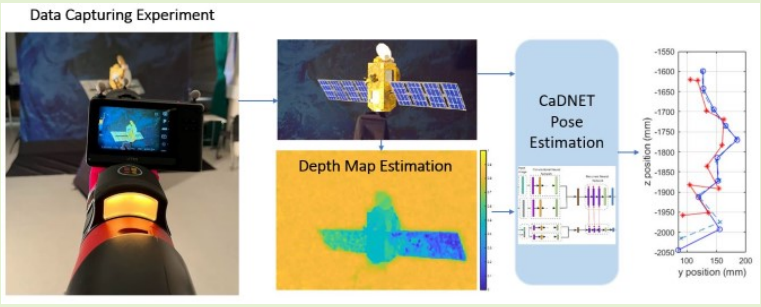
CaDNET: An End-to-End Plenoptic Camera-Based Deep Learning Pose Estimation Approach For Space Orbital Rendezvous

Zakaria Chekakta and Nabil Aouf

***Abstract*— This paper presents a novel deep learning-based approach for relative pose estimation using a focused Plenoptic camera for space rendezvous operations of On-Orbit Servicing (OOS) applications. Plenoptic cameras, also known as light-field cameras, are similar to traditional cameras but have an array of microlenses in front of the sensor. This configuration offers several advantages, such as software-based refocusing and increased image quality in low-light conditions while maintaining an extended depth of field. Moreover, it enables the derivation of 3D depth images from the same light field, making it possible to use a single camera as a stereo vision system for autonomous space rendezvous navigation challenges. We propose a robust deep learning solution suitable for uncooperative close-range rendezvous missions, such as debris removal, based on a Bidirectional Long Short-Term Memory (BiLSTM) network and a Convolutional Neural Network (CNN), to accurately estimate the target's pose from images captured by a Plenoptic camera mounted rigidly on the chaser satellite. We validate the proposed approach, named Cascaded Deep Network (CaDNET), using on-ground data obtained from a designed experimental setup. Through the quality experimental results achieved, we demonstrate the feasibility of adopting the Plenoptic camera as an AI-based relative navigation solution for space rendezvous missions.**



***Index Terms*— Plenoptic Camera, Deep Learning, Navigation, Space Rendezvous**

I. Introduction

RELATIVE navigation algorithms for space close-range rendezvous (RV) such as in On-Orbit Servicing (OOS) have been empirically proven to be essential to guarantee collision-free and reliable space operations such as docking, grasping, refueling, debris removal and inspection. One of the main OOS objectives is to reduce the further accumulation of space debris and due to the necessary high accuracy requirements in terms of target approaching in all OOS missions, adequate autonomous relative navigation algorithms capable of determining the target's movements in space relative to a chaser satellites/spacecraft are essential. The OOS task will be substantially complicated if the target is deemed to be uncooperative and has no supporting equipment for the operation such as in Active Debris Removal (ADR) operations. Conventionally, optical sensors, are typically used to achieve mission autonomy [1], therefore, compact and lightweight passive cameras have been introduced and become the norm as the low-cost sensor for the orbital rendezvous task [2]–[5].

In autonomous rendezvous, the chaser spacecraft carrying the camera should be able to estimate the target spacecraft's relative pose using an onboard navigation algorithm without human intervention. Adopting optical camera sensors onboard the chaser, it is then natural for the pose estimation algorithms to be vision-based. The monocular camera setup tends to be selected as the onboard relative navigation sensor and is chosen over stereo ones [6], [7]. This is because of its simplicity and the limitation of space onboard the chaser satellite in the case of space deployment. Further, during close-range operations, the images must be well-focused throughout the whole mission to allow for reliable tracking, where the Depth of Field (DoF) is required to be relatively large [8].

To overcome the limitation of conventional cameras and fulfill the requirement in terms of accuracy and real-time computation for space rendezvous operations, Plenoptic cameras could be a serious candidate to be considered instead. Due to their enhanced focus range and depth estimation capabilities with a more open aperture, which would allow

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

Z. Chekakta is a Postdoctoral Research Fellow in Intelligent Navigation for Space with the Department of Engineering at City, University of London, EC1V 0HB, UK (e-mail: zakaria.chekakta@city.ac.uk).
N. Aouf is a Professor of Robotics and Autonomous Systems with the Department of Electrical and Electronic Engineering at City, University of London, EC1V 0HB, UK.

sharp and well-focused images during a larger range operation under different lighting conditions, adopting such a camera for orbital rendezvous becomes an attractive camera alternative. Indeed, using only one Plenoptic camera gives the possibility to create 2D images and 3D range images simultaneously [9]. As a result, a Plenoptic camera is considered a passive monocular that can replace a stereo camera offering a 3D range capability.

The concept of Plenoptic cameras is presented in Figure 1(a) and Figure 1(b). It is based on the use of a microlens array (MLA) between the main sensor of the camera lens and its imaging lens. The difference between the focused and the unfocused Plenoptic camera is in the way they capture and process light rays. A focused Plenoptic camera uses microlenses to focus the incoming light onto the sensor, resulting in a sharp image with a well-defined depth of field. In contrast, an unfocused Plenoptic camera captures all the light rays within the field of view, resulting in a low-resolution image with information about light rays coming from multiple directions. The captured light field can be processed to generate different views of the same scene by refocusing the image at different depths.

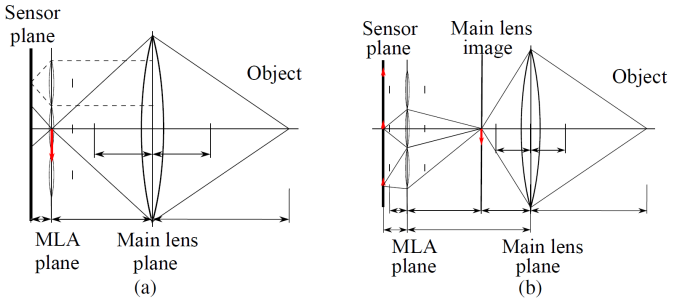


Fig. 1. Plenoptic Camera Concept: (a) Unfocused Plenoptic, (b) Focused Plenoptic

Recently, Plenoptic cameras attracted the interest of researchers around the world. European Space Agency (ESA) has been developing vision-based navigation solutions using various sensors [10], including the option to investigate the merits of adopting the Plenoptic camera as an alternative onboard sensor for future rendezvous missions [11]. The objective of this work is to investigate the benefits of using the Plenoptic camera for close-range uncooperative rendezvous navigation. The Plenoptic sensor technology was initially available for commercial use by Lytro, and later by Raytrix. Significant work related to the Plenoptic camera is focused on the sensor calibration process, its use in the estimation of depth, and the simulation of the Plenoptic camera functionality since acquiring Plenoptic cameras can be quite expensive [12], [13].

In this paper, we propose the following contributions:

- A deep learning architecture for pose estimation using Plenoptic camera sensor data for uncooperative close-range rendezvous operations with on-ground experiment validation.
- An innovative experimental light-field close-range dataset that is well suited for space rendezvous navigation scenarios including different target positions and a range of variable trajectories.

The rest of this paper is organized as follows: Section II introduces the work on Plenoptic cameras in robotics applications, provides existing works on pose estimation in space, and presents the depth map optimization strategy used in our work. Section III details the deep learning relative navigation solution proposed in this work. Section IV is devoted to the experiments and dataset collection, describes the scenarios, and presents the results obtained. Conclusion and remarks are given in Section V.

II. Plenoptic Camera and Pose Estimation

A. Related work

In the literature, research adopting a Plenoptic camera for robotics applications is rare, and more specifically for space applications. Dansereau et al. [14] show that visual odometry plays a crucial role in autonomous robot navigation and how it can be improved for underwater robotics by deploying a Plenoptic Light Field camera. Zeller et al. [15] present a narrow field-of-view vision-based odometry method for indoor robot navigation. Their approach fused the depth data obtained from a monocular Simultaneous Localization and Mapping (SLAM) algorithm with the depth estimated by the Plenoptic camera. In [8], M. Lingenauber et al. discuss the application of Plenoptic cameras for enhancing robot vision during close-range on-orbit servicing maneuvers. Their work highlights the camera's ability to capture 4D light fields, enabling high-quality 2D and 3D depth images that assist in precise robotic arm movements for tasks like grasping and docking with millimeter accuracy under low-light orbital conditions.

Other works show the benefit of using Plenoptic cameras in low-light conditions. Dansereau et al. [16] demonstrate that Plenoptic cameras, through the use of volumetric focus, significantly enhance image quality in low-light conditions by improving the signal-to-noise ratio and maintaining focus across various depths. Their method outperforms traditional imaging techniques, showcasing its efficacy in challenging lighting environments. The same authors show the effectiveness of light field cameras in low-light conditions by using a linear 4D Frequency-Hyperfan filter to enhance image quality. Their approach leverages the redundant information captured by light field cameras to effectively reduce noise while preserving detail and depth of field, making it suitable for challenging lighting environments [17]. S. Zhang et al. [18] explore using Plenoptic cameras to improve depth map estimations in challenging scenarios such as low-light conditions. They highlight that the unique capabilities of Plenoptic cameras allow for capturing depth information more accurately by using the rich structure of light fields, even in environments with minimal illumination.

Concurrently, the field of pose estimation is expanding to machine learning (ML) algorithms. Researchers are leveraging ML techniques, particularly deep learning such as Convolutional Neural Networks (CNNs), to improve the accuracy and robustness of pose estimation algorithms. These advances are crucial for various applications, including terrestrial robotics

vehicles but more importantly, spacecraft rendezvous operations, where precise and reliable pose estimation is essential for the success of space missions.

By 2019, CNNs began to be applied in the realm of spacecraft relative pose estimation for rendezvous and docking (RV), as demonstrated in the ESA Kelvins' Satellite Pose Estimation Challenge [19]. This marked a shift in applying AI to space navigation using vision-based systems. Notably, the most successful approaches in this challenge employed an indirect method of pose estimation, utilizing CNNs primarily for feature extraction combined with Machine Learning techniques, such as a PnP solver, for final pose determination [20], [21](See Figure 2). This strategy, leveraging predefined natural landmarks on the target's surface, suits uncooperative scenarios and represents a departure from traditional image processing methods that require customized feature detection and matching techniques for each scenario. Hence, in the recent development of rendezvous visual-based pose estimation approaches, the Image Processing (IP) step has been completely shifted to a deep learning model able to generalize for different image conditions [5].

In addition to the indirect pose estimation approaches, there has also been a development in end-to-end DNN models that directly generate a pose from image inputs. This methodology has been explored for close-range rendezvous scenarios, tested with various types of imagery and continuous trajectories [5], [22]. End-to-end methods are particularly advantageous because they do not require supplementary ML pipelines to determine the pose, streamlining the process significantly. Furthermore, these methods facilitate the incorporation of Deep Learning-based temporal modeling, which enhances solutions for time-series data by exploiting correlations between successive relative poses. This capability represents a key advancement in applying AI to dynamic and complex scenarios like space navigation.

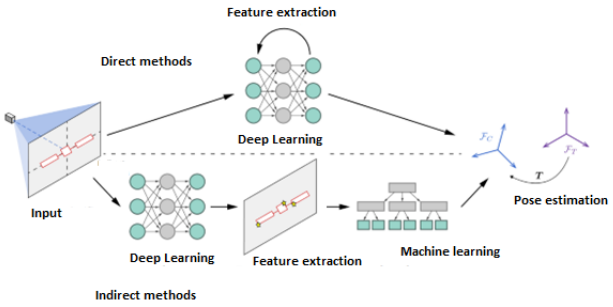


Fig. 2. Direct versus indirect methods for DL-based pose estimation [20]

Despite advancements in the on-ground experiment setup and camera calibration procedures [23], challenges remain in accurately estimating the camera-target relative pose while replicating realistic illumination conditions and other constraining conditions that are required to be met in real rendezvous missions. Numerous studies aim to tackle the domain shift problem [24]. Tobin et al. [24] show that by training a CNN on a diverse set of random, yet unrealistic, textures, it can generalize from synthetic to real-world environments, allowing CNNs to adapt to new domains. Building upon this idea, Jackson et al. [25] and Geirhos et al. [26] mention that randomizing textures during training helps CNNs to focus on learning object shapes rather than textures, thereby increasing the network's robustness. The work of L.P. Cassinis et al. [27] details the development of a CNN-based monocular pose estimation system designed to enhance the accuracy of spacecraft pose estimation in on-orbit servicing and debris removal missions. Their approach leverages a unique on-ground testbed to simulate space-like conditions, focusing on bridging the gap between synthetic and real-world imagery to improve the robustness and accuracy of pose estimates under challenging conditions.

Other studies explore the effects of simple training augmentation on CNN performance using lab-generated images from the SPEED dataset. In [28] authors examine the application of texture randomization in training, which enhances the network's performance on spaceborne images, demonstrating substantial improvements in recognizing and processing images. K. Black et al. [29] introduce a novel CNN-based monocular pose estimation system designed for real-time, flight-ready applications in non-cooperative spacecraft scenarios. Their method demonstrates improvements in achieving state-of-the-art accuracy with low computational demands, effectively generalizing from synthetic training data to real in-space imagery, and optimizing for performance on low-power flight-like hardware. S. Zhang et al. [30] introduce a novel neural network approach called Deep Coherent Point Drift (DeepCPD), for 6D pose estimation of noncooperative spacecraft using point cloud data. The DeepCPD enhances the registration of unorganized scan point clouds to their reference models by replacing the traditional Expectation-Maximization step with a neural network, improving performance and accelerating the process while maintaining robustness against various data imperfections.

In an attempt to assess the performance of a CNN-based pose estimation system in more challenging scenarios, we [5] present a deep learning pipeline named ChiNet [1] using a combination of a CNN and long short-term memory (LSTM). It uses three different training strategies to improve feature learning and end-to-end pose estimation through regressions. The pipeline also fuses thermal infrared data with RGB inputs to mitigate the effects of artifacts from imaging space objects. The capabilities of the proposed framework are demonstrated on a synthetic dataset and validated on experimental data. In the studies, [31], [32], we develop a deep learning-based navigation architecture that combines a CNN with a Recurrent Neural Network (RNN) using LSTM layers. This hybrid architecture is called a Deep Recurrent Convolutional Neural Network (DRCNN). The DRCNN is adopted for 3D LiDAR data. The data collected from the LiDAR sensor are transformed into multi-projected images keeping depth information. The approach is evaluated for a space orbital robotics rendezvous

[1]Pronounced "kai-net," the first term is an abbreviation of the Greek word "chimera," meaning "something made up of parts of things that are different from each other."

relative navigation, and a space landing scenario using both simulation and on-ground validation data.

B. Plenoptic Depth Map Optimization

To this day, the landscape of software tools handling Plenoptic content has been marked by diversity. In 2012, Lytro, a prominent camera manufacturer, released a highly developed and influential software application, the Lytro Desktop Software (LDS) to handle image processing for the Lytro I first-generation and second-generation Illum models. Export options included depth maps (in greyscale), fully focused images, perspective-shift images, a brief video clip of the captured scene, and stereo image pairs. The Lytro I generation and Illum cameras stored images in `.lfp` and `.lfr` formats respectively. However, Lytro's image processing pipeline remains proprietary and has not been publicly maintained since the company ceased operations in 2018. During Lytro's earlier years, independent programmers reverse-engineered the company's file formats to create binary file decoders such as `lfptools` and `python-lfp-reader`. Regrettably, these tools cannot execute essential light-field rendering functions like refocusing. In Matlab, various researchers have published methods that focus on algorithmic calibration and decoding of Lytro's Plenoptic camera [33]–[35]. These approaches utilize provided metadata and concentrate on tasks such as detecting the micro image center, rearranging the 4D data, and rectifying radial lens distortions. More recently, research has shifted towards successfully recovering physical information at the boundaries of light fields [36], [37]. In [13], the author presented a framework for enhancing the decoding process. It incorporates advanced techniques such as scale-space analysis, centroid grid fitting, de-vignetting, illumination channel correction, micro-image resampling, and computational refocusing. `PlenoptiCam v1.0` [13] stands out by supporting cameras with arbitrary dimensions, handling footage from custom-built prototypes and Lytro cameras, and achieving improved accuracy and performance. Its ability to suppress noise, address color variances, and provide accurate angular sampling ensures high-quality results. By offering extensive capabilities and delivering exceptional performance, `PlenoptiCam` emerges as the ideal software for our research in Plenoptic imaging. Figure 3 illustrates the depth estimation strategy employed in our work, highlighting the optimization process.

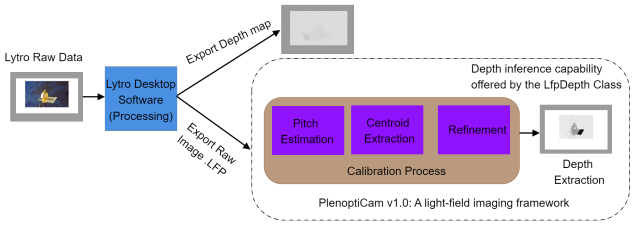


Fig. 3. Depth estimation process

In the study of our research, the captured frames from the Lytro camera undergo processing using the Lytro Desktop Software. Subsequently, the raw images are exported as `lfp` frames. We employed `PlenoptiCam` software, which integrates advanced algorithms for depth estimation from Plenoptic camera data. As illustrated in Figure 3, the depth estimation process is facilitated by the `LfpDepth` class within `PlenoptiCam`, which executes a sequence of calibrated operations to produce depth maps. The process begins with pitch estimation, where the software evaluates the scale-space of micro images to accurately gauge the pitch of the MLA. Following the pitch estimation, the Centroid Extraction process uses the Levenberg-Marquardt optimization algorithm for reducing least-squares errors globally while detecting microlens centers. It simultaneously determines centroid spacing and projective mapping, providing the necessary data for depth calculation. The centroids extracted are then refined to improve their accuracy. This refinement process adjusts the centroid coordinates, enhancing the precision of the subsequent depth extraction process. With the refined centroids, the `LfpDepth` class proceeds to extract depth information. It calculates depth by analyzing the disparities and geometric relations between the different micro images based on their centroid alignments. Further details regarding the process can be found in [13].

III. Cascaded Deep Network (CaDNet) Relative Pose Estimation

The objective of the monocular orbital space navigation solution is to estimate the target frame (R) relative position and orientation to the chaser camera frame (C) using a single monocular image captured by the onboard chaser camera. The relative position between the frames is represented by a translation vector t_{RC}, which specifies the displacement from the origin of the camera frame C to the origin of the target frame R. The relative orientation between the frames is represented by a rotation matrix R_{RC}, which defines the orientation of the target frame with respect to the camera frame.

Figure 4 provides a visual representation of the target and camera reference frames, as well as the position and orientation variables. It shows the relationship between the reference frames and the position and orientation variables.

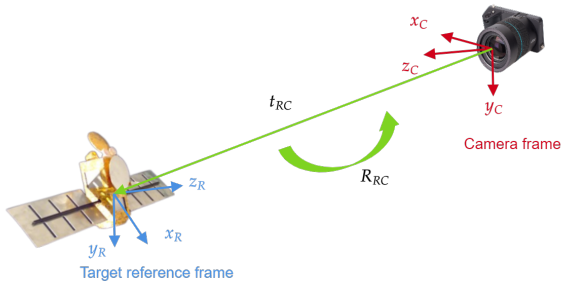


Fig. 4. The target reference frame (R), camera reference frame (C), relative position ($\boldsymbol{t_{RC}}$), and relative attitude ($\boldsymbol{R_{RC}}$).

The process of estimating the pose of a new camera frame relative to a reference target frame can be mathematically formulated as follows:

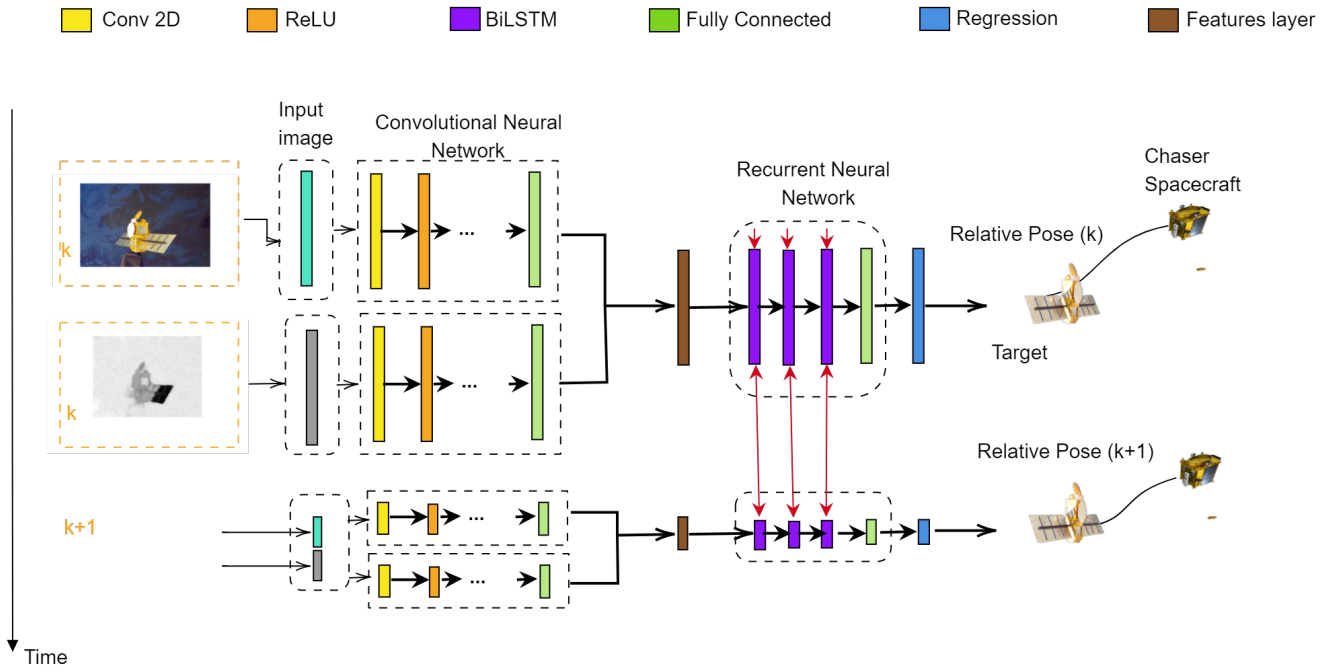


Fig. 5. CaDNET overview. The architecture performs end-to-end spacecraft pose estimation from RGB and the depth map obtained by the Plenoptic camera.

Let $\mathbf{x}_R = [x_R, y_R, z_R, 1]^T$ be a point in the reference frame, and let $\mathbf{x}_C = [x_C, y_C, z_C, 1]^T$ be the corresponding point in the camera frame, where $\mathbf{x}_R$ and $\mathbf{x}_C$ are expressed in homogeneous coordinates. Then, the rigid transformation $G \in \mathrm{SE}(3)$ (Special Euclidean Group in three dimensions) that maps $\mathbf{x}_R$ to $\mathbf{x}_C$ can be represented as:

$$\mathbf{x}_C = G\mathbf{x}_R \tag{1}$$

where G is a 4×4 homogeneous transformation matrix of the form:

$$G = \begin{bmatrix} R_{ot} & t \\ 0_{1\times 3} & 1 \end{bmatrix}, \quad R_{ot} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \quad t = \begin{bmatrix} t_X \\ t_Y \\ t_Z \end{bmatrix} \tag{2}$$

$R_{ot} \in SO(3)$ (Special Orthogonal Group in three dimensions) is a rotation matrix and t is a translation vector.

To estimate the pose of a new camera frame relative to a reference target frame, we need to find the values of the rotation matrix R_{ot} and the translation vector t that best describe the transformation G. This is primarily achieved by minimizing the re-projection error, which is the discrepancy between observed 2D points in the camera frame and their predicted 2D locations obtained by projecting 3D points from the reference frame using the estimated pose. For systems that can capture depth directly, such as cameras equipped with depth sensors or LIDAR, the pose estimation can also leverage the depth information to enhance accuracy by aligning 3D points directly

Let $P_R = p_1, p_2, \ldots, p_n$ be a set of n 3D points in the reference target frame, and let $P_C = q_1, q_2, \ldots, q_n$ be the same set of points in the camera frame. Then, we can define the registration error as the difference between the observed points in the camera frame and their corresponding points in the reference frame after applying the estimated transformation G.

$$e_i = q_i - Gp_i, \quad i = 1, 2, \ldots, n. \tag{3}$$

The goal is to find the values of R and t that minimize the sum of the squares of the registration error:

$$\min_{R,t} \sum_{i=1}^{n} ||q_i - Gp_i||^2. \tag{4}$$

This is a non-linear optimization problem that can be solved using various techniques, such as Gauss-Newton, Levenberg-Marquardt, or gradient descent.

In practice, the estimation of the relative camera pose is usually done incrementally, using a sequence of frames of the reference target frame. This allows for tracking the camera's motion over time. However, to handle scale ambiguity, the camera pose between consecutive frames is often represented as a 3D similarity transformation, which includes a scaling factor to account for changes in the camera's distance from the scene.

To solve this relative pose estimation problem, this paper proposes the adoption of a hybrid deep learning network composed of an RNN module to process the features extracted

by another network module built as a Cascaded CNN network. The resulting CaDNET architecture is shown in Figure 5 to provide a smooth and precise estimate of the 6-Degree Of Freedom (DOF) poses. The CaDNET represents an end-to-end Deep Learning Monocular Pose Estimator where one CNN network contains a set of convolutional layers to extract features from the input RGB image, thus exploiting the natural ability of CNNs to autonomously extract features from images. The other depth CNN network is also composed of a set of convolutional layers in order to extract features from the depth map. The outputs of both networks are concatenated in a *feature layer* $=$ *fully connected*$_1$ $||$ *fully connected*$_2$, where $||(.)$ denotes the process of vertical concatenation. The RNN network is then constructed using bidirectional LSTM (BiLSTM) layers to regress the relative poses in close-range rendezvous application between the target satellite and the chaser satellite. The detailed architecture of the proposed CaDNET and its parameters are presented in Table I.

The main reason to adopt the BiLSTM cells instead of a standard LSTM (presented in Figure 6) in a pose estimation network is to take advantage of the information in the input sequence in both forward and backward directions. While a standard LSTM network processes the input sequence in a forward direction, a BiLSTM network processes the input sequence in both the forward and backward directions, capturing the dependencies in the data in both directions. This can provide a more comprehensive representation of the input data, leading to improved performance in tasks such as pose estimation. During space rendezvous, predicting the current pose accurately can benefit significantly from knowing the subsequent positions, which could indicate the trajectory's direction and velocity.

It is important to clarify that while the future data for a current timestep is not available during prediction (in real-time scenarios), during the training phase, the model learns to utilize the complete sequence context effectively. The BiLSTM, therefore, is trained on full sequences, allowing it to establish patterns that involve both preceding and succeeding elements in the data. BiLSTMs manage to use future information, especially in testing and real-world deployment scenarios where real-time processing is not necessary (e.g., processing batches of data every few seconds or minutes), the incoming data can be buffered until enough future context is collected. This approach is common in applications where a slight delay is acceptable for the benefit of more accurate predictions. In our work, when future data cannot be delayed, the CaDNET reverts to using only past data for real-world testing, utilizing only the unidirectional LSTM approach while still employing BiLSTM for non-real-time analysis.

For each new Plenoptic camera frame, its pose with respect to the reference target frame has to be estimated using the transformation between the reference target coordinate system $\mathbf{x}_R$ and the camera coordinates of the new frame $\mathbf{x}_C$ as given in (2). The matrix G has six degrees of freedom which need to be estimated. Those are the three rotation angles ϕ, θ and ψ as well as the coefficients of the translation vector t_X, t_Y, and t_Z. In our approach, the Euler angles are ordered in a ZYX, Yaw around the Z-axis, followed by pitch around the Y-axis,

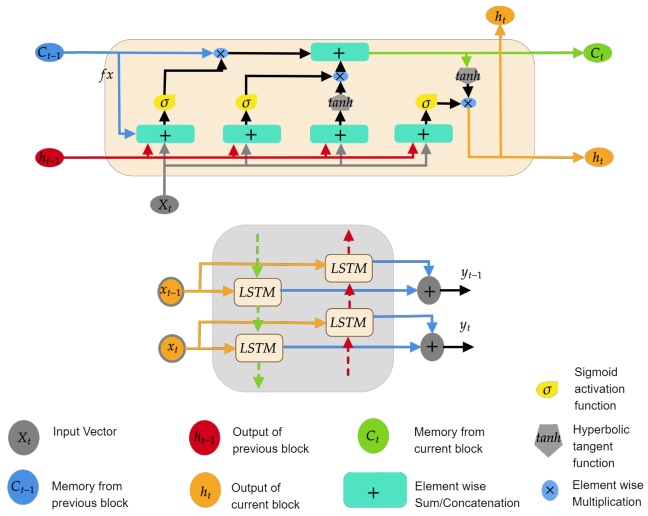


Fig. 6. Top: Long short-term memory (LSTM). Bottom: Bidirectional LSTM (BiLSTM).

and concluding with roll around the X-axis. This convention was chosen because it is commonly used in spacecraft and it effectively represents the rotational dynamics and orientation changes of the spacecraft during rendezvous operations.

Using CaDNET, those quantities are estimated based on the RGB images and depth map both obtained from the onboard Plenoptic camera. In our architecture, detailed in Table I, the output dimension (1024) is being optimized to enhance performance. The elements $1-12$ of the output are extracted. Specifically, the regression vector is defined as:

$$\textit{Regression} = [\, r_{11}\ r_{21}\ r_{31}\ r_{12}\ r_{22}\ r_{32}\ r_{13}\ r_{23}\ r_{33}\ t_X\ t_Y\ t_Z]$$

The training scheme for the CaDNET model ensures that it learns the SO(3) constraints of the rotation matrix from the known ground truth rotation matrix. During training, CaDNET naturally optimizes the estimation of the rotation matrix to meet these constraints because the ground truth rotation matrix adheres to them. As a result, CaDNET usually produces rotation matrices that respect the SO(3) constraints during testing. If an estimated rotation matrix fails to meet these constraints during testing, the solution is rejected and recalculated. However, this is rare since the training process effectively enforces these constraints.

Initially, we investigated using a regression output that directly outputs a 12×1 vector corresponding to the pose estimation. However, empirical results indicated that extending the output layer to 1024 components before keeping only the necessary 12 leads to better performance. By extending the output layer to 1024 components, even though the final pose estimation only requires a 12×1 vector, this mechanism encourages the network to learn a broader range of features and dependencies. This expansion forces the layers preceding the final output, especially the last fully connected layer, to handle and process a more complex feature space. This also helps prevent overfitting to the pose estimation with more generalized learning, which is lost with a smaller output layer. The unused components from 13 to 1024 do not contribute directly to the pose estimation but do enhance the learning

TABLE I
CADNET ARCHITECTURE PARAMETERS

RGB CNN		Depth CNN	
Layer Type	Variables	Layer Type	Variables
Input Layer	RGB image [281 405 3]	Input Layer	Depth map [141 203]
$Conv_1$	Filter size 5×5, padding 3, stride 2, channels 64	$Conv_1$	Filter size 5×5, padding 3, stride 2, channels 64
$ReLU_1$	-	$ReLU_1$	-
$Conv_2$	Filter size 5×5, padding 2, stride 2, channels 128	$Conv_2$	Filter size 5×5, padding 2, stride 2, channels 128
$ReLU_2$	-	$ReLU_2$	-
$Conv_3$	Filter size 5×5, padding 2, stride 2, channels 256	$Conv_3$	Filter size 5×5, padding 2, stride 2, channels 256
$ReLU_3$	-	$ReLU_3$	-
$Conv_4$	Filter size 5×5, padding 2, stride 2, channels 512	$Conv_4$	Filter size 5×5, padding 2, stride 2, channels 512
$ReLU_4$	-	$ReLU_4$	-
$Conv_5$	Filter size 5×5, padding 2, stride 2, channels 1024	$Conv_5$	Filter size 5×5, padding 2, stride 2, channels 1024
$ReLU_5$	-	$ReLU_5$	-
Fully Connected$_1$	1024×1 matrix	Fully Connected$_1$	1024×1 matrix

	Layer Type	Variables
RNN	Features layer	2048×1
	$BiLSTM_1$	hidden values 1000
	$BiLSTM_2$	hidden values 1000
	$BiLSTM_3$	hidden values 1000
	Fully Connected$_2$	1024×1 matrix
	Regression	1024×1 matrix use 12 first variables to predict the pose

process.

IV. EXPERIMENTS AND RESULTS

To evaluate our Plenoptic camera deep learning architecture for pose estimation, several experiments considering different rendezvous scenarios, including different backgrounds behind the target, are performed. Since our main goal is to show the Plenoptic depth map of the scene and how it can improve the pose estimation accuracy when used as an additional input, the ground truth for the scenario's trajectories is measured. All our experiments are performed using the Plenoptic Lytro Illum camera (Parameters in Table II).

TABLE II
PARAMETERS OF THE PLENOPTIC LYTRO ILLUM CAMERA

Parameter	**Value**
Camera Model	Plenoptic Lytro Illum
Resolution	1404×2022 pixels
Focal Length	35 mm
Field of View (FOV)	$18°$ horizontally and vertically

A. Experiments and platform setup

The main goal of the experiments is to create a complete dataset of a mock-up satellite for close-range rendezvous using a Plenoptic technology for training, testing, and validating our deep learning relative pose estimation solution. Motivated by the lack of the Plenoptic camera extension in Space simulation software, we devoted our effort to creating real-world data suitable for the space rendezvous environment and providing realistic data. The setup for acquiring the real data is based on the Testbed facility at the City University of London as shown in Figure 7.

The setup attaches the Lytro camera to the end-effector of the arm robot (Sawyer in this case) to allow simulation of the approaching phase of a spacecraft's close-range rendezvous operation. Using the Computer-Aided Design (CAD) model of Jason-1 as a basis, a laboratory mock-up was created, with a 1:4 scale replica of the original full size of the satellite, with 1-DOF in rotation. A light source has been used to emulate sun illumination. The setup is placed inside an area equipped with an Optitrack system to allow the tracking of all the bodies and calculating the ground truth data of the relative motion between the onboard chaser camera and the target satellite. Additionally, the camera has been calibrated (See II-B) to provide better depth estimation of the scene as shown in Figure 8.

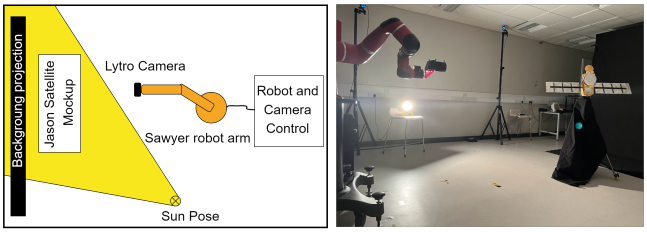


Fig. 7. Overview of the experimental setup. Left : an arm robot, a mock-up of the target satellite, and a simulated sun with a black background for a similar space conditions. Right: A real setup scene using a mock-up of the Jason satellite, and the Lytro camera is attached to the Sawyer robot.

The motion capture system for recording the ground truth measures approximately $5 \times 5 \times 3$ m. OptiTrack can record 6-DOF pose data of rigid and flexible bodies by detecting, tracking, and triangulating passive near infrared markers placed on targets. The data can be saved or streamed over a local network in real-time. The OptiTrack setup at the City University of London consists of six PrimeX 13 cameras with a resolution of 1280×1024 px running at a native framerate of 240 Hz, capable of achieving positional errors less than 0.20 mm and rotational errors less than 0.5 deg.

Fig. 8. Left: Camera frame, Right: depth estimation

B. Dataset generation

It is worth mentioning that the Jason 1 mock-up is attached to a stand and can be rotated and placed in a different orientation which gives us the opportunity to create multiple scenarios based on the target's initial pose and the different backgrounds. We propose and create two scenarios for the data collection experiments.

1) Scenario 1: In this scenario, we utilize the Optitrack software to record the initial pose of the target by attaching markers to it. The scenario considers the target as floating in space without any Earth background in the scene. Additional markers were also placed on the chaser Plenoptic camera to track its movement. The collected data consists of 10 trajectories, with the chaser moving approximately half a meter closer to the target. For each trajectory, 10 frames are captured.

2) Scenario 2: In this scenario, the Earth is introduced in the background. As previously mentioned, the Jason 1 target can be rotated and its initial orientation adjusted. To expand the diversity of the dataset, we re-recorded all previous trajectories with different initial poses and with the Earth as the background. The camera depth estimation can be observed in Figure 8.

The dataset is primarily divided into three groups: one group 60% for training, the second 20% for validation, and the last one 20% for testing. The entire dataset consists of $1,000$ Plenoptic Lytro camera frames, 900 frames from scenario 2, and 100 frames from scenario 1. The frames were captured using the following steps: initially, we positioned the target mock-up satellite with no orientation displacement relative to the camera frame, which aligns with the world frame. We then recorded 10 different trajectories from this baseline pose. Subsequently, we systematically adjusted the target's orientation by rotating it around the Y-axis by increments of $\frac{2\pi}{10}$ radians for each subsequent set of trajectories. This procedure was repeated until the target had been rotated by a full 2π radians, ensuring comprehensive coverage of potential observation directions. Each increment aimed to simulate distinct observational angles of the target, enhancing the robustness and generalization capability of the network across varying pose estimations. In this study, the light source was stationary and maintained a constant position relative to the target throughout the dataset creation process.

C. Training phase

The training of the CaDNET model involves optimizing the model's parameters to minimize the error between the predicted and the ground truth poses. This process is done through an iterative optimization algorithm called stochastic gradient descent (SGD), where the model is updated based on the error of the prediction on a batch of training data. The training process is monitored using the root mean squared error (RMSE) between the predicted and ground truth poses, as well as the loss function that is used to optimize our CaDNET model. It is important to keep in mind that overfitting, where the model memorizes the training data but fails to generalize to unseen testing data, can occur during the training process. To avoid this situation few hyperparameters of the network training have been optimized leading to the final values given in Table III. The computational time of the CaDNET to run an estimation of one frame is $0.389440(s)$ on an Intel® Core™ i5-7400 CPU, with NVIDIA GeForce GTX 1060 6GB.

TABLE III
TRAINING PARAMETERS FOR CADNET

Parameter	CNN	DepthCNN	RNN
Max Epochs	3000	1000	300
Mini Batch Size	10	10	5
Initial Learning Rate	0.002	0.002	0.1
Gradient Threshold	1	1	1

In the training phase, each component of the CaDNET architecture—CNN, DepthCNN, and RNN—is trained independently with specific parameters optimized for its role. The label data used to train the three components is the ground truth pose associated with each frame. The CNN, being a crucial element for initial feature extraction, is trained for 3000 epochs to ensure robust feature detection. This training allows for a more thorough learning process, using a mean squared error (MSE) loss function to accurately predict continuous output variables.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2, \tag{5}$$

where N is the number of responses, t_i is the target output, and y_i is the network's prediction for response i.

Similarly, DepthCNN, which focuses on capturing depth-related features, is also trained for 1000 epochs. On the other hand, the RNN is trained in the same way with 300 epochs.

After training each network component with the outlined parameters, the CaDNET model integrates these trained models as described in Table III. This modular training approach allows for specialized optimization of different network parts, ensuring that each component performs its function effectively within the entire network architecture.

D. Estimation results

In Figure 9, we present the 3D estimation of the camera pose relative to the target along a single test trajectory from scenario 1. On the left, the results of our proposed CaDNET approach are compared to the performance of a CNN-RNN(RGB) architecture without depth information. For visualization and comparison, the ground truth trajectory (GT) is plotted as well. The right figure displays the 2D y-z camera motion, with the CaDNET pose estimation being compared to the CNN-RNN(RGB).

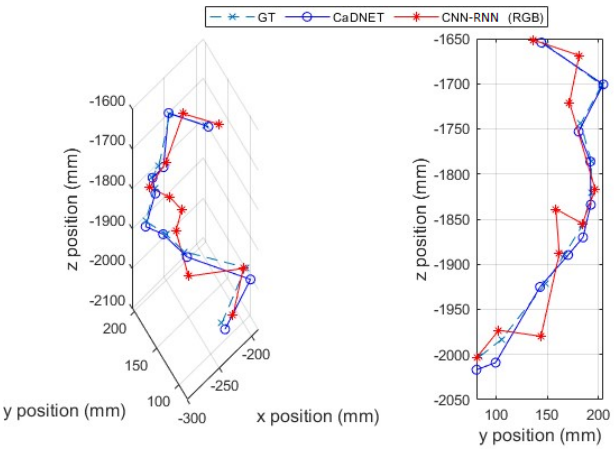


Fig. 9. First test trajectory estimation

Figure 10 showcases the estimation of the translation along the X, Y, and Z axes throughout the entire trajectory. The CaDNET approach outperforms the CNN-RNN(RGB) estimation without depth information and provides more accurate estimation of the translational pose.

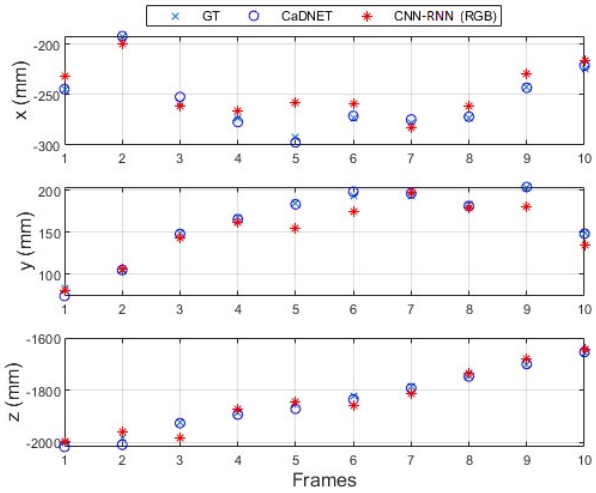


Fig. 10. Translational motion estimation for the first test trajectory

Similarly, the estimation results for the second scenario test trajectory are presented in Figure 11, and Figure 12.

The accuracy of our pose estimation improves as the camera approaches the target. In Figures 9 and 11, the trajectories start from approximately $-2000(mm)$ from the target and progress to about $-1600(mm)$. Initially, the accuracy at the start distances is lower. As the camera moves closer, the BiLSTM module becomes more effective. The BiLSTM is designed to leverage sequential data, enhancing its capability to refine pose estimations by incorporating temporal consistency and learning from the progression of frames. This demonstrates the important role of the RNN in adapting and optimizing the pose estimation performance.

Based on Figures 13, 14 that compare the translational and orientation error for the entire test trajectories of the CaDNET approach with the CNN-RNN(RGB) approach without depth information, respectively, several possible observations can be

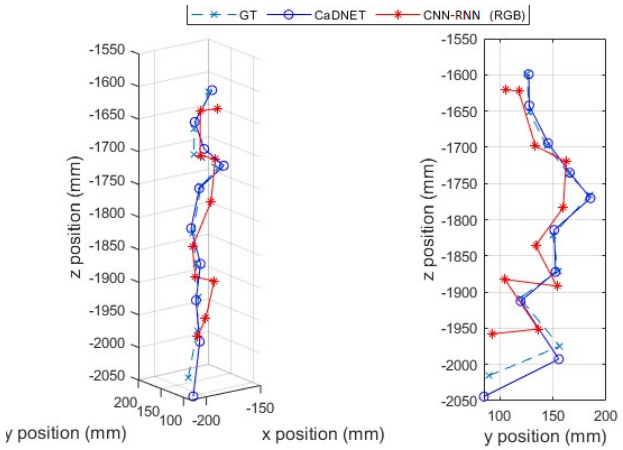


Fig. 11. Second test trajectory estimation

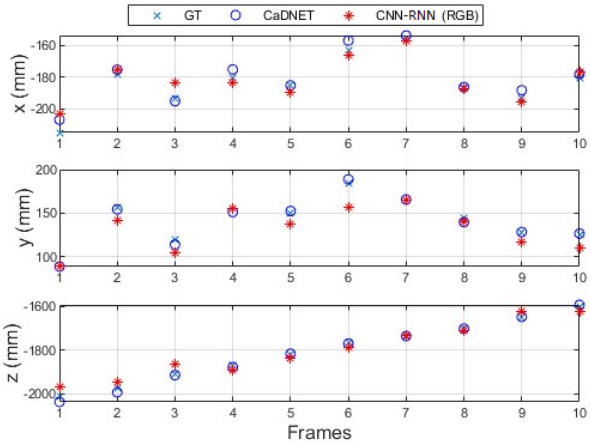


Fig. 12. Translational motion estimation for the second test trajectory

made: the performance of the CaDNET approach compared to CNN-RNN(RGB) approach in terms of the position error and orientation error is highlighted and the CaDNET approach outperforms the CNN-RNN(RGB) approach. The CaDNET is more accurate in the orientation estimation as it varies from approximately between $\pm 7(^{\circ})$ reaching the maximum when complex motions between two consecutive frames, as evidenced in test trajectory 1 between frames 2 and 3. Overall, the figures show valuable insights into the performance of the CaDNET compared to the CNN-RNN(RGB) approach and the impact of Plenoptic depth information on the pose estimation process.

Table IV shows the results of the two approaches, CaDNET and CNN-RNN(RGB), for the test trajectories in terms of the drift metric, which is measured in millimeters. In our trials, the drift is the RSS "Root-Sum-Square", between the estimated point and the ground truth (GT) point. The table shows that CaDNET has a lower drift compared to CNN-RNN(RGB) for all three axes.

Table V compares the performance of the architectures in terms of the global RMSE. The results show that CaDNET outperforms CNN-RNN(RGB) with a lower global RMSE value of 255.1 compared to 552.2.

Figure 15 illustrates the translational error across randomly

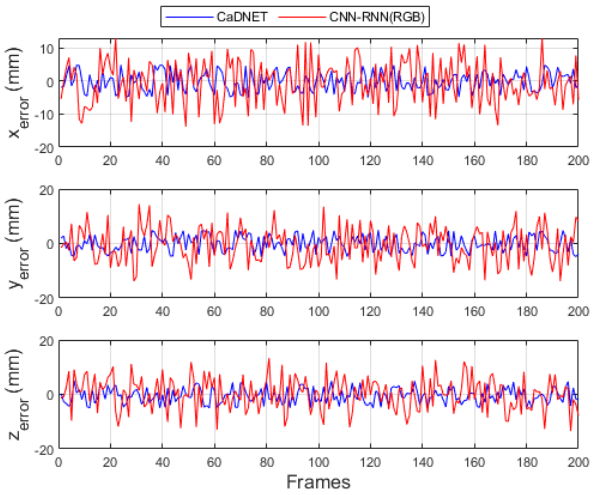


Fig. 13. Translational error throughout the entire test dataset

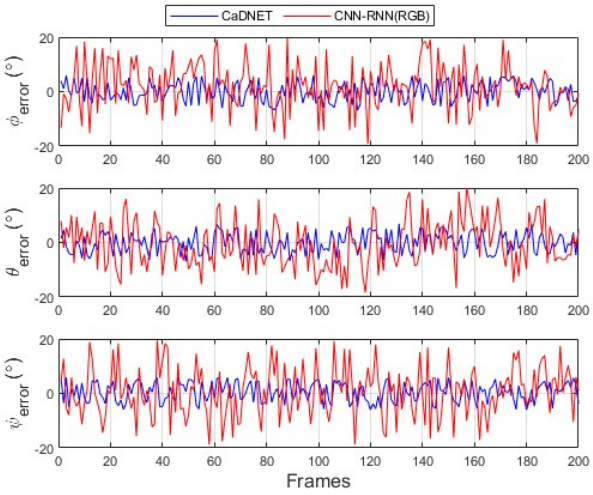


Fig. 14. Orientation error throughout the entire test dataset

TABLE IV
TEST TRAJECTORIES DRIFT

	Drift$_X$ (mm)	**Drift$_Y$ (mm)**	**Drift$_Z$ (mm)**
CaDNET	234.2	244.2	776.9
CNN-RNN(RGB)	793.4	735.4	1582.7

TABLE V
GLOBAL RMSE

Global RMSE in (mm)	**CaDNET**	**CNN-RNN(RGB)**
$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2}$	255.1	552.2

selected frames from the test dataset. It compares the translational estimation performance of the proposed CaDNET against the baseline cascaded CNNs architecture, which includes Depth CNN and RGB CNN without the RNN component. The comparative analysis highlights the importance of incorporating the RNN module, specifically a BiLSTM, in our model. The RNN's ability to effectively model the dynamics of navigation kinematics is evident, as demonstrated by the reduced translational errors with the CaDNET approach. Without the RNN, the CNN-only model exhibits noticeably higher errors, showing the RNN's role in enhancing the accuracy of sequential data processing within the context of pose estimation.

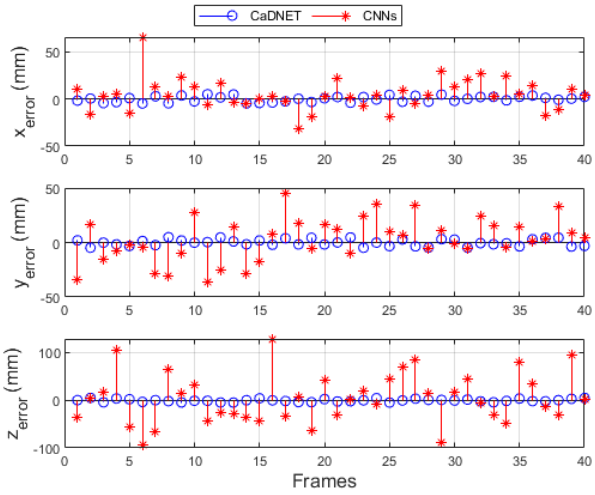


Fig. 15. Translational error throughout randomly selected frames

V. CONCLUSION

This paper highlights the effectiveness of a deep learning approach called Cascaded Deep Network (CaDNET) for pose estimation in space rendezvous operations using a focused Plenoptic camera. In this 'end-to-end' method, the network is trained directly on images to predict the relative pose. This simplifies the process and directs the focus more toward designing the network architecture and optimizing the parameters. Additionally, the use of a Plenoptic camera provides several advantages, including improved image quality and the ability to generate enhanced depth maps, making it a good option for autonomous navigation challenges.

The experimental validation conducted using a Plenoptic Lytro camera demonstrates the feasibility of the proposed approach for close-range, uncooperative rendezvous missions. The results support the adoption of a Plenoptic camera in AI-based relative navigation solutions for space rendezvous missions. However, it is important to acknowledge certain limitations of the CaDNET approach. The system's performance is heavily dependent on the quality and variability of the training data. In scenarios where the camera faces extreme lighting conditions or highly reflective surfaces, the accuracy of pose estimation may degrade. Additionally, the computational requirements for processing high-resolution Plenoptic images in real-time can be substantial, posing challenges for onboard spacecraft systems with limited processing capabilities.

Future work will focus on enhancing the robustness of the CaDNET model under varied and challenging environmental conditions. We plan to integrate more diverse datasets, including those simulated under extreme conditions. Furthermore, efforts will be directed toward optimizing the computational efficiency of the model, making it feasible for deployment on spacecraft with constrained computational resources.

REFERENCES

[1] Wigbert Fehse. *Automated rendezvous and docking of spacecraft*, volume 16. Cambridge university press, 2003.

[2] Özgün Yılmaz, Nabil Aouf, Elena Checa, Laurent Majewski, and Manuel Sanchez-Gestido. Thermal analysis of space debris for infrared-based active debris removal. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 233(3):811–822, 2019.

[3] O Yilmaz, Nabil Aouf, L Majewski, M Sanchez-Gestido, and G Ortega. Using infrared based relative navigation for active debris removal. 2017.

[4] Duarte Rondao and Nabil Aouf. Multi-view monocular pose estimation for spacecraft relative navigation. In *2018 AIAA Guidance, Navigation, and Control Conference*, page 2100, 2018.

[5] Duarte Rondao, Nabil Aouf, and Mark A Richardson. Chinet: Deep recurrent convolutional learning for multimodal spacecraft pose estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 2022.

[6] Sumant Sharma, Jacopo Ventura, and Simone D'Amico. Robust model-based monocular pose initialization for noncooperative spacecraft rendezvous. *Journal of Spacecraft and Rockets*, 55(6):1414–1429, 2018.

[7] Lorenzo Pasqualetto Cassinis, Robert Fonod, and Eberhard Gill. Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft. *Progress in Aerospace Sciences*, 110:100548, 2019.

[8] Martin Lingenauber, Klaus H Strobl, Nassir W Oumer, and Simon Kriegel. Benefits of plenoptic cameras for robot vision during close range on-orbit servicing maneuvers. In *2017 IEEE Aerospace Conference*, pages 1–18. IEEE, 2017.

[9] Christian Perwass and Lennart Wietzke. Single lens 3d-camera with extended depth-of-field. In *Human vision and electronic imaging XVII*, volume 8291, pages 45–59. SPIE, 2012.

[10] Olivier Dubois-Matra, Massimo Casasco, Manuel Sanchez Gestido, and Irene Huertas Garcia. Esa technology developments in vision-based navigation. In *IUTAM Symposium on Optimal Guidance and Control for Autonomous Systems*, pages 39–50. Springer, 2023.

[11] Martin Lingenauber, Florian A Fröhlich, Ulrike Krutz, Christian Nissler, and Klaus H Strobl. In-situ close-range imaging with plenoptic cameras. In *2019 IEEE Aerospace Conference*, pages 1–16. IEEE, 2019.

[12] Tim Michels, Arne Petersen, Luca Palmieri, and Reinhard Koch. Simulation of plenoptic cameras. In *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2018.

[13] Christopher Hahne and Amar Aggoun. Plenopticam v1. 0: A light-field imaging framework. *IEEE Transactions on Image Processing*, 30:6757–6771, 2021.

[14] Donald G Dansereau, Ian Mahon, Oscar Pizarro, and Stefan B Williams. Plenoptic flow: Closed-form visual odometry for light field cameras. In *2011 IEEE/RSJ international conference on intelligent robots and systems*, pages 4455–4462. IEEE, 2011.

[15] N Zeller, F Quint, and U Stilla. Narrow field-of-view visual odometry based on a focused plenoptic camera. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):285, 2015.

[16] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams. Linear volumetric focus for light field cameras. *ACM Trans. Graph.*, 34(2):15–1, 2015.

[17] Donald G Dansereau, Daniel L Bongiorno, Oscar Pizarro, and Stefan B Williams. Light field image denoising using a linear 4d frequency-hyperfan all-in-focus filter. In *Computational Imaging XI*, volume 8657, pages 176–189. SPIE, 2013.

[18] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016.

[19] Mate Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Märtens, and Simone D'Amico. Satellite pose estimation challenge: Dataset, competition design, and results. *IEEE Transactions on Aerospace and Electronic Systems*, 56(5):4083–4098, 2020.

[20] Jianing Song, Duarte Rondao, and Nabil Aouf. Deep learning-based spacecraft relative navigation methods: A survey. *Acta Astronautica*, 191:22–40, 2022.

[21] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[22] Pedro F Proença and Yang Gao. Deep learning for spacecraft pose estimation from photorealistic rendering. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6007–6013. IEEE, 2020.

[23] Michele Bechini, Michèle Lavagna, and Paolo Lunghi. Dataset generation and validation for spacecraft pose estimation via monocular images processing. *Acta Astronautica*, 2023.

[24] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[25] Philip TG Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, volume 6, pages 10–11, 2019.

[26] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[27] Lorenzo Pasqualetto Cassinis, Alessandra Menicucci, Eberhard Gill, Ingo Ahrns, and Manuel Sanchez-Gestido. On-ground validation of a cnn-based monocular pose estimation system for uncooperative spacecraft: Bridging domain shift in rendezvous scenarios. *Acta Astronautica*, 196:123–138, 2022.

[28] Tae Ha Park, Sumant Sharma, and Simone D'Amico. Towards robust learning-based pose estimation of noncooperative spacecraft. *arXiv preprint arXiv:1909.00392*, 2019.

[29] Kevin Black, Shrivu Shankar, Daniel Fonseka, Jacob Deutsch, Abhimanyu Dhir, and Maruthi R Akella. Real-time, flight-ready, non-cooperative spacecraft pose estimation using monocular imagery. *arXiv preprint arXiv:2101.09553*, 2021.

[30] Shaodong Zhang, Weiduo Hu, Wulong Guo, and Chang Liu. Neural-network-based pose estimation during noncooperative spacecraft rendezvous using point cloud. *Journal of Aerospace Information Systems*, pages 1–11, 2023.

[31] Odysseas Kechagias-Stamatis, Nabil Aouf, Vincent Dubanchet, and Mark A Richardson. Deeplo: Multi-projection deep lidar odometry for space orbital robotics rendezvous relative navigation. *Acta Astronautica*, 177:270–285, 2020.

[32] Zakaria Chekakta, Abdelhafid Zenati, Nabil Aouf, and Olivier Dubois-Matra. Robust deep learning lidar-based pose estimation for autonomous space landers. *Acta Astronautica*, 201:59–74, 2022.

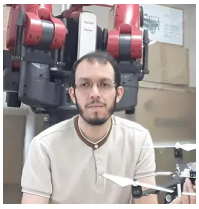
[33] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1027–1034, 2013.

[34] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):287–300, 2016.

[35] Mikael Le Pendu and Aljosa Smolic. High resolution light field recovery with fourier disparity layer completion, demosaicing, and super-resolution. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020.

[36] Pierre David, Mikaël Le Pendu, and Christine Guillemot. White lenslet image guided demosaicing for plenoptic cameras. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2017.

[37] Pierre Matysiak, Mairéad Grogan, Mikaël Le Pendu, Martin Alain, and Aljosa Smolic. A pipeline for lenslet light field quality enhancement. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 639–643. IEEE, 2018.



Zakaria Chekakta received his Ph.D. in Electronics and Industrial System Control from the National Polytechnic School of Oran, Algeria, in 2021. He is currently a Postdoctoral Research Fellow in Intelligent Navigation for Space with the Robotics and Machine Intelligence Group at City, University of London, U.K. Zakaria has more than three years of experience in the space sector, having contributed to various high-profile projects funded by the European Space Agency (ESA) and the UK Space Agency. His work includes leading the development of AI-based autonomous refueling systems for satellites and advancing deep neural network-based navigation for space landing operations.

Nabil Aouf received the Ph.D. degree in robust control for aerospace vehicles from the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, USA, in 2002. He is currently a Professor of Autonomous Systems and Machine Intelligence with the City University of London, London, U.K., where he is also the Director of the Systems, Autonomy and Control Centre and the co-Director of the London Space Institute. He also leads the Robotics, Autonomy and Machine Intelligence Group.