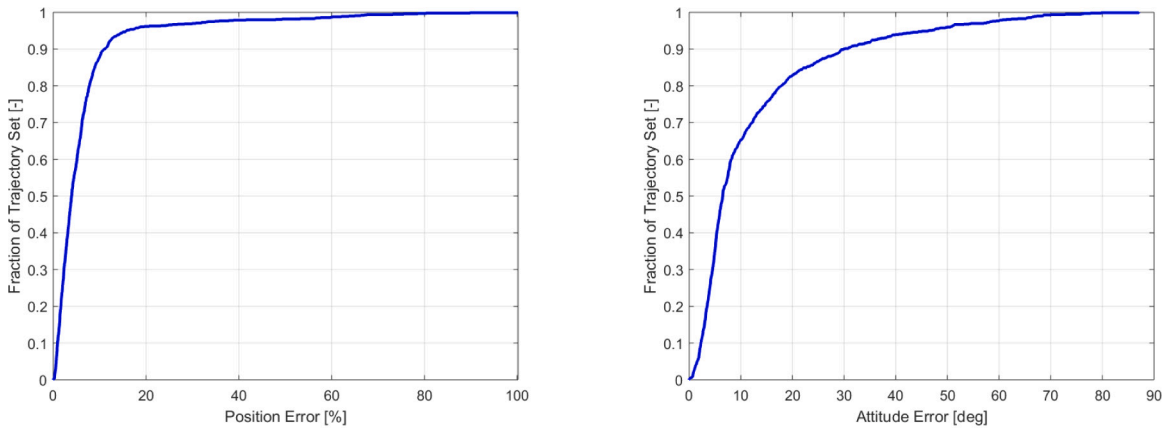
In this section, the pose estimation results are presented for the V-bar trajectory images generated at ESTEC's GRALS testbed. Two separate error metrics are adopted in the evaluation, in accordance with Kisantal et al. [15]. Firstly, the translational error E_T between the estimated relative position $\hat{\boldsymbol{t}}^C$ and the ground truth $\boldsymbol{t}^C$ is computed as

$$E_T = \left\| \boldsymbol{t}^C - \hat{\boldsymbol{t}}^C \right\|. \tag{10}$$

Secondly, the attitude error E_R is measured in terms of the Euler axis-angle error between the estimated quaternion $\hat{\boldsymbol{q}}$ and the ground truth $\boldsymbol{q}$,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_s & \boldsymbol{\beta_v} \end{pmatrix} = \boldsymbol{q} \otimes \hat{\boldsymbol{q}} \tag{11}$$

$$E_R = 2 \arccos(|\beta_s|). \tag{12}$$



(a) (b)

Fig. 16. Cumulative distribution function for the position (a) and attitude (b) errors. As can be seen, an initial steep increase in both curves highlights that most of the images are characterized by relatively low pose errors.

To categorize the pose estimation error, the following definitions are introduced:

- High accuracy: $E_T < 5\%$, $E_R < 2°$,
- Medium accuracy: $E_T < 10\%$, $E_R < 5°$,
- Low accuracy: $E_T < 10\%$, $E_R < 10°$,

in which the position error is expressed as a percentage of the ground truth relative position $\boldsymbol{t}^C$. Moreover, if the number of keypoints within the defined detection threshold falls below the minimum number of features required by the EPnP to solve for the pose, no pose is returned.

6.1. High exposure, 40° Illumination Azimuth

Table 4 lists the categorized pose estimation results as a percentage of the High Exposure, 40° Illumination Azimuth V-bar trajectory images. As can be seen, 59% of the trajectory images are characterized by position errors $E_T < 10\%$ and attitude errors $E_R < 10°$. Moreover, medium and high accuracies are achieved in 31% and 3% of the images, respectively. Furthermore, Fig. 16 shows the cumulative distribution function for both the attitude and position errors across the V-bar trajectory. This function is convenient in that it captures which fraction of the trajectory images returns a certain pose estimation accuracy.

Overall, these results highlight a remarkable performance of the proposed pose estimation system. Despite the limitations in the achievable calibration accuracies reported in Section 4.5, the results demonstrate that a CNN trained on augmented, purely synthetic images can adapt to a previously unseen domain, and perform accurate keypoints detections. Specifically, the inclusion of a texture randomization step within the data augmentation pipeline guarantees that the CNN focuses on the shape of the target rather than on its textures. This improves the detection robustness against illumination conditions and material reflections that were not part of the training dataset.

To help analyze the CNN detection performance prior to pose estimation, Fig. 17 illustrates four representative CNN detections for each pose accuracy category. First of all, a scenario characterized by an adverse MLI reflection is shown for which no pose estimate is returned. These unfavorable reflections are very challenging to handle by the CNN, resulting in a highly scattered and inaccurate keypoints detection. Next, the pose estimate scenarios are displayed. Notably, a lower detection accuracy can be inferred for the upper corners in the high and medium accuracy estimates. This is deemed to be a direct consequence of instruments occlusion, which is not properly captured in the training dataset due to the differences between the target mockup and the rendering model. Furthermore, a low accuracy in the detected SAR antenna keypoints can be observed in the low accuracy scenario.

Table 4
Pose estimation results for the high exposure, 40° Azimuth V-bar trajectory. Position results are scaled from the ORGL setup to the real orbital distances by accounting for the real dimensions of the target spacecraft. The remaining 6% of the trajectory images are characterized by pose errors above the threshold set for the low accuracy.

Scenario	No pose	High accuracy	Medium accuracy	Low accuracy
High exp. 40° Az.	1%	3% $\bar{E}_T = 0.2$ m $\bar{E}_R = 0.8°$	31% $\bar{E}_T = 0.8$ m $\bar{E}_R = 3°$	59% $\bar{E}_T = 1.4$ m $\bar{E}_R = 4.6°$

The SAR antenna corners are generally easier to detect than other target keypoints, mostly due to a higher similarity with the rendering model and a lack of adverse MLI reflection. As a result, they are retained as the main contributors to high and medium pose accuracies, leading to lower pose accuracies when not accurately detected.

To further investigate the overall performance of the proposed system, the pose estimation results are extended to the entire V-bar trajectory. Firstly, Fig. 18 shows the camera boresight component of the estimated position against the ground truth relative distance. Overall, it can be seen that the estimation follows the true value with error peaks scattered throughout the trajectory. Notably, a larger number of outliers is observed for mid-far ranges, suggesting an increase of pose estimation error with distance. Next, Fig. 19 illustrates the pose estimation results after averaging with a moving mean with a window size $k = 100$. This is done in order to capture the relation between the mean estimation error and the relative distance between the monocular camera and the target. As a validation, both the position and attitude errors exhibit the typical trend observed in monocular pose estimation systems [9,15,28], with larger mean position errors at larger distances and fairly equal attitude errors unless the target is very close to the camera. Furthermore, the obtained mean attitude errors are comparable to the ones obtained by other pose estimation systems on the lab-generated images of the SPEED dataset [15]. This is remarkable since, although the SPEED dataset includes 300 images under several relative poses, the adopted illumination source consisted of light boxes resembling the Earth albedo in no-eclipse scenarios. This is an illumination condition far less extreme than a direct, high intensity Sun lamp, due to the patterned flare introduced by the sun lamp and intense surface glow due to high reflectivity and overexposure of the camera. As a result, a smaller domain adaptation is required from the synthetic SPEED dataset compared to the GRALS trajectory images.

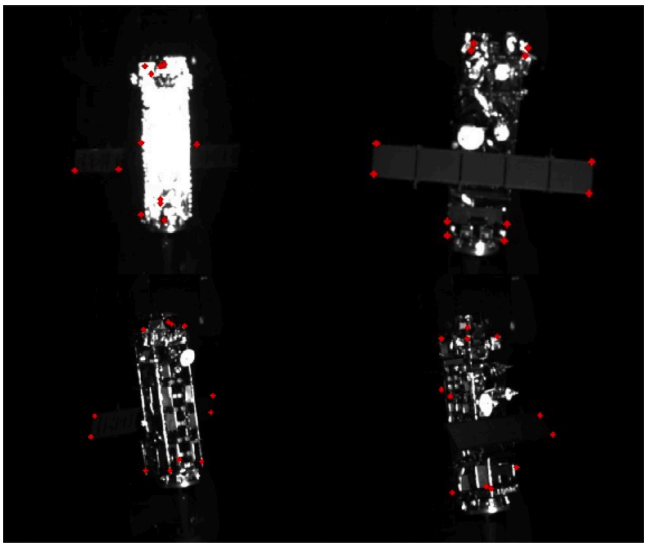


Fig. 17. Upper left: No pose due to adverse MLI reflection. **Upper right:** High accuracy. **Lower left:** Medium accuracy due to instruments occlusion on upper corners. **Lower right:** Low accuracy due to improper SAR corners detection and offset in the lower corners.

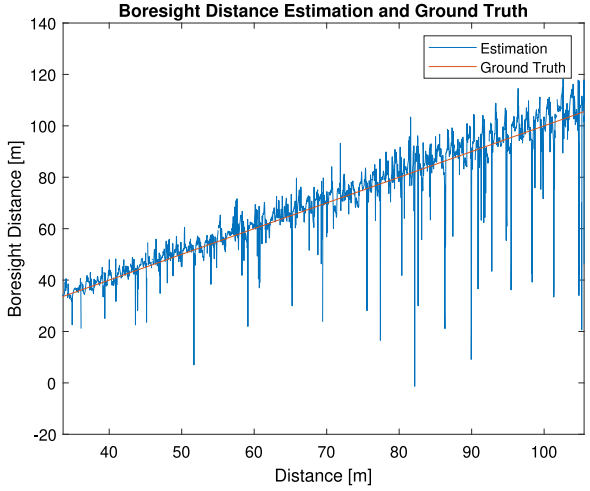


Fig. 18. Boresight estimation from the CNN+EPnP pipeline compared to the ground truth pose from calibration.

6.2. Low exposure, 60°–90° Illumination Azimuth

Table 5 lists the categorized pose estimation results as a percentage of the low Exposure, 60°–90° Illumination Azimuth V-bar trajectory images. As anticipated in Section 4.6, these trajectories are characterized by more adverse illumination conditions as well as by a much lower exposure of the monocular camera, in order to stress-test the CNN performance on extreme scenarios. As can be seen, the pose estimation accuracy drops considerably compared to the results observed in Table 4, indicating that the adopted training data augmentation is not enough to bridge this synthetic-lab domain gap. However, the severe illumination conditions in these two scenarios suggest that the main cause of a larger domain adaptation could be traced back to the randomization of the main light source locations recreated during training. In other words, the extremely low pose accuracies are not expected to be a direct result of an insufficient texture randomization, and further improvements in the CNN training shall aim at extending the illumination scenarios.

Table 5
Pose estimation results for the low exposure, 60°–90° Azimuth V-bar trajectory.

Scenario	No pose	High accuracy	Medium accuracy	Low accuracy
Low exp. 60° Az.	38%	0.3%	3%	5%
Low exp. 90° Az.	75%	0	0.5%	2%

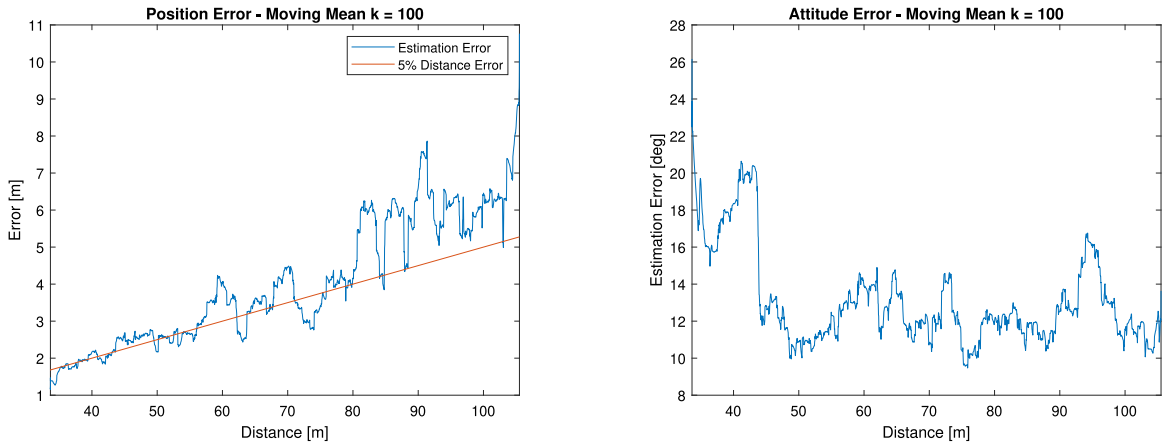
6.3. Pose error analysis

The pose estimation analysis in Section 6.1 returned important insights on the performance of the CNN in the high exposure V-bar scenario, proving its capability to return satisfactory pose estimates for over half of the images. Yet, it is also important to investigate the scenarios in which the pose estimate considerably drifts from the ground truth, i.e. the large errors observed in Fig. 18. Although the majority of the pose outliers are related to a poor keypoints detection, there might be cases in which the large estimation error stems from solving the PnP problem rather than from the CNN detection step. Fig. 20a illustrates a representative example: as can be seen, the CNN performs an accurate detection of most of the keypoints. However, the attitude error associated to this detection amounts to $E_T > 50°$. Notably, the fact that the position error is relatively small suggests that the estimation algorithm is correctly locating the target but confusing its orientation. Generally, this is an indication that there could be a wrong 2D-3D correspondence before solving Eqs. (1)–(2). Specifically, a wrong correspondence would happen if the CNN suddenly switches the keypoints detection order, as this would associate the wrong 2D features to the 3D model points. Since heatmaps are a good indicator of the CNN confidence on each detection, it could be possible to correlate the heatmaps' shape and intensity with a potential features switch. Following this line of reasoning, Fig. 20 shows a representative example in which inaccurate heatmaps detected at image frame k (b) lead to a large pose estimation despite a correct location of the 2D features at image frame k + 1 (a). Ideally, such large heatmaps dispersions could be used to trigger a 2D-3D mismatch flag at step k + 1. In this case, the SoftPosIT algorithm [48] could be used to solve for the relative pose, exploiting the fact that this algorithm assumes unknown feature correspondences. As SoftPosIT is an iterative solver, the estimated pose at frame k would be used during initialization.

Results for the selected scenario indicate that the estimated pose can be refined once the correspondences are handled by SoftPosIT. Specifically, an attitude error $E_R < 2°$ can be achieved, proving that the proposed method could be used to refine the pose under wrong 2D–3D correspondences. Unfortunately, the validation of the proposed method over the entire image sequence of the V-bar trajectory showed multiple scenarios in which a large heatmaps dispersion does not correlate with a feature switch, leading to even worse accuracies after the iterative refinement. As such, different correlations shall be investigated to assess the robustness of this method. Nevertheless, the pose error analysis showed that the CNN can confuse similar features, when tested on a domain which is very different than the training one.

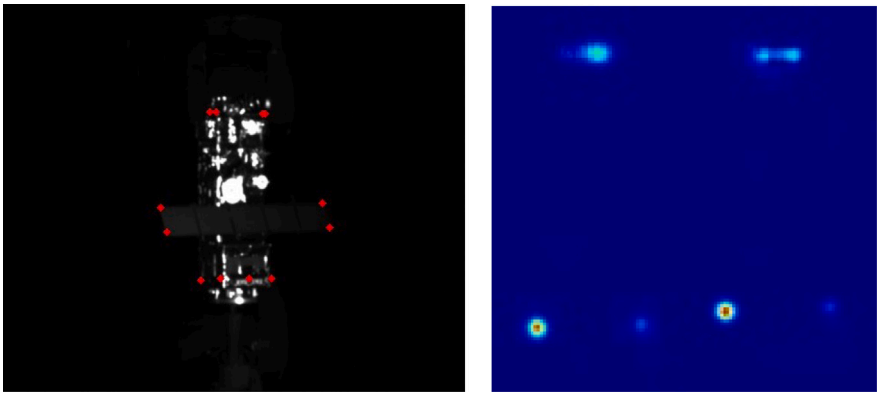
7. Conclusions

This paper introduces a novel on-ground validation framework to assess the performance of a monocular CNN-based pose estimation system on lab-generated space imagery, whilst providing a systematic introduction of ESTEC's GNC Rendezvous, Approach and Landing Simulator (GRALS) for close-proximity operations around uncooperative spacecraft. The performance of the proposed system is evaluated on a representative V-bar trajectory around a 1:25 mockup model of the Envisat spacecraft by recreating space-like illumination conditions and simultaneously operating two KUKA robots, in order to recreate the



(a) (b)

Fig. 19. Moving average trends of the estimated relative position (a) and attitude (b).



(a) CNN keypoints - frame k+1 (b) CNN Heatmaps - frame k

Fig. 20. Example of an accurate keypoints detection (a) leading to a large pose estimation error as a result of features switch (b).

translational motion of the camera as well as the rotational motion of the target spacecraft. Thanks to the reconfigurability of the robotic arms, the proposed setup is capable of recreating realistic rendezvous trajectories under multiple camera-target geometries. Moreover, the proposed calibration procedure guarantees accurate reference pose labels associated to each image of the generated trajectory, allowing a reliable validation of the CNN pose estimation performance.

The domain shift problem typical of CNNs is tackled by introducing a novel data augmentation pipeline which includes both light and texture randomization. Results on the high exposure, 40° illumination Azimuth scenario indicate that over half of the V-bar trajectory is characterized by pose accuracies $E_T < 10\%$, $E_R < 10°$, an impressive result given the large domain gap between the synthetic training images and the GRALS-generated images. Specifically, these results highlight that texture randomization during training increases the CNN robustness against previously unseen target features, forcing the CNN to rely on the target shape instead of its textures. Moreover, preliminary analyses on the large pose estimation scenarios indicate that the adopted CNN undergoes feature switching when challenged with large domain shifts, suggesting that an iterative SoftPosIT refinement, triggered by monitoring the heatmaps' dispersion pattern, could further improve the pose estimation accuracy.

Further work is still required on different levels of the proposed pipeline. First of all, the synthetic illumination conditions adopted during training could be further randomized to guarantee reliable pose estimates in the low exposure scenarios. Also, more data augmentation techniques should be explored so to refine the CNN detections. This is expected to improve the mean attitude error on the high exposure scenario. Besides, the performance of the proposed pipeline should be tested on less extreme domain variations, in order to evaluate the pose estimation accuracy in less adverse scenarios. Next, the estimation performance should be evaluated at a navigation filter level to assess if low accuracy measurements can be handled by the filter without leading to divergence. Finally, the overall calibration procedure should be upgraded in order to improve the model reprojection error at closer ranges.

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 study is funded and supported by the European Space Agency and Airbus Defence and Space under Network/Partnering Initiative (NPI) program with grant number NPI 577 - 2017. The first author would like to thank Martin Schwendener and Irene Huertas for the help during the image acquisition campaign at ORGL, and Kuldeep Barad for the adaptation of the HRNet.

References

[1] A. Tatsch, N. Fitz-Coy, S. Gladun, On-orbit Servicing: A Brief Survey, in: Proceedings of the 2006 Performance Metrics for Intelligent Systems Workshop, 2006, pp. 21–23.

[2] M. Wieser, H. Richard, G. Hausmann, J.-C. Meyer, S. Jaekel, M. Lavagna, R. Biesbroek, E.deorbit mission: OHB debris removal concepts, in: ASTRA 2015-13th Symposium on Advanced Space Technologies in Robotics and Automation, Noordwijk, The Netherlands, 2015.

[3] J. Davis, H. Pernicka, Proximity operations about and identification of non-cooperative resident space objects using stereo imaging, Acta Astronaut. 155 (2019) 418–425, http://dx.doi.org/10.1016/j.actaastro.2018.10.033.

[4] V. Pesce, M. Lavagna, R. Bevilacqua, Stereovision-based pose and inertia estimation of unknown and uncooperative space objects, Adv. Space Res. 59 (2017) 236–251, http://dx.doi.org/10.1016/j.asr.2016.10.002.

[5] R. Opromolla, G. Fasano, G. Rufino, M. Grassi, Uncooperative pose estimation with a LIDAR-based system, Acta Astronaut. 110 (2015) 287–297, http://dx.doi.org/10.1016/j.actaastro.2014.11.003.

[6] S. Segal, P. Gurfil, K. Shahid, In-orbit tracking of resident space objects: A comparison of monocular and stereoscopic vision, IEEE Trans. Aerosp. Electron. Syst. 50 (1) (2014) 676–688, http://dx.doi.org/10.1109/TAES.2013.120006.

[7] S. Sharma, J. Ventura, S. D'Amico, Robust model-based monocular pose initialization for noncooperative spacecraft rendezvous, J. Spacecr. Rockets 55 (6) (2018) 1–16, http://dx.doi.org/10.2514/1.A34124.

[8] L. Pasqualetto Cassinis, R. Fonod, E. Gill, Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft, Prog. Aerosp. Sci. 110 (2019) http://dx.doi.org/10.1016/j.paerosci.2019.05.008.

[9] S. Sharma, S. D'Amico, Comparative assessment of techniques for initial pose estimation using monocular vision, Acta Astronaut. 123 (2015) 435–445, http://dx.doi.org/10.1016/j.actaastro.2015.12.032.

[10] S. D'Amico, M. Benn, J. Jorgensen, Pose estimation of an uncooperative spacecraft from actual space imagery, Int. J. Space Sci. Eng. 2 (2) (2014) 171–189, http://dx.doi.org/10.1504/IJSPACESE.2014.060600.

[11] S. Sharma, C. Beierle, S. D'Amico, Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks, in: IEEE Aerospace Conference, Big Sky, MT, USA, 2018, http://dx.doi.org/10.1109/AERO.2018.8396425.

[12] S. Sharma, S. D'Amico, Neural network-based pose estimation for noncooperative spacecraft rendezvous, IEEE Trans. Aerosp. Electron. Syst. 56 (6) (2020) http://dx.doi.org/10.1109/TAES.2020.2999148.

[13] J. Shi, S. Ulrich, S. Ruel, Cubesat simulation and detection using monocular camera images and convolutional neural networks, in: 2018 AIAA Guidance, Navigation, and Control Conference, Kissimmee, FL, USA, 2018, http://dx.doi.org/10.2514/6.2018-1604.

[14] S. Sonawani, R. Alimo, R. Detry, D. Jeong, A. Hess, H. Ben Amor, Assistive relative pose estimation for on-orbit assembly using convolutional neural networks, in: AIAA Scitech 2020 Forum, Orlando, FL, USA, 2020, http://dx.doi.org/10.1109/AERO.2018.8396425.

[15] M. Kisantal, S. Sharma, T. Park, D. Izzo, M. Martens, S. D'Amico, Satellite pose estimation challenge: Dataset, competition design and results, IEEE Trans. Aerosp. Electron. Syst. (2020) http://dx.doi.org/10.1109/TAES.2020.2989063.

[16] Y. Huo, Z. Li, F. Zhang, Fast and accurate spacecraft pose estimation from single shot space imagery using box reliability and keypoints existence judgments, IEEE Access 8 (2020) http://dx.doi.org/10.1109/ACCESS.2020.3041415.

[17] L. Pasqualetto Cassinis, A. Menicucci, E. Gill, I. Ahrns, J. Gil Fernandez, On-ground validation of a CNN-based monocular pose estimation system for uncooperative spacecraft, in: 8th European Conference on Space Debris, Darmstadt, Germany, 2021.

[18] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: Int. Conf. Intelligent Robots and Systems, 2017, pp. 23–30, http://dx.doi.org/10.1109/IROS.2017.8202133.

[19] M. Wilde, C. Clark, M. Romano, Historical survey of kinematic and dynamic spacecraft simulators for laboratory experimentation of on-orbit proximity maneuvers, Prog. Aerosp. Sci. 110 (2019) http://dx.doi.org/10.1016/j.paerosci.2019.100552.

[20] M. Zwick, I. Huertas, L. Gerdes, G. Ortega, ORGL - ESA's test facility for approach and contact operations in orbital and planetary environments, in: International Symposium on Artificial Intelligence, Robotics and Automation in Space, Madrid, Spain, 2018.

[21] H. Krúger, S. Theil, TRON - hardware-in-the-loop test facility for lunar descent and landing optical navigation, in: IFAC-ACA 2010 Automatic Control in Aerospace, 2010.

[22] A. Valmorbida, M. Mazzucato, M. Pertile, Calibration procedures of a vision-based system for relative motion estimation between satellites flying in proximity, Measurement 151 (2020) http://dx.doi.org/10.1016/j.measurement.2019.107161.

[23] T. Park, J. Bosse, S. D'Amico, Robotic testbed for rendezvous and optical navigation: Multi-source calibration and machine learning use cases, in: AAS/AIAA Astrodynamics Specialist Conference, Big Sky, MT, USA, 2021.

[24] P. Jackson, A.-A. A., S. Bonner, T. Breckon, B. Obara, Style augmentation: Data augmentation via style randomization, in: Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[25] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, W. F.A., W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, in: International Conference on Learning Representations, New Orleans, LA, USA, 2019.

[26] J. Donahue, P. Krahenbuhl, T. Darrell, Adversarial feature learning, in: Int. Conf. Learning Representation, 2017.

[27] M. Ghifary, W. Kleijn, M. Zhang, D. Balduzzi, Deep reconstruction-classification networks for unsupervised domain adaptation, in: Int. Conf. Computer Vision, 2016.

[28] T. Park, S. Sharma, S. D'Amico, Towards robust learning-based pose estimation of noncooperative spacecraft, in: AAS/AIAA Astrodynamics Specialist Conference, Portland, ME, USA, 2019.

[29] K. Black, S. Shankar, D. Fonseka, J. Deutsch, A. Dhir, M. Akella, Real-time, flight-ready, non-cooperative spacecraft pose estimation using monocular imagery, in: 31st AAS/AIAA Space Flight Mechanics Meeting, 2021.

[30] L. Pasqualetto Cassinis, R. Fonod, E. Gill, Evaluation of tightly- and loosely-coupled approaches in CNN-based pose estimation systems for uncooperative spacecraft, Acta Astronaut. 182 (2021) 189–202, http://dx.doi.org/10.1016/j.actaastro.2021.01.035.

[31] H. Curtis, Orbital Mechanics for Engineering Students, Elsevier, 2005.

[32] M.A. Fischer, R. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395, http://dx.doi.org/10.1145/358669.358692.

[33] B. Chen, J. Cao, A. Parra, T. Chin, Satellite pose estimation with deep landmark regression and nonlinear pose refinement, in: International Conference on Computer Vision, Seoul, South Korea, 2019.

[34] S. Sharma, S. D'Amico, Reduced-dynamics pose estimation for non-cooperative spacecraft rendezvous using monocular vision, in: 38th AAS Guidance and Control Conference, Breckenridge, CO, USA, 2017.

[35] Lepetit, F. Moreno-Noguer, P. Fua, Epnp: an accurate O(n) solution to the PnP problem, Int. J. Comput. Vis. 81 (2009) 155–166, http://dx.doi.org/10.1007/s11263-008-0152-6.

[36] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, X. Savatier, A study of vicon system positioning performance, Sensors 17 (7) (2017) 1591, http://dx.doi.org/10.3390/s17071591.

[37] A. Tabb, K. Yousef, Solving the robot-world hand-eye (s) calibration problem with iterative methods, Mach. Vis. Appl. (2017) 569–590, http://dx.doi.org/10.1007/s00138-017-0841-7.

[38] L. Hou-Yuan, Z. Chang-Yin, An estimation of Envisat's rotational state accounting for the precession of its rotational axis caused by gravity-gradient torque, Adv. Space Res. 61 (1) (2018) 182–188, http://dx.doi.org/10.1016/j.asr.2017.10.014.

[39] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[40] S. Ren, K. He, R. Girshick, J. Sun, Object detection via region-based fully convolutional networks, Adv. Neural Inf. Process. Syst. (2016) 379–387.

[41] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, ArXiv Preprint, arXiv:1704.04861.

[42] G. Pavlakos, X. Zhou, A. Chan, K. Derpanis, K. Daniilidis, 6-DoF Object pose from semantic keypoints, in: IEEE International Conference on Robotics and Automation, 2017.

[43] A. Newell, K. Yang, J. Deng, Stacked Hourglass networks for human pose estimation, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision - ECCV 2016, Vol. 9912, Springer, Cham, 2016, pp. 483–499.

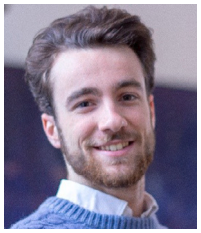
[44] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[45] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019.

[46] K. Black, S. Shankar, D. Fonseka, J. Deutsch, A. Dhir, M. Akella, Real-time, flight-ready, non-cooperative spacecraft pose estimation using monocular imagery, in: 31st AAS/AIAA Space Flight Mechanics Meeting, Virtual, 2021.

[47] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference for Learning Representations, San Diego, CA, USA, 2015, http://dx.doi.org/10.2514/6.2018-2100.

[48] P. David, D. DeMenthon, R. Duraiswami, H. Samet, SoftPOSIT: simultaneous pose and correspondence determination, Int. J. Comput. Vis. 59 (3) (2004) 259–284, http://dx.doi.org/10.1023/B:VISI.0000025800.10423.1f.



Lorenzo Pasqualetto Cassinis is a Ph.D. candidate in the Space System Engineering section of TU Delft and currently a researcher in the GNC section of ESA's European Space Research and Technology Centre and a Visiting Researcher at Stanford's Space Rendezvous Laboratory. He graduated with a M.Sc. in Aerospace Engineering in 2017 from TU Delft, and prior to that with a B.Sc. in Aerospace Engineering from the University of Padua in 2015. During his M.Sc., he worked as systems engineer on flight correlation of the DIDO-II CubeSat. From December 2017 until May 2018, he also worked at GMV on the validation of the GNC software for the PROBA-3 Mission. His current Ph.D. research relates to monocular-based navigation in debris removal scenarios, with special focus on CNN-based systems. This research is a collaboration between TU Delft, ESA, and Airbus Defence and Space.

Dr. **Alessandra Menicucci** received a Laurea diploma in Particle Physics in 2000 from the University of Rome La Sapienza and a Ph.D. degree in Physics in 2004 from University of Rome Tor Vergata working on the space-borne experiment PAMELA for the detection of antimatter in cosmic rays. From 2005 to 2006 she was a Marie-Curie post-doc fellowship at AGFA-Gevaert (Belgium) and then joined the space environment and effects section at ESA-ESTEC where she provided support the numerous ESA missions in the field of radiation environment and effects analysis and lead various R&D activities focusing on radiation monitors developments. In 20015 she moved to the Space System Engineering section of the Aerospace Engineering faculty of the Delft University of Technology where she holds a tenured assistant professor position. Her research interests span over miniaturized sensors, picosatellites, space system engineering, radiation environment and effects analysis. She is author of more than 40 peer reviewed conference and journal papers.

Dr. **Eberhard Gill** received a diploma in physics and holds a Ph.D. in theoretical astrophysics from the Eberhard-Karls-University of Tuebingen, Germany. He holds a M.Sc. of Space Systems Engineering from the Delft University of Technology. He has been working as researcher at the German Aerospace Center (DLR) from 1989 to 2006 in the field of precise satellite orbit determination, autonomous navigation and formation flying. He has developed a GPS-based onboard navigation system for the BIRD microsatellite. Dr. Gill has been Co-Investigator on several international missions, including Mars94-96, Mars-Express, Rosetta, Equator-S and Champ, and acted as Principal Investigator on the PRISMA mission. Since 2007, he holds the Chair of Space Systems Engineering at the Faculty of Aerospace Engineering of the Delft University of Technology. In 2013, he has been appointed also as department head Space Engineering at the faculty. Since 2015, he is founding Director of the TU Delft Space Institute.

Dr. **Ingo Ahrns** received his diploma in computer science in 1996 from the University of Kiel, Germany, and his doctoral thesis in 2000 from the Daimler research center in Ulm, Germany, on the topic of biologically-inspired robot vision. In 2000, he joined the space-robotics department of Airbus Defence and Space. Among many research and development projects, he worked on EUROBOT, DEOS, and the European Lunar Lander, mainly on camera- and LIDAR-based navigation and robot vision activities. In 2018, he started to investigate the use of Deep Learning in space-robotics activities, with focus on CNN based pose-estimation. Since 2013, he became a robot vision expert at Airbus Defence and Space. He currently leads the Autonomous Systems team at Airbus Defence and Space in Bremen, Germany, and coordinates the Airbus Defence and Space robotics R&T cluster.

Manuel Sanchez-Gestido received a M.Sc. in Aerospace Engineering from Technical University of Madrid (UPM). He is GNC System Engineer at the Guidance, Navigation and Control section in the European Space Agency (ESA) at ESTEC (European Space and Technology Research Center). Prior to his current position he was working for several years in Galileo and GNSSs within the Navigation directorate in ESA and before that in Mission Analysis and Mission Planning for Earth Observation missions.