IEEE Transactions on Aerospace and Electronic Systems, Volume 59, Issue 2, April 2023, pp. 937-949
DOI:10.1109/TAES.2022.3193085

ChiNet: Deep Recurrent Convolutional Learning for Multimodal Spacecraft Pose Estimation

Duarte Rondao, Nabil Aouf, and Mark A. Richardson

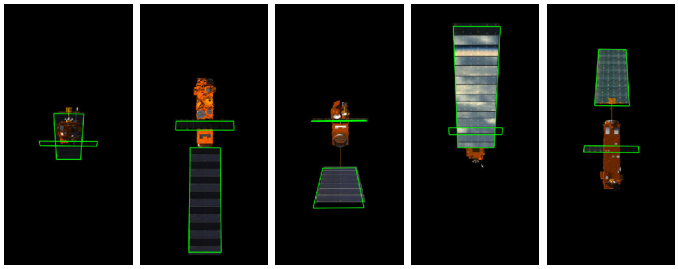
***Abstract*—This paper presents an innovative deep learning pipeline which estimates the relative pose of a spacecraft by incorporating the temporal information from a rendezvous sequence. It leverages the performance of long short-term memory (LSTM) units in modelling sequences of data for the processing of features extracted by a convolutional neural network (CNN) backbone. Three distinct training strategies, which follow a coarse-to-fine funnelled approach, are combined to facilitate feature learning and improve end-to-end pose estimation by regression. The capability of CNNs to autonomously ascertain feature representations from images is exploited to fuse thermal infrared data with electro-optical red-green-blue (RGB) inputs, thus mitigating the effects of artifacts from imaging space objects in the visible wavelength. Each contribution of the proposed framework, dubbed ChiNet, is demonstrated on a synthetic dataset, and the complete pipeline is validated on experimental data.**

I. Introduction

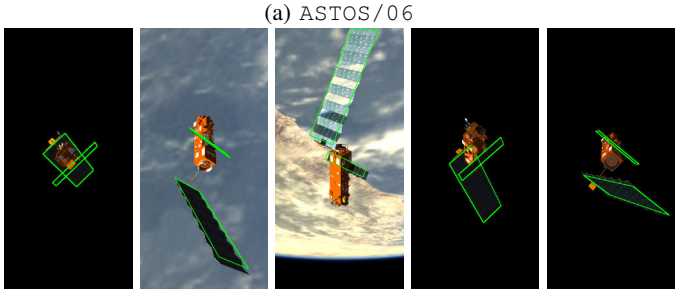
SPACECRAFT relative pose estimation is the problem of determining the rigid transformation between two space bodies – one of which is controllable and carries the navigation sensors – in terms of their relative position and attitude. This is a requirement for close-range rendezvous (RV) which has traditionally been solved using active sensors such as lidar [1]; the task is significantly hampered when the target is said to be non-cooperative, i.e. it does not bear any supportive equipment towards the RV [2].

Non-cooperative rendezvous (NCRV) operations involve the management of large relative velocities and minimal reaction times, justifying the need for autonomous operations and redundant sensors. As such, compact and lightweight passive digital cameras have become the cost-effective sensor for the task. Accordingly, the last couple of decades have focused on the development of robust image processing (IP) and machine learning (ML) techniques to accurately estimate the target's six degree-of-freedom (DOF) pose from images obtained aboard the chaser [3]. As the target is generally known beforehand, the followed strategies often choose to solve the model-to-image registration problem, under which the pose is retrieved via perspective-n-point (PnP [4]) and RANSAC-based (random sample consensus [5]) methods from correspondences between two-dimensional image features and three-dimensional model points. The challenge lies in robustly retrieving these correspondences in the face of hindering conditions such as shadows and sun glare, tumbling targets, or unknown initial poses. The former have been tackled in ground-based systems through multimodal sensing, but the fusion of each wavelength typically requires hand-crafted features, making its execution challenging [6, 7].

On the other hand, it represents an area with the potential of largely benefiting from DNN-based (deep neural network) estimation methods. In particular, convolutional neural networks (CNNs [8]) are naturally tailored to process such image inputs: here, the IP task is shifted completely to the network, and the effort becomes concentrated towards parameter optimisation and data modelling, allowing for the generalisation of the model to a wider swath of imaging conditions. The popularity of CNNs permeated onto the field of spacecraft relative pose estimation for rendezvous near the end of the past decade, mainly due to the European Space Agency (ESA) Kelvins Satellite Pose



(a) ASTOS/06



(b) ASTOS/13

Fig. 1: Qualitative results of the proposed method on two simulated NCRV sequences with Envisat from the Astos dataset. ChiNet provides continuous and robust pose estimation throughout the whole trajectories, explicitly taking into account information from previous frames.

D. Rondao is a Postdoctoral Research Fellow with the Department of Electrical and Electronic Engineering at City, University of London, EC1V 0HB, UK (e-mail: duarte.rondao@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.
M.A. Richardson is a Professor of Electronic Warfare with the Centre for Electronic Warfare, Information and Cyber at Cranfield University, Defence Academy of the United Kingdom, SN6 8LA, UK.

Estimation Challenge (SPEC),[1] where the vast majority (if not all) of the competitors used DNN-based approaches. SPEC benchmarked the participating algorithms on the Spacecraft PosE Estimation Dataset (SPEED [9]), which consists of images of the Tango satellite generated under unrelated randomised poses. However, during an RV sequence, it is expected that the pose of the observed target continually varies as the operation progresses, i.e. the poses are correlated through time.

This paper proposes the adoption of a recurrent neural network (RNN) module to process the features extracted by a CNN front-end model and to exploit this temporal correlation between acquired image frames in NCRV sequences. The resulting deep recurrent convolutional neural network (DRCNN) architecture, dubbed ChiNet,[2] is shown to provide a smoother and lower-error estimate of the 6-DOF pose when compared to a single CNN (Fig. 1 illustrates qualitative results on two NCRV sequences). Furthermore, ChiNet proposes a new three-step training regimen to learn features in a coarse-to-fine manner, which is inspired from traditional ML approaches. Lastly, ChiNet also explores the impact of multimodal sensing in the pose estimating by augmenting the number of input channels to the network with images from a long wavelength infrared (LWIR) camera, thus exploiting the natural ability of CNNs to autonomously extract features from images. The following contributions are proposed, to the best of the authors' knowledge: 1) The work represents the first use of RNNs, in particular long short-term memories (LSTMs), to tackle the problem of spacecraft pose estimation for RV using 2D images as the sole navigation input; 2) It is also the first to explore the potential benefit of a multimodal sensor input for the task, in particular in the visible and LWIR modalities, leveraging the power of deep learning to formulate it as an optimal process and surpassing the hurdles of classical approaches; and 3) A multi-step optimisation approach to DNN training is devised to facilitate the learning and reduce the overall estimation error.

The paper is organised as follows. Section II surveys the literature to highlight pertinent related work. Section III thoroughly details the methodology of each proposed contribution. Section IV presents the attained results. Lastly, Section V shares the conclusions of the work.

II. Related Work

This section briefly summarises the existing model-based literature on spacecraft pose estimation with monocular cameras, i.e. when the target is known. It is broadly divided into two categories: methods based on geometry and methods based on learning (with a focus on DNNs).

A. Geometry-based Methods

These methods estimate the 4×4 relative pose matrix $\boldsymbol{T} = \boldsymbol{T}_{ct} \in \mathrm{SE}(3)$ (the Special Euclidean Group in three dimensions) relating the target body-fixed reference frame $\mathcal{F}_t$ to the camera frame $\mathcal{F}_c$, which is attached to the chaser, from model points $\boldsymbol{p}^{(i)} \in \mathbb{R}^3$ expressed in $\mathcal{F}_t$ and their image plane projections $\boldsymbol{z}^{(i)} \in \mathbb{R}^2$ expressed in $\mathcal{F}_c$, which are related according to the perspective projection model [4]:

$$\boldsymbol{z}^{(i)} = \pi\left(\boldsymbol{K}\begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}\begin{bmatrix}\boldsymbol{p}^{(i)} \\ 1\end{bmatrix}\right), \tag{1}$$

where $i \in \{1, \ldots, N\}$ and the projective function $\pi(\boldsymbol{a}) \coloneqq \boldsymbol{a}_{1:2}/a_3$ has been defined. Here, $\boldsymbol{R}, \boldsymbol{t}$ are the 3×3 attitude matrix and 3×1 position vector composing $\boldsymbol{T}$, and $\boldsymbol{K}$ is the 3×3 intrinsic camera matrix accounting for the focal length f obtained *a priori* via calibration.

Equation (1) can be solved in closed form for $N \geq 4$ using a PnP [4] solver, while using RANSAC [5] to reject spurious matches. Alternatively, it can be solved iteratively via robust estimation [10]. Arguably, the biggest challenge resides in matching $\boldsymbol{z}^{(i)}$ to $\boldsymbol{p}^{(i)}$.

Tracking by recursion [11] was initially adopted as a popular solution in which 3D control points from a computer-aided design (CAD) model of the target are projected onto the image using the expected pose accompanied by a gradient-based scan to locate the corresponding 2D feature. Initially limited to edge features [12], the technique was later adapted to include other features such as colours [13] and keypoints [14] at the expense of requiring hardware acceleration to deal with complex models.

Conversely, tracking by detection entails an offline stage where a database of target feature points, whose positions on the surface are known, is built. Matching is then performed using heuristics exploiting the grouping of local model features and multiple hypotheses [15, 16]; the pose and correspondence problems may also be solved concurrently at a higher computational cost [17]. An alternative approach constructs a database by discretising the 3D object into 2D keyframes representing multiple viewpoints [18], and then using local keypoint detectors and descriptors (e.g. SIFT [19], SURF [20], or the more modern ORB [21]) to obtain the matches.

Both tracking by detection and by recursion have been applied to spacecraft pose estimation in the LWIR [22, 23], and to model-free estimation in general [24]. While the latter leverages the increased repeatability of LWIR features with respect to the visible band [25], the former applications do not explicitly make use of such advantages, leaving a gap in the literature for this modality.

B. Learning-based Methods

These methods also estimate the pose $\boldsymbol{T}$ but do not necessarily make use of Eq. (1) or local features, instead exploring patterns in training data to generalise towards previously unseen query images. A coarser estimation of $\boldsymbol{T}$ can also be considered in order to initialise tracking by recursion methods or to reduce the search-space in tracking by detection.

Generally, global features (e.g. bags of keypoints, shapes, or even raw images) have been preferred for combination with a variety of ML techniques ranging from nearest neighbour search [26] to unsupervised clustering [27], principal component analysis [28], Bayesian classification [29], and deep learning [30].

[1] https://kelvins.esa.int/satellite-pose-estimation-challenge.

[2] 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".

The recent prevalence of the latter with respect to the others originated from SPEC in 2019. As reported by Kisantal *et al.* [9], the majority of the participating teams used CNNs to directly predict the relative pose of the target in an end-to-end, regressive fashion from each raw image (e.g. [31]). The attitude estimation was noted to be the most challenging, and was improved in approaches which first included a target localisation step (e.g. [32]). However, the best-performing entries, including those who won 1st and 2nd places, followed instead an indirect approach where the role of the CNN was relayed completely towards the prediction of keypoints in the image, pre-selected by the human-in-the-loop, which were then used with PnP to recover the pose [33].

After SPEC, published DNN-based work has seldom considered actual rendezvous trajectories [34], continuing to focus instead on individual greyscale images of SPEED [35–37]. In either case, the proposed strategies consist in using a CNN for keypoint detection for use with PnP. Additionally, the contribution of modalities beyond the visible remains to be fully investigated [38].

Contrary to the above examples, ground-based applications have recently adopted the use of RNNs combined with features extracted by CNN front-ends to model the intrinsic motion dynamics from sequences of imaging data rather than individual inputs [39, 40]; more specifically, these proposed LSTM-based [41] DRCNNs for visual odometry (VO) to estimate a car's egomotion. Kechagias-Stamatis *et al.* [42] introduced DeepLO, which followed the same philosophy for lidar-based relative navigation with a non-cooperative space target. Lidar data was preprocessed by quantisation and projection onto each plane in the target body frame of reference, thus creating three 2D depth images to be processed by a regular CNN.

III. Methodology

This section describes in detail the proposed DRCNN framework for end-to-end spacecraft pose estimation. The CNN and RNN modules are both described, as well as the multistage optimisation strategy to train them.

A. System Architecture

The results from SPEC have shown promising results in the use of CNNs for the task. However, the current literature treats each incoming image as a separate input, thus ignoring the intrinsic temporal correlation between them. Therefore, the main focus here is the investigation of the feasibility of a DRCNN for estimating the pose in rendezvous sequences. The problem has been previously studied by Kechagias-Stamatis *et al.* [42] for VO with lidar map inputs, but not for images. Furthermore, VO is concerned with estimating the motion between two time-consecutive images, but during an RV a single acquired image contains enough information relating $\mathcal{F}_t$ to $\mathcal{F}_c$. This work recognises this as a requirement and as such considers it for the DRCNN formulation.

The architecture of the proposed framework is schematically depicted in Figure 2. The pipeline takes a four-dimensional red-green-blue-thermal (RGBT) image formed from the channel-wise concatenation of a visible image and a LWIR image. This multimodal image is then processed by a CNN, whose learned output features are modelled temporally (along the vertical axis in the figure) with an RNN. Two fully connected (FC) layers convert the output into position and attitude values forming the 6-DOF pose. Note that, unlike in DeepVO or DeepLO, ChiNet receives only a single target snapshot at a time, thus predicting the complete relative pose for each time-step $\tau = \tau_\kappa$. Additionally, since the front-end is fully convolutional, the network is capable of receiving inputs of arbitrary spatial dimensions (i.e. any width and height).

B. Deep Feature Extraction with Convolutions

CNN front-ends for feature extraction are typically chosen to be large but powerful architectures, such as ResNet [43] or Inception-v3 [44], and SPEC-admitted architectures were no exception. On the other hand, these networks are also characterised by elevated processing times and are potentially prone to overfitting due to their high number of parameters.

To mitigate this, ChiNet adopts the Darknet-19 architecture (backbone of the YOLO object detector [45]), with some modifications (Fig. 2, centre). First, the 3×3 kernel size on the first convolutional layer is replaced by a 7×7 one to adapt to image inputs larger than $224\,\text{px} \times 224\,\text{px}$. Second the network is modernised (bringing it closer to Darknet-53 [46]) by replacing all max pooling layers with a stride of 2 in the preceding convolution. Whereas the former is a fixed operation, the latter is learned, which further contributes to the adaptability of the network to the task at hand. In addition, residual connections are introduced but only in the channel expansion-contraction layers (green blocks in Fig. 2), thus avoiding the need to add extra 1×1 convolutions to keep the dimensions consistent. Lastly, a dropout layer [47] with probability $p = 0.2$ is added to further prevent overfitting.

1) Optimal Low-Level Sensor Fusion: ChiNet preprocesses images acquired separately by each camera via concatenation along the channel dimension, forming a four-channel RGBT image which the network takes as input. The first convolutional layer entails a weighted sum of the pixels in each channel, outputting new activation maps that effectively encompass the fused information. This is equivalent to a pixel (or low-level) fusion of the inputs resulting in a series of multimodal images upon which feature extraction is to be performed. Furthermore, these weights are not predefined but learned in the context of the network training procedure, thus being optimal in the sense of minimising the objective loss. This philosophy has been previously explored in VO applications using traditional IP techniques such as intensity level thresholding and discrete wavelet transforms, showing promising results [48]. ChiNet's approach, however, bypasses the need of manually developing a potentially sub-par weighting strategy to combine the multiple input modalities.

C. Temporal Sequence Modelling with LSTMs

The features learned by the CNN are post-processed by a deep RNN module that models the intrinsic temporal correlations coming from an ordered sequence of image inputs. This addition is expected to be beneficial to the problem of

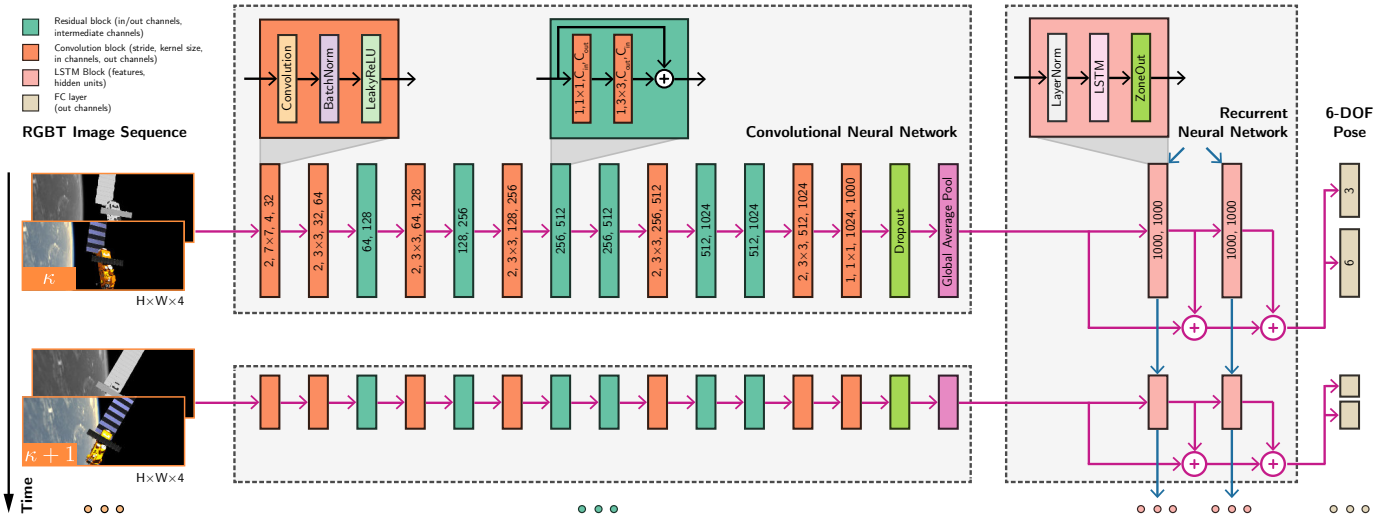


Fig. 2: ChiNet system overview. The proposed DRCNN architecture performs end-to-end spacecraft pose estimation from a sequence of multimodal RGBT image inputs of arbitrary size.

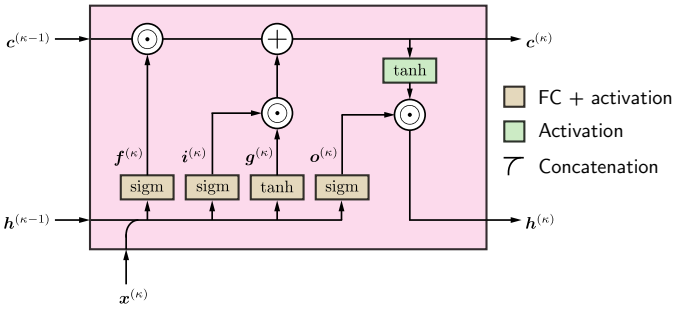


Fig. 3: Block diagram of a LSTM recurrent memory unit. sigm and tanh denote the sigmoid and hyperbolic tangent activation functions, respectively; $\odot$ and $+$ denote element-wise product and addition, respectively.

spacecraft pose estimation due to the inherent relative motion dynamics entailed, and the estimate of the solution for the current frame can benefit from the knowledge of previous frames: even more than in ground-based applications, the perceived motion of a space target during RV is not likely to change abruptly but can be represented as a smooth function of the previous states.

ChiNet's recurrent feature post-processing module is based on the LSTM architecture [41]. LSTMs were designed in an attempt to combat vital flaws in the capability of vanilla recurrent cells to model long sequences, as they suffered from vanishing and exploding gradients. The LSTM's ability to learn long-term dependencies is owed to its gated design that determines which sectors of the previous hidden state should be kept or discarded in the current iteration. This is achieved not only in combination with the current input, processed by four different units, but also by a cell state which acts as an "information motorway" that bypasses the cells. The LSTM structure is illustrated in Figure 3.

The design of the RNN is schematically depicted in Figure 2 (right). The CNN features are fed to two stacked LSTM layers with 1000 hidden states each; stacked LSTM layers have been previously adopted for architectures such as DeepVO [40] and DeepLO [42] and shown empirically to help in modelling complex motion dynamics.

Unlike FC or convolutional layers, data normalisation in LSTMs must be done internally due to the gated system topology. Batchnorm would be impractical both in terms of time and memory consumption since since this would require fitting one layer per time-step and storing the statistics of each one during training. In opposition, layer normalisation [49] is instead employed by computing the mean and variance across all the features of the i-th layer rather than across the batch dimension.

A second nuanced aspect pertains to dropout, typically applied as a binary mask to randomly nullify some of a layer's activations. In the case of LSTMs, however, stochasticity should be applied in the recurrent loop. More than that: rather than following a potentially naive dropout philosophy, ChiNet employs zoneout [50], which was specifically designed for RNNs. In zoneout, the values of the hidden state $\boldsymbol{h}^{(\kappa)}$ and memory cell $\boldsymbol{c}^{(\kappa)}$ are randomly expected to either maintain their previous value or are updated in the usual manner. The modified LSTM equations thus become:

$$\begin{bmatrix} \tilde{\boldsymbol{f}}^{(\kappa)} \\ \tilde{\boldsymbol{i}}^{(\kappa)} \\ \tilde{\boldsymbol{o}}^{(\kappa)} \\ \tilde{\boldsymbol{g}}^{(\kappa)} \end{bmatrix} = \mathrm{LN}\left(\boldsymbol{W}^{h}\boldsymbol{h}^{(\kappa-1)};\gamma_1,\beta_1\right) + \mathrm{LN}\left(\boldsymbol{W}^{x}\boldsymbol{x}^{(\kappa)};\gamma_2,\beta_2\right), \tag{2}$$

$$\begin{aligned} \boldsymbol{c}^{(\kappa)} = \boldsymbol{d}^{c,(\kappa)} \odot \boldsymbol{c}^{(\kappa-1)} + \left(\mathbf{1} - \boldsymbol{d}^{c,(\kappa)}\right) \\ \odot \left(\boldsymbol{f}^{(\kappa)} \odot \boldsymbol{c}^{(\kappa-1)} + \boldsymbol{i}^{(\kappa)} \odot \boldsymbol{g}^{(\kappa)}\right), \end{aligned} \tag{3}$$

$$\begin{aligned} \boldsymbol{h}^{(\kappa)} = \boldsymbol{d}^{h,(\kappa)} \odot \boldsymbol{h}^{(\kappa-1)} + \left(\mathbf{1} - \boldsymbol{d}^{h,(\kappa)}\right) \\ \odot \left(\boldsymbol{o}^{(\kappa)} \odot \tanh\left\{\mathrm{LN}\left[\boldsymbol{c}^{(\kappa)};\gamma_3,\beta_3\right]\right\}\right), \end{aligned} \tag{4}$$

where $\boldsymbol{f}, \boldsymbol{i}, \boldsymbol{o}, \boldsymbol{g}$ are the forget, input, output, and modulation gates, respectively; $\boldsymbol{h}$ is the hidden state; $\boldsymbol{x}$ is the input; $\boldsymbol{W}^{h^\top} = [\,\boldsymbol{W}^{hf^\top}\ \boldsymbol{W}^{hi^\top}\ \boldsymbol{W}^{ho^\top}\ \boldsymbol{W}^{hg^\top}\,]$ is the recurrent weights matrix; $\boldsymbol{W}^{x^\top} = [\,\boldsymbol{W}^{xf^\top}\ \boldsymbol{W}^{xi^\top}\ \boldsymbol{W}^{xo^\top}\ \boldsymbol{W}^{xg^\top}\,]$ is the input weights matrix; $\boldsymbol{f} = \mathrm{sigm}(\tilde{\boldsymbol{f}})$; $\boldsymbol{i} = \mathrm{sigm}(\tilde{\boldsymbol{i}})$; $\boldsymbol{o} = \mathrm{sigm}(\tilde{\boldsymbol{o}})$; $\boldsymbol{g} = \tanh(\tilde{\boldsymbol{g}})$; sigm is the sigmoid nonlinear activation function; tanh is the hyperbolic tangent activation function; LN denotes layer normalisation with scale γ and offset β; $\boldsymbol{d}^c, \boldsymbol{d}^h$ are the binary cell and hidden state zoneout masks, respectively; $\mathbf{1}$ is a vector of ones of appropriate length; the superscript (κ) denotes a variable at time-step $\tau = \tau_\kappa$; and $\odot$ denotes an element-wise product operation.

Residual connections have also been implemented (see Fig. 2, right), drawing inspiration from the CNN front-end itself. During preliminary experiments, it was found that the addition of residual connections to the LSTMs in ChiNet resulted in faster training convergence and overall lower pose estimation error.

D. Multistage Optimisation

Instead of pursuing an indirect approach (i.e. DNN to predict keypoints followed by PnP), ChiNet provides an end-to-end, direct method to retrieve the pose. The former has been shown to produce the lowest error estimates in SPEC, suggesting that the latter may be harder to train. To mitigate this and lower the overall error in end-to-end approaches, a multistage, coarse-to-fine approach is proposed and described in this section.

Stage 1

The objective of Stage 1 is to emulate the benefits of transfer learning [51], in which the network is pre-trained on a set of tasks involving a large dataset and then used to initialise a same-sized network to solve the purported task that generally has fewer training examples. Transfer learning is advantageous for CNNs as these normally entail millions of parameters and thus may converge towards a suboptimal solution if the training data is not diverse enough.

A subset of 1000 object categories of ImageNet [52] is the typical go-to choice for pre-trained networks. However, the data is composed of red-green-blue (RGB) images and thus cannot be expanded for use with multimodal data. As such, a strategy to pre-train a CNN by artificially augmenting the number of samples based only on the actual training dataset is proposed.

This stage bypasses the RNN and the two FC layers are connected directly to the CNN's output. The procedure thus aims to first train the CNN on a simpler task to learn coarse features in terms of a discretised pose representation. The Special Orthogonal Group in three dimensions $\mathrm{SO}(3)$, or simply the attitude space, is divided into a spherical grid of discrete azimuth and elevation steps, centred on the target, of fixed radius, i.e. a 2-sphere S^2, or viewsphere. Each square on the grid then represents an attitude class $a_i \in \mathbb{Y}_{\mathrm{S}^2}$, $i = \{1, \ldots, K_{\mathrm{S}^2}\}$ with K_{S^2} possible classes depending on the square size. For the sake of succinctness, the reader is directed to Rondao *et al.* [29] for further details on the viewsphere. The position component is estimated in terms of the relative depth $\|\boldsymbol{t}\|$, thus maximising the joint conditional probability:

$$\boldsymbol{\theta}^{*(\mathrm{S1})} = \underset{\boldsymbol{\theta}^{(\mathrm{S1})}}{\arg\max}\, p\left(\|\boldsymbol{t}^{(\kappa)}\|, \boldsymbol{a}^{(i,\kappa)} \,\middle|\, \mathbf{I}^{(\kappa)}; \boldsymbol{\theta}^{(\mathrm{S1})}\right), \tag{5}$$

where $\boldsymbol{\theta}^{(\mathrm{S1})}$ are the CNN parameters learned in Stage 1, $\boldsymbol{a}^{(i)}$ is the one-hot vector encoding of a_i, and $\mathbf{I}$ is the image input. Note that thus far the learning depends only on each individual input at time $\tau = \tau_\kappa$, not yet exploiting the temporal correlation in the data.

Sequential images from an RV training sequence are preprocessed as follows. 1) First, the attitude space is discretised into the set $\mathbb{Y}_{\mathrm{S}^2}$ with K_{S^2} classes as mentioned above, discarding any unrepresented class. 2) Define a number N_{S^2} of desired observations per attitude class. 3) Similarly, K_t bins are defined for the relative position $\boldsymbol{t}$, selecting the edges according to the minimum and maximum values observed in the dataset, thus creating the set $\mathbb{Y}_t$. 4) For each attitude class $a_c \in \mathbb{Y}_{\mathrm{S}^2}$: 4-a) identify the subset $\mathbb{Y}'_t \subseteq \mathbb{Y}_t$ of K'_t depth bins that contain at least one observation; 4-b) randomly sample N_{S^2}/K'_t observations with attitude label a_c equally for each of the K'_t depth bins according to the position ground truth. Oversample if necessary.

The resulting Stage 1 dataset will have a total of $N_{\mathrm{S}^2} \cdot K_{\mathrm{S}^2}$ observations with equal representation. For the present application, N_{S^2} was chosen such that $N_{\mathrm{S}^2} \cdot K_{\mathrm{S}^2} = 10\,000$. It was found that having balanced attitude classes was paramount to prevent overfitting. To increase data variance in the case of oversampling, an online data augmentation pipeline was implemented, both in terms of visual filtering and small perturbations to the pose.

The loss is formulated as a multi-task learning problem with the attitude component represented by a cross-entropy function and the position component by a regression function, respectively, for each observation i:

$$\mathcal{L}_{\mathrm{S}^2}^{(\mathrm{S1})} = -\sum_i \sum_{c=1}^{K_{\mathrm{S}^2}} a_c^{(i)} \log\left(\hat{a}_c^{(i)}\right), \tag{6}$$

$$\mathcal{L}_t^{(\mathrm{S1})} = \sum_i \frac{\|\boldsymbol{t}^{(i)} - \hat{\boldsymbol{t}}^{(i)}\|}{\|\boldsymbol{t}^{(i)}\|}, \tag{7}$$

where $\hat{\boldsymbol{a}}^{(i)} = [\,\hat{a}_1^{(i)}\ \ \ldots\ \ \hat{a}_{K_{\mathrm{S}^2}}^{(i)}\,]^\top$ is the predicted attitude class encoding, and $\hat{\boldsymbol{t}}^{(i)} \in \mathbb{R}^3$ is the predicted position. In VO, the multi-task loss is typically achieved via linear combination of each component using manually tuned weights; however, as shown by Kendall *et al.* [53], this is a sub-optimal approach. Instead, ChiNet models each weight $\{\sigma_{\mathrm{S}^2}, \sigma_t\}$ as learnable task-specific variances of a Boltzmann distribution and a Gaussian distribution, respectively, yielding the combined loss:

$$\mathcal{L}^{(\mathrm{S1})} = \frac{1}{2}\mathcal{L}_{\mathrm{S}^2}^{(\mathrm{S1})} \exp\left(-2\hat{\sigma}_{\mathrm{S}^2}\right) + \mathcal{L}_t^{(\mathrm{S1})} \exp\left(-2\hat{\sigma}_t\right) + \hat{\sigma}_{\mathrm{S}^2} + \hat{\sigma}_t. \tag{8}$$

The reader is directed to Kendall *et al.* [53] for the details on the derivation[3] of Eq. (8).

Stage 2

Stage 2 represents ChiNet's nominal training phase of the whole structure, using the normal, non-modified dataset. The full DRCNN pipeline is trained to maximise the conditional probability of a series of time-sequential poses $\{\boldsymbol{T}^{(1)}, \dots, \boldsymbol{T}^{(\kappa)}\} \in \mathrm{SE}(3)$ given a sequence of RGBT images, i.e.:

$$\boldsymbol{\theta}^{*(\mathrm{S2})} = \underset{\boldsymbol{\theta}^{(\mathrm{S2})}}{\arg\max}\, p\left(\boldsymbol{T}^{(1)}, \dots, \boldsymbol{T}^{(\kappa)} \,\middle|\, \mathbf{I}^{(1)}, \dots, \mathbf{I}^{(\kappa)}; \boldsymbol{\theta}^{(\mathrm{S2})}\right), \tag{9}$$

where the CNN weights are initialised with the results of Stage 1. Special care must be taken for the representation of the attitude to ensure it remains a member of some group isomorphic to $\mathrm{SO}(3)$. A common approach is to admit the unit quaternion representation $\boldsymbol{q}$ (e.g. [31, 54]) due to the lack of singularities. However, this representation is not continuous due to its antipodal ambiguity (i.e. $\boldsymbol{q} = -\boldsymbol{q}$), which has been shown to introduce learning difficulties into the DNN and higher convergence errors.

Instead, ChiNet employs the 6D attitude representation $\boldsymbol{r} \in \mathbb{R}^6$ proposed by Zhou *et al.* [55] which admits a continuous mapping $\mathbb{R}^6 \leftarrow \mathrm{SO}(3)$. The transform $\boldsymbol{r} \mapsto \boldsymbol{R}$ entails reshaping $\boldsymbol{r}$ into a 3×2 matrix followed by Gram-Schmidt orthogonalisation;[4] the inverse transform thus consists in removing the right-most column of $\boldsymbol{R}$. This approach is similar to directly estimating the 9 parameters of $\boldsymbol{R}$ followed by incorporation of the orthogonalisation procedure inside the network, except with the major advantage of not having to estimate 3 superfluous parameters.

The Stage 2 loss is a combined loss based on the L^2 norm regression of $\boldsymbol{r}$ and $\boldsymbol{t}$:

$$\mathcal{L}_r^{(\mathrm{S2})} = \sum_{\kappa=1}^{T} \|\hat{\boldsymbol{r}}^{(\kappa)} - \boldsymbol{r}^{(\kappa)}\|, \;\; \mathcal{L}_t^{(\mathrm{S2})} = \sum_{\kappa=1}^{T} \|\hat{\boldsymbol{t}}^{(\kappa)} - \boldsymbol{t}^{(\kappa)}\|, \tag{10}$$

$$\mathcal{L}^{(\mathrm{S2})} = \mathcal{L}_r^{(\mathrm{S2})} \exp\left(-2\hat{\sigma}_r\right) + \mathcal{L}_t^{(\mathrm{S2})} \exp\left(-2\hat{\sigma}_t\right) + 2\left(\hat{\sigma}_r + \hat{\sigma}_t\right), \tag{11}$$

where Eq. (11) is derived similarly to Eq. (8) for two Gaussian distributions, and the temporal component has been highlighted in terms of the training sequence length T. Training very long sequences involves high memory requirements, so a truncated backpropagation through time (BPTT) procedure is adopted instead. This entails unfolding the sequence for a predefined number of time-steps T smaller than the full sequence length $\tilde{T}$, performing one training iteration, and then moving on to the next partition. In order to keep continuity while still allowing the network to learn long sequences, ChiNet follows the approach in [39] whereby the training is carried out with a sliding window over the sequence, where consistency is established by appropriately initialising the LSTMs's hidden states with those computed in the previous iteration.

Stage 3

The final training stage consists in a geometric refinement of the output from Stage 2, following the reprojection of 3D model points using the ground truth and predicted relative pose first proposed by Kendall *et al.* [56] for camera pose estimation in urban scenarios:

$$\boldsymbol{\theta}^{*(\mathrm{S3})} = \underset{\boldsymbol{\theta}^{(\mathrm{S3})}}{\arg\max}\, p\left(\boldsymbol{T}^{(1)}, \dots, \boldsymbol{T}^{(\kappa)} \,\middle|\, \mathbf{I}^{(1)}, \dots, \mathbf{I}^{(\kappa)}, \mathbb{P}; \boldsymbol{\theta}^{(\mathrm{S3})}\right), \tag{12}$$

where $\mathbb{P} = \{\boldsymbol{p}^{(1)}, \dots, \boldsymbol{p}^{(N)}\}$ is a manually selected set of $N \geq 4$ target model points expressed in $\mathcal{F}_t$. The loss is straightforwardly defined as:

$$\mathcal{L}^{(\mathrm{S3})} = \sum_{\kappa=1}^{T} \sum_{i=1}^{N} \left\| \boldsymbol{z}^{(i,\kappa)} - \pi\left(\boldsymbol{K}, \boldsymbol{T}, \boldsymbol{p}^{(i)}\right) \right\|, \tag{13}$$

where $\mathbb{Z}_\kappa = \{\boldsymbol{z}^{(1,\kappa)}, \dots, \boldsymbol{z}^{(N,\kappa)}\}$ is the set of projected keypoints corresponding to $\mathbb{P}$ at time $\tau = \tau_\kappa$, and π follows from Eq. (1). Similarly to Stage 2, the 6D attitude representation is used. Eq. (13) thus learns the pose implicitly via the minimisation of the reprojection error, which naturally balances the contributions of the position and attitude branches, and does not require defining explicit weights unlike Stages 1 and 2. This is advantageous for datasets in which the position depth has a high variance, since each contribution is weighed differently due to parallax, as reported in [56]. On the other hand, the loss formulation requires a good initialisation of the parameters $\boldsymbol{\theta}^{(\mathrm{S3})}$ to converge, hence why it is used as a refinement stage.

IV. Experimental Results

In this section, the performance of the proposed end-to-end DRCNN pipeline is evaluated on both synthetic and experimental data.

A. Synthetic Dataset

1) Description: The framework is initially validated on the Astos dataset, consisting of 14 different rendezvous trajectories with the failed satellite Envisat, featuring three distinct guidance profiles (GPs), three tumbling modes, and two approach vectors. The images are synthetically generated using the Astos Camera Simulator[5] with emulated visible and thermal cameras of characteristics expressed in Table I. The visible and LWIR images are aligned and resized to a resolution of $640\,\mathrm{px} \times 512\,\mathrm{px}$ for both training and testing. The reader is directed to Rondao *et al.* [25] for details on the chosen Envisat orbital parameters and image generation. Figure 4a illustrates the three considered GPs of the chaser expressed in the target's local-vertical-local-horizontal (LVLH) frame

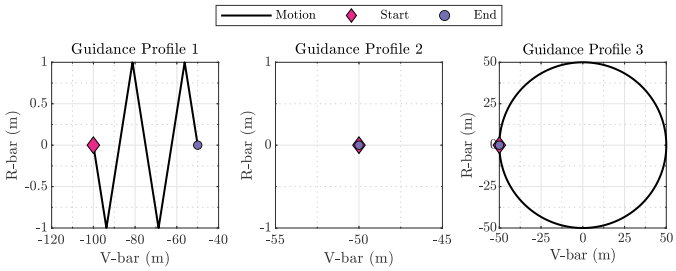
[3] Despite Eq. (7) not strictly representing the L^2 component of a Gaussian PDF due to the division by $\|\boldsymbol{t}^{(i)}\|$, the formulation of Eq. (8) yields good results in practice.

[4] This happens only at inference time and is not needed for training.

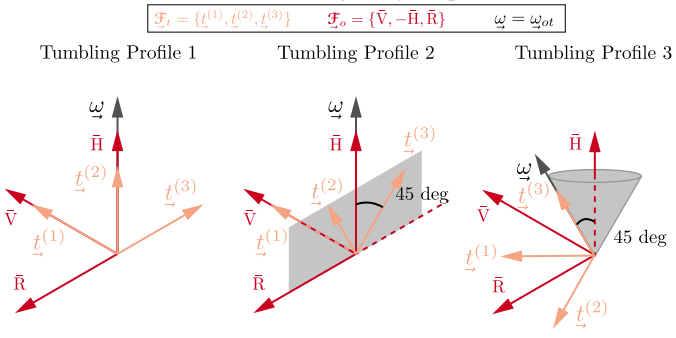
[5] https://www.astos.de/products/camsim.

TABLE I: Simulated camera parameters for the Astos dataset.

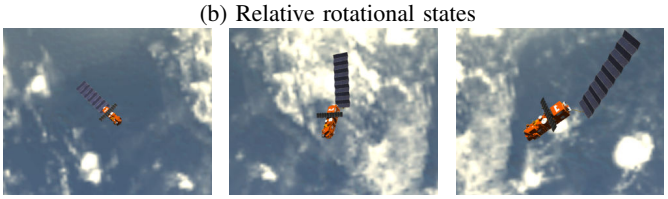
Parameter	Unit	Visible	LWIR
Camera	[-]	mvBlueFOX-MLC 202b	FLIR Tau2
Frame rate	[Hz]	10	10
Focal length	[mm]	5	13
Sensor width	[mm]	4.8	10.875
Sensor height	[mm]	3.6	8.7
Bit depth	[-]	24	8



(a) Relative trajectory shapes



(b) Relative rotational states



(c) ASTOS/13, $\|\boldsymbol{t}\| = 100\,\mathrm{m}$ (d) ASTOS/13, $\|\boldsymbol{t}\| = 75\,\mathrm{m}$ (e) ASTOS/13, $\|\boldsymbol{t}\| = 50\,\mathrm{m}$

Fig. 4: Characteristics of the synthetic Astos dataset.

$\mathcal{F}_o$ [1]. Figure 4b depicts the considered rotational states for Envisat; a note is made relative to tumbling profile (TP) 3, in which the spin axis is configured at a 45 deg angle with H-bar but is simultaneously fixed in the inertial frame. Since Envisat's orbit is approximately circular, this results in the spin axis demonstrating an axial precession with period equal to the orbital period, or 3.59 deg min^{-1}. Apart from the guidance and tumbling profiles, additional variation is added via the approach vector (see Tab. II), where the V-bar case features a black, deep-space background, and the R-bar case contains Earth in the field of view (FOV). Lastly, Figures. 4c–4e display uncropped frames of one Astos dataset sequence following GP1 at different relative ranges, to illustrate the apparent size of the target with respect to the FOV.

2) Training and Testing: A train-test split is performed on the Astos dataset according to Table II, where one half of the sequences are used for training and the other half for testing.

TABLE II: Sequence key for the Astos dataset.

Sequence	GP	TP	Approach Vector	Selection	Length (s)
00	1	1	V-bar	Train	125
01	1	1	R-bar	Test	125
02	1	2	V-bar	Test	125
03	1	2	R-bar	Train	125
04	1	3	V-bar	Train	125
05	1	3	R-bar	Test	125
06	2	1	V-bar	Test	309
07	2	1	R-bar	Train	309
08	2	2	V-bar	Train	216
09	2	2	R-bar	Test	216
10	2	3	V-bar	Test	216
11	2	3	R-bar	Train	216
12	3	1	N/A	Train	200
13	3	2	N/A	Test	200

The split was performed so that the network is trained at least once on each GP and TP, but the tests include different combinations thereof.

The sequences are further partitioned for training according to randomly sampled lengths of $\{64, 128, 256, 512\}$ seconds. Clark *et al.*'s [39] method is used to train the RNN module whereby each sequence is fed to the network according to a sliding window. In the present experiments, a window length of 8 frames with a stride of 4 was utilised.

Image augmentation is performed online (i.e. during training) on the data in terms of image processing (e.g. random brightness and contrast, Gaussian blur and noise, random pixel dropout, etc.) and camera perturbations by manipulating the image according to a homography computed through a pure rotation.

Stages 1 and 2 are trained for 100 epochs with a cyclical learning rate decay of 5 cycles, whereas Stage 3 is trained for 66 epochs with early stopping and a step learning rate decay every 9 epochs. Stage 1 samples the dataset for a total of 10 000 images. The CNN and RNN modules are trained separately, but sequentially. The Adam optimiser [57] is used. The final pipeline uses a dropout probability of 0.2, and hidden and cell states zoneout factors of 0.15 for both.

The DRCNN is implemented from the ground up on MATLAB version R2019b. The pipeline is trained on a NVIDIA® Turing® V100 Tensor Core graphics processing unit (GPU) with a minibatch size of 128.

3) Evaluation: The test results are presented in terms of the position and attitude error metrics, respectively:

$$\delta\tilde{t} := \|\hat{\boldsymbol{t}} - \boldsymbol{t}\|, \tag{14}$$

$$\delta\tilde{q} := 2\arccos\left(\hat{\boldsymbol{q}}^{-1} \otimes \boldsymbol{q}\right)_4, \tag{15}$$

where $\hat{\bullet}$ denotes the estimated quantity, $\otimes$ denotes quaternion multiplication, and the subscript "4" refers to the scalar element of the quaternion. Additionally, the position error is also assessed in terms of the relative range:

$$\delta\tilde{t}_\mathrm{r} := \frac{\delta\tilde{t}}{\|\boldsymbol{t}\|}. \tag{16}$$

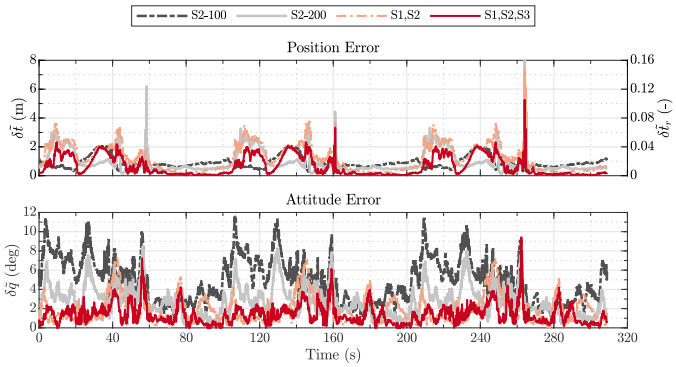
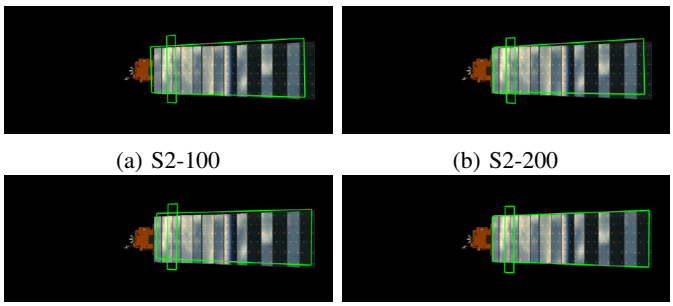


Fig. 5: Comparison of estimated position and attitude errors over time in terms of training stages used for ASTOS/06. The position error is presented in both unnormalised and normalised scales. All models are trained on a CNN taking RGB inputs. *(S2-100)* Stage 2 trained for 100 epochs. *(S2-200)* Stage 2 trained for 200 epochs. *(S1,S2)* Stage 1 and Stage 2. *(S1,S2,S3)* Stage 1, Stage 2, and Stage 3.



(a) S2-100 (b) S2-200

(c) S1,S2 (d) S1,S2,S3

Fig. 6: Effect of multistage optimisation illustrated on ASTOS/06 at time $\tau = 59.4\,\mathrm{s}$. Each stage progressively improves the pose estimate in the presence of spurious reflections, as shown by the model reprojection in green. See Fig. 5 for key.

For succinctness, the ASTOS/06 sequence is used as a representative case study, where the errors are plotted as a function of time, whereas the results for the remaining sequences are summarised for the complete pipeline in terms of their mean and median statistics. Since for this sequence the relative distance is constant (see Fig. 4), $\delta\tilde{t}_\mathrm{r}$ is illustrated by adding a second vertical axis to the position error plots.

Evaluation of Multistage Optimisation

To assess each contribution in the proposed multistage optimisation scheme, the CNN module on its own is first considered, and trained according to four different schemes: 1) Stage 2 only for 100 epochs [S2-100]; 2) Stage 2 only for 200 epochs [S2-200]; 3) Stages 1 and 2 [S1,S2]; and 4) Stages 1, 2, and 3 [S1,S2,S3]. The RNN is not considered for this test.

Figure 5 depicts the results of the benchmark on the baseline. From the overall shape of the plot lines, the periodicity of the tumbling motion can be clearly discerned. An initial period

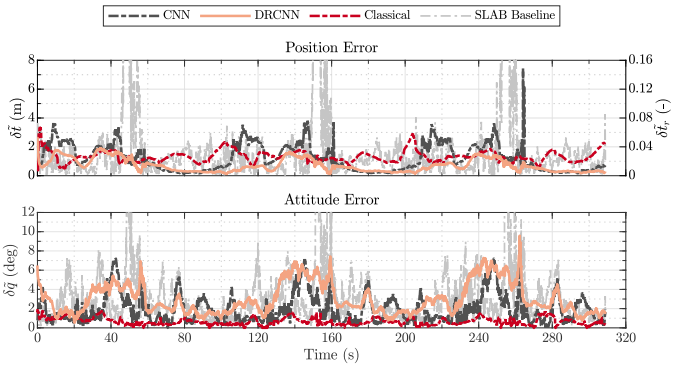


Fig. 7: Comparison of estimated position and attitude errors over time in terms of recurrence for ASTOS/06, benchmarking the plain CNN against the complete DRCNN, both trained on Stages 1 and 2. The position error is presented in both unnormalised and normalised scales. "Classical" refers to the algorithm in [29]. "SLAB Baseline" refers to the algorithm in [58]. All models are trained on RGB inputs.

approximately covering the interval $\tau \in [0\ ;\ 60[\,\mathrm{s}$ is first noted, during which the target performs slightly over half a revolution and the errors are overall higher, culminating in a local peak at which the solar array reflects Earth's rim (see Fig. 6). It is then followed by a second period covering $\tau = [60\ ;\ 103[\,\mathrm{s}$ where the main body (also known as "bus") comes back into view and both shadows and reflections are minimised, hence driving down the errors. This pattern is repeated twice more throughout the plot as the target performs a total of three revolutions.

Regarding the position error, the S1,S2 strategy is essentially on par with S2-100 and S2-200 for the first period, and performs better than both on the second period. Notably, the benefit of the dual-stage training can be observed specifically at times $\tau = \{60, 160, 260\}\,\mathrm{s}$, where a mitigation of the error spikes is seen. Training on the three stages (S1,S2,S3) reduces these peaks even further.

The gains of adopting the proposed method become clearer looking at the attitude error plot. S2-100 exhibits the higher error throughout, followed by S2-200. The dual-stage S1,S2 approach further reduces the error, except for peaks at $\{45,147,250\}$ s, where it is comparable to the previous mode; this corresponds to the segments where the target nearly completes half a revolution and the solar array begins to cover the main bus. The triple-stage approach can be seen to provide the steadiest performance. It is also noted that the highest error peaks for the attitude correspond to those identified for the position, which S1,S2,S3 mitigates, but does not completely eliminate.

Evaluation of Recurrent Module

In this section, the performance of the CNN is compared to the complete DRCNN; Figure 7 plots the estimation results over time, where the training regime consisted of S1,S2, and RGB inputs are considered. The DRCNN is successful in overwhelmingly mitigating the localised position error peaks, which correspond to points in the trajectory where the solar array reflections are most intense or it occludes the main

bus, as mentioned in the previous section. This is due to the LSTM states taking into account the preceding images, thus preventing sudden jumps in the solution. The mean position error is reduced approximately by half, bringing the mean range-normalised error to approximately 1.40 %.

The mean values for the attitude errors, however, are slightly worse for the RNN-based architecture. Overall, an increase of 0.5 deg–1 deg in the mean error and 1 deg in the median error is observed. It can be argued that this is an acceptable loss in performance given the benefit seen for the position estimation. However, the pipeline could instead be modified to output an attitude estimate from the CNN alone while processing the position with the RNN. This is left as future work.

To compare the proposed approach with an indirect formulation, Stanford's Space Rendezvous Laboratory (SLAB) CNN [58] was trained on the same Astos dataset split and benchmarked on this trajectory. As reported in [9], it was the baseline contribution of the SPEC authors, being only surpassed by three other entries. Therefore, it is considered a representative algorithm for the class of indirect approaches. It consists of two subnetworks: an object detector and a keypoint regressor. As expected, the error profiles are noisy since there are no temporal constraints applied. In particular, it oscillates more than ChiNet's CNN profile, suggesting that depending on two subnetworks introduces an additional error source. More importantly, the SLAB baseline solution "blows up" around the three periods highlighted previously (see the analysis of Fig. 5), reaching errors as high as 340 m for the position and 180 deg for the attitude. These correspond to frames where the number of visible keypoints is minimal, highlighting the disadvantages of this type of approach in such situations.

Finally, the peformance is compared against the classic ML-based algorithm developed by the authors in [29] (herein referred to as "classical"), which uses a combination of IP-based feature detection and description techniques in the visible modality, robust optimisation, and Kalman filtering. For the position benchmark, the classical solution reaches a maximum error of 2.5 m only and exhibits fewer fluctuations compared to the CNN. However, the mean error of the former ($\delta\tilde{t}_\mathrm{r} = 3.59\,\%$) is 2.14 percent points higher than the latter's. The DRCNN benchmark curve stays below both throughout virtually the whole sequence. Regarding the attitude benchmark, however, the deep learning-based solutions display lower robustness compared to the classical method, which sees a decrease in the mean error by 1.28 deg and 2.36 deg with respect to the CNN and DRCNN, respectively.

Evaluation of Multimodal Inputs

In this section, the influence of augmenting the RGB input produced by regular camera with an image in the LWIR, thus creating a four channel multimodal RGBT input, is assessed. Two models are trained for comparison, one with inputs exclusively on the visible modality, and another with multimodal inputs. Both models are trained on Stages 1 and 2. Again, the RNN is not considered for this test so as to separate the effect of each contribution. The results are depicted in Figure 8.

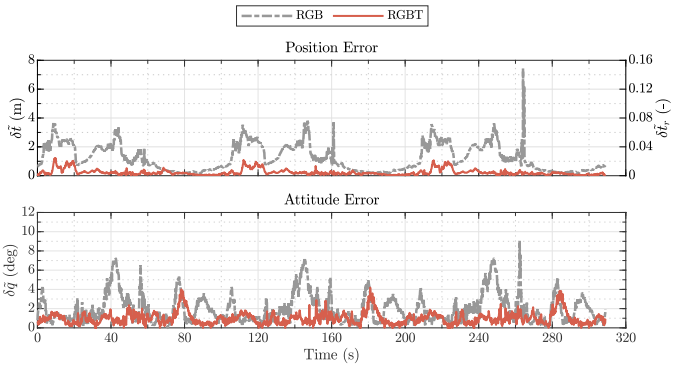


Fig. 8: Comparison of estimated position and attitude errors over time in terms of imaging modality for ASTOS/06, benchmarking RGB inputs against the multimodal RGBT. The position error is presented in both unnormalised and normalised scales. All models are trained on a CNN and Stages 1 and 2.

TABLE III: Summary of position and attitude error statistics on all Astos dataset rendezvous test sequences for the complete DRCNN pipeline, trained on Stages 1, 2, and 3. All tests use multimodal RGBT inputs. "SD" denotes standard deviation.

Sequence	$\delta\tilde{t}$ (m)			$\delta\tilde{t}_r$ (-)			$\delta\tilde{q}$ (deg)		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
01	3.45	1.75	3.51	0.0455	0.0222	0.0472	7.49	6.89	4.80
02	4.05	1.49	4.23	0.0583	0.0210	0.0616	8.67	10.46	4.53
05	3.09	1.79	3.12	0.0465	0.0329	0.0455	14.12	15.46	8.63
06	0.24	0.14	0.23	0.0048	0.0027	0.0046	1.85	0.70	1.80
09	0.33	0.39	0.24	0.0065	0.0075	0.0048	2.09	2.10	1.26
10	0.67	0.36	0.63	0.0133	0.0070	0.0125	10.61	6.92	9.02
13	0.29	0.25	0.21	0.0058	0.0049	0.0041	3.52	2.25	2.77

The contribution of the multimodality can be seen immediately from the figure, where the plots of both position and attitude errors in time exhibit more stability for RGBT inputs compared to RGB inputs. Notably, not only are the reflection-induced peaks mitigated, but the errors corresponding to the approximate first half of the tumbling period are as well. Overall, the mean position error is reduced in almost 80 % by using multimodal inputs, granting a mean range-normalised position error below 0.5 %, compared to 2.5 % for visible only. The mean attitude error is halved, becoming slightly lower than 1 deg.

Summary of Performance

Table III compiles the error statistics for the performance of the complete multimodal DRCNN framework on the entire Astos test dataset. The performance on the nominal sample sequence is plotted for comparison with the previous sections. For completeness, the results on ASTOS/02 are also graphed against time to evaluate ChiNet's behaviour on a sub-optimal case. Both plots are juxtaposed in Figure 9. Lastly, results on ASTOS/06 are illustrated qualitatively in Figure 1a (Fig. 1b showcases those on ASTOS/13).

Beginning with ASTOS/06, it can be seen that ChiNet provides an estimate of the position with an error bound at 0.6 m, scoring on average a mean $\delta\tilde{t}_\mathrm{r} = 0.49\,\%$. The classical solution [29], on the other hand, reached maximum values of 2.5 m (refer again to the red curve in Fig. 7). For this trajectory, ChiNet presents an improvement of around 2.2 percentage

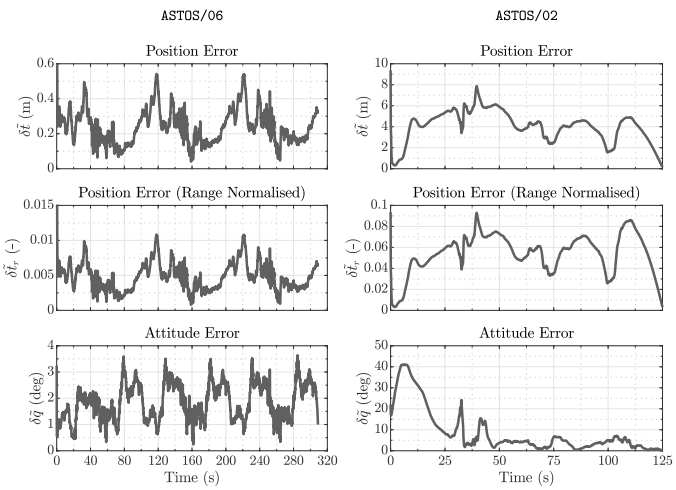


Fig. 9: Estimated position and attitude errors over time for the complete multimodal DRCNN on two trajectories. *(Left)* Good fit. *(Right)* Challenging fit.

points in terms of mean range-normalised position error. The classical solution performs better in terms of mean attitude error (0.78 deg). Still, ChiNet produces a solution not exceeding 2 deg in error.

Considering the remaining sequences within GP2 (fixed relative range), it can be seen that the quality of the solution degrades as more challenging rotation modes are considered. The estimation of the attitude appears to be more affected by this factor. For mode TP2 (two-axis rotation), the pose errors are comparable to TP1, even despite the benchmark of the former being performed on an R-bar approach vector (i.e. with Earth in the FOV). Mode TP3 (precession) experiences by far the largest degradation, with the mean attitude error exceeding 10.5 deg. On sequences featuring this rotation mode, the edge of the solar array leaves the FOV for a considerable amount of time, which could explain the higher error.

Overall, GP1 trajectories (forced translation) exhibits reduced performance when compared to GP2. For this profile, ChiNet produces estimates of the position with mean $\delta\tilde{t}_\mathrm{r}$ not exceeding 5.5 %. The mean attitude error is less affected by the change in guidance profile, being 1.5–4 × higher with respect to GP2. This was expected since the network sees far more examples of the relative pose at a distance of 50 m than at larger distances. The drop in performance can be noted from the benchmark curves of `ASTOS/02`. This sequence exhibit the highest position estimation errors in the test set, oscillating between 3–9 % in relative units throughout most of the sequence, before greatly decreasing towards the end when the target-chaser distance nears 50 m. Interestingly, the error is also lower during the first tenth of the trajectory, which could be explained by the initial attitude configuration (and resultant projection of the object on the FOV) facilitating the position estimation.

Indeed, a position-attitude coupling can be ascertained since `ASTOS/02` shares the same TP with `ASTOS/09` but a different GP, and showcases a transient attitude error reaching ∼40 deg. Nevertheless, it is noted that, once past the peak at time $\tau = \sim 6$ s, the attitude estimate begins to converge, reaching a steady state error at $\tau = \sim 45$ s which is comparable to the other good fits. Despite being a challenging fit, ChiNet's modular approach prevents the solution from fully diverging. In terms of mean values for `ASTOS/02`, ChiNet's $\delta\tilde{t}_\mathrm{r} = 5.47$ % is approximately 2.7 percent points higher than the output of the classical algorithm. The mean attitude error higher by a factor of 3.4 ×.

TABLE IV: Sequence key for the City dataset.

Sequence	GP	Initial dist. (m)	Final dist. (m)	Rotation (rev)	Length (s)
`00`	Fixed	3.8	3.8	2	120
`01`	Fixed	1.1	1.1	2	120
`02`	Translation	3.8	1.1	0.5	30
`03`	Translation	3.8	2	0.5	30

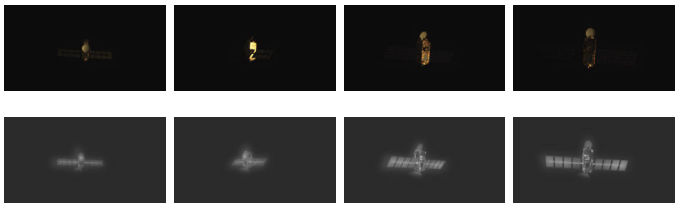


Fig. 10: Sample images from the City dataset (cropped for visualisation purposes). *(Top row)* Visible modality. *(Bottom row)* LWIR modality.

B. Experimental Dataset

1) Description: Lastly, the performance of the complete ChiNet pipeline is assessed on real data acquired from the Autonomous Systems and Machine Intelligence Laboratory (ASMIL) at City, University of London (herein referred to as "City dataset"). This test provides insight on how well the deep learning framework can adapt to data captured by actual sensors, and to the sources of error a laboratory setup brings (e.g. camera calibration; ground truth measurement; camera misalignments; camera synchronisation; sensor noise). It also evaluates how the network fares against previously unseen motion when trained on reduced amounts of data.

The City dataset consists of a multimodal collection of four rendezvous sequences with a 1:4 scale mock-up of the Jason-1 satellite. The mock-up rotates along its vertical axis at a constant rate of 6 deg s^{-1}. Despite having a different form factor, Jason-1 is similar to Envisat in terms of components (i.e. main bus coated in multi-layer insulation [MLI], thermal radiators, solar array, radiometric instruments). In total, four trajectory types are considered. Table IV summarises the characteristics of each sequence, and Figure 10 shows some sample images from the dataset in each modality.

Trajectories are acquired for simulation of both sunlight and eclipse conditions. On the visible spectrum, this is controlled respectively by aiming a floodlight directly at the target, or by aiming it at a nearby wall, creating a dimly lit environment. On the LWIR spectrum, the model's temperature is controlled by internal resistor heaters in the main bus and by an external heater. The thermal signature of the model is made to coarsely match that of Envisat in both illumination conditions. Images

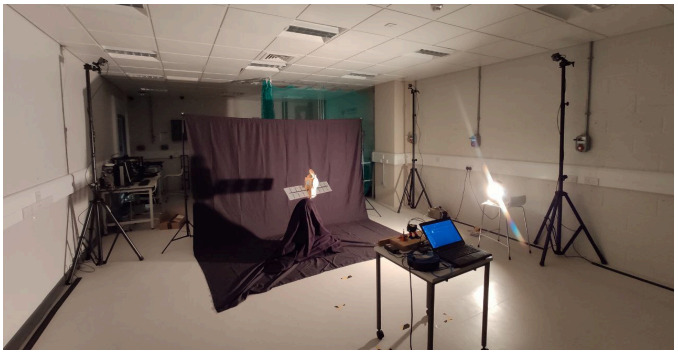


Fig. 11: Validation setup of the ASMIL at City, University of London.

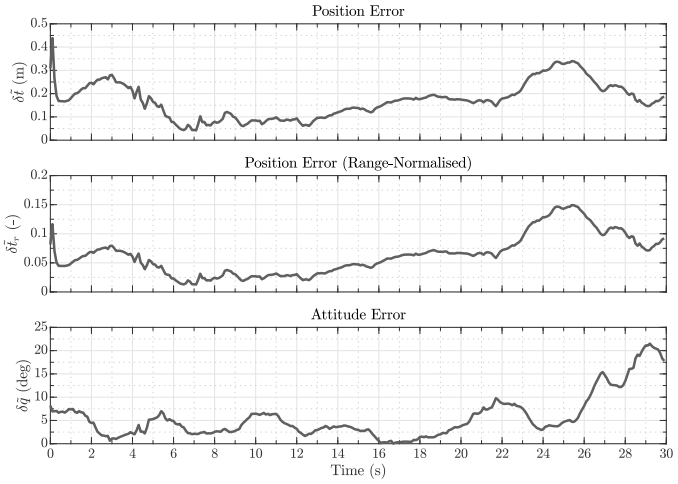


Fig. 12: Estimated position and attitude errors over time on the CITY/03 laboratory test rendezvous sequences. The model is trained on the full DRCNN pipeline with multimodal RGBT inputs and on Stages 1, 2, and 3.

are acquired at a resolution of 744 px × 490 px and frequency of 10 Hz (software synchronised). The visible and thermal cameras are aligned and set up in a stereo configuration with a very short baseline to minimise disparity. The ground truth is recorded with a six-camera OptiTrack motion capture system. Using the ground truth and the CAD model of the target, the background is digitally masked out to simulate a deep space background. Figure 10 depicts some sample frames of the dataset, whereas Figure 11 showcases the experiment setup at ASMIL.

2) Training and Testing: The methodology follows analogously from Section IV-A. The pipeline is trained on CITY/00, CITY/01, and CITY/02, and is evaluated on CITY/03.

Figure 12 illustrates the evolution in time of the position and attitude estimation errors for the test sequence CITY/03. Figure 13 qualitatively illustrates these results. It can be observed that the position error is bounded at 35 cm throughout the trajectory, except for the initial transient period. The mean and median error are shown to be approximately half of that, which corresponds to a figure below 6.5 % of range. The attitude error is kept below 10 deg for the first 85 % of the

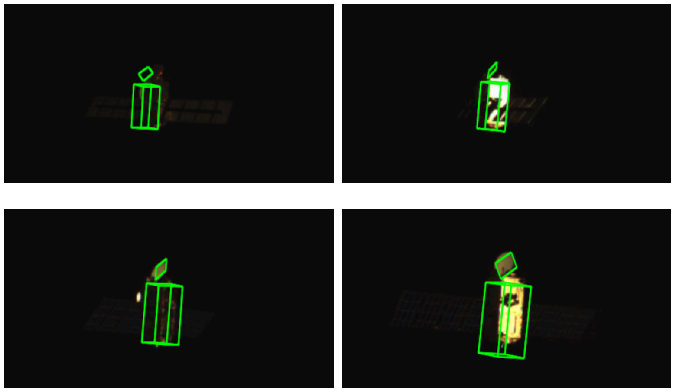


Fig. 13: Qualitative pose estimation performance on frames of the CITY/03 laboratory test rendezvous sequences.

sequence, demonstrating that the network is mostly able to separate the translational motion from the rotational one; a degradation of the estimate is observed during the last 4 s, when the target reaches a rotation of 180 deg around the spin axis and the error peaks at about 20 deg, which can be explained by the fact that the training data is biased towards an observation of that specific attitude for larger relative distances. The mean error is approximately 5.5 deg (resp. 3.97 deg median).

V. Conclusion

This paper presented ChiNet: a contribution towards deep learning-based, end-to-end, multimodal spacecraft pose estimation for orbital NCRV. The proposed method employs a CNN as a front-end feature extractor and applies an LSTM-based RNN back-end to model the temporal relationship between incoming frames from an optical camera. Furthermore, RGB images are augmented with those captured in the LWIR band, granting a feature-rich input beyond the visible. The full pipeline is trained according to an innovative multistage optimisation scheme that categorises the learning process in a coarse to fine fashion.

Each of the proposed contributions was individually tested on realistic synthetic data. The addition of the coarse training stage was demonstrated to mitigate spikes in the pose estimation errors originating from sharp reflections of both Earth and sunlight on the solar array. Including the keypoint-based refinement stage improved the average position and attitude errors. The recurrent module eliminated sharp jumps in the estimate of the position, reducing the mean error by half. The inclusion of multimodal RGBT image inputs was shown to improve the mean position error in nearly 80 % and to reduce the mean attitude error in half.

Overall, ChiNet was shown to generalise well to unseen trajectories, benchmarking a mean range-normalised position error of 2.5 % per average trajectory and a mean attitude estimation error of 6.9 deg per average trajectory on the sequences of the Astos dataset. The simplest case was shown to be comparable to the classical solution developed in [29], even surpassing it in terms of position estimation performance. In contrast, the solution generated from an indirect-based network was shown to blow up under the same conditions. Challenging cases were identified for which sub-optimal fits were attained,

showing room for improvement, but the proposed methodology was demonstrated to avoid divergence nonetheless. The pipeline required no localisation or segmentation preprocessing to produce an accurate solution. Lastly, the proposed work was benchmarked on experimental data, demonstrating the capability of the network to learn novel situations under a reduced training regime.

Future work might investigate the robustness of the framework towards non-nominal illumination conditions. Another potential avenue to investigate could be tackling the problem of bridging the domain gap in the context of spacecraft pose estimation via adaptation or randomisation, whereby a deep network is trained with synthetic images and tested on real data, as the latter are typically scarce prior to the actual mission, but the former can be generated in large quantities.

REFERENCES

[1] Fehse, W., *Automated Rendezvous and Docking of Spacecraft*, 1st ed. Cambridge, UK: Cambridge University Press, 2003, pp. 1, 3, 8, 32–33, 114, 272–277.

[2] Wertz, J. R. and Bell, R., "Autonomous Rendezvous and Docking Technologies — Status and Prospects," in *Space Systems Technology and Operations*, Peter Tchoryk, J. and Shoemaker, J., Eds., vol. 5088, Orlando, FL: SPIE, Aug. 2003.

[3] Cassinis, L. P., Fonod, R., and Gill, E., "Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft," *Progress in Aerospace Sciences*, vol. 110, p. 100 548, Oct. 2019.

[4] Szeliski, R., *Computer Vision: Algorithms and Applications*, 1st ed. London, UK: Springer-Verlag, 2011, pp. 44–49, 284–286.

[5] Fischler, M. A. and Bolles, R. C., "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[6] Mouats, T., Aouf, N., and Richardson, M. A., "A Novel Image Representation via Local Frequency Analysis for Illumination Invariant Stereo Matching," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2685–2700, 2015.

[7] Beauvisage, A., Ahiska, K., and Aouf, N., "Multimodal tracking framework for visual odometry in challenging illumination conditions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 11 133–11 139.

[8] LeCun, Y., Jackel, L., Boser, B., *et al.*, "Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 41–46, Nov. 1989.

[9] Kisantal, M., Sharma, S., Park, T. H., *et al.*, "Satellite Pose Estimation Challenge: Dataset, Competition Design and Results," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–1, 2020.

[10] Stewart, C. V., "Robust parameter estimation in computer vision," *SIAM Review*, vol. 41, no. 3, pp. 513–537, Jan. 1999.

[11] Drummond, T. and Cipolla, R., "Real-time visual tracking of complex structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 932–946, 2002.

[12] Kelsey, J., Byrne, J., Cosgrove, M., Seereeram, S., and Mehra, R., "Vision-based relative pose estimation for autonomous rendezvous and docking," in *2006 IEEE Aerospace Conference*, IEEE, 2006.

[13] Petit, A., Marchand, E., and Kanani, K., "A robust model-based tracker combining geometrical and color edge information," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Nov. 2013.

[14] ——, "Combining complementary edge, keypoint and color features in model-based tracking for highly dynamic scenes," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, May 2014.

[15] Cropp, A., "Pose Estimation and Relative Orbit Determination of a Nearby Target Microsatellite using Passive Imagery," Ph.D. dissertation, University of Surrey, United Kingdom, 2001.

[16] D'Amico, S., Benn, M., and Jørgensen, J. L., "Pose estimation of an uncooperative spacecraft from actual space imagery," *International Journal of Space Science and Engineering*, vol. 2, no. 2, p. 171, 2014.

[17] Shi, J.-F., Ulrich, S., and Ruel, S., "Spacecraft Pose Estimation Using a Monocular Camera," in *67th International Astronautical Congress*, Paper IAC–16–C1.3.4, Guadalajara, Mexico: International Astronautical Federation (IAF), 2016.

[18] Rondao, D. and Aouf, N., "Multi-view monocular pose estimation for spacecraft relative navigation," in *2018 AIAA Guidance, Navigation, and Control Conference*, Kissimmee, FL: American Institute of Aeronautics and Astronautics, Jan. 2018.

[19] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[20] Bay, H., Tuytelaars, T., and Van Gool, L., "SURF: Speeded Up Robust Features," in *European Conference on Computer Vision – ECCV 2006*, Springer Berlin Heidelberg, 2006, pp. 404–417.

[21] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., "ORB: An Efficient Alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 2564–2571.

[22] Shi, J.-F., Ulrich, S., Ruel, S., and Anctil, M., "Uncooperative Spacecraft Pose Estimation Using an Infrared Camera During Proximity Operations," in *AIAA SPACE 2015 Conference and Exposition*, Paper AIAA 2015-4429, Pasadena, CA: American Institute of Aeronautics and Astronautics, Aug. 2015.

[23] Gansmann, M., Mongrard, O., and Ankersen, F., "3D Model-Based Relative Pose Estimation for Rendezvous and Docking Using Edge Features," in *10th International ESA Conference on Guidance, Navigation and Control Systems*, Salzburg, Austria: ESA, 2017.

[24] Yılmaz, Ö., Aouf, N., Majewski, L., Sanchez-Gestido, M., and Ortega, G., "Using Infrared Based Relative Navigation for Active Debris Removal," in *10th International ESA Conference on Guidance, Navigation and Control Systems*, Salzburg, Austria: ESA, 2017, pp. 1–16.

[25] Rondao, D., Aouf, N., Richardson, M. A., and Dubois-Matra, O., "Benchmarking of local feature detectors and descriptors for multispectral relative navigation in space," *Acta Astronautica*, vol. 172, pp. 100–122, Jul. 2020.

[26] Comellini, A., Le Ny, J., Zenou, E., Espinosa, C., and Dubanchet, V., "Global descriptors for visual pose estimation of a non-cooperative target in space rendezvous," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–1, 2021.

[27] Petit, A., Marchand, E., Sekkal, R., and Kanani, K., "3d object pose detection using foreground/background segmentation," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, May 2015.

[28] Shi, J.-F., Ulrich, S., and Ruel, S., "Spacecraft Pose Estimation using Principal Component Analysis and a Monocular Camera," in *AIAA Guidance, Navigation, and Control Conference*, Paper 2017-1034, American Institute of Aeronautics and Astronautics, Jan. 2017.

[29] Rondao, D., Aouf, N., Richardson, M. A., and Dubanchet, V., "Robust On-Manifold Optimization for Uncooperative Space Relative Navigation with a Single Camera," *Journal of Guidance, Control, and Dynamics*, vol. 44, no. 6, pp. 1157–1182, Jun. 2021.

[30] Sharma, S., Beierle, C., and D'Amico, S., "Pose Estimation for Non-Cooperative Spacecraft Rendezvous Using Convolutional Neural Networks," in *2018 IEEE Aerospace Conference*, IEEE, Mar. 2018.

[31] Proença, P. F. and Gao, Y., *Deep Learning for Spacecraft Pose Estimation from Photorealistic Rendering*, 2019. arXiv: 1907.04298 [cs.CV].

[32] Sharma, S. and D'Amico, S., "Neural network-based pose estimation for noncooperative spacecraft rendezvous," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 6, pp. 4638–4658, 2020.

[33] Chen, B., Cao, J., Parra, A., and Chin, T.-J., "Satellite Pose Estimation with Deep Landmark Regression and Nonlinear Pose Refinement," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Oct. 2019.

[34] Pasqualetto Cassinis, L., Fonod, R., Gill, E., Ahrns, I., and Gil-Fernández, J., "Evaluation of tightly- and loosely-coupled approaches in cnn-based pose estimation systems for uncooperative spacecraft," *Acta Astronautica*, vol. 182, pp. 189–202, 2021, ISSN: 0094-5765.

[35] Huo, Y., Li, Z., and Zhang, F., "Fast and accurate spacecraft pose estimation from single shot space imagery using box reliability and keypoints existence judgments," *IEEE Access*, vol. 8, pp. 216 283–216 297, 2020.

[36] Piazza, M., Maestrini, M., and Di Lizia, P., "Deep Learning-Based Monocular Relative Pose Estimation of Uncooperative Spacecraft," in *8th European Conference on Space Debris*, Flohrer, T., Lemmens, S., and Schmitz, F., Eds., ESA Space Debris Office, vol. 8, Darmstadt, Germany, May 2021.

[37] Garcia, A., Musallam, M. A., Gaudilliere, V., *et al.*, "LSPnet: A 2D Localization-oriented Spacecraft Pose Estimation Neural Network," in *AI4Space 2021 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Jun. 2021. arXiv: 2104.09248 [cs.CV].

[38] Hogan, M., Rondao, D., Aouf, N., and Dubois-Matra, O., *Using Convolutional Neural Networks for Relative Pose Estimation of a Non-Cooperative Spacecraft with Thermal Infrared Imagery*, 2021. arXiv: 2105.13789 [cs.CV].

[39] Clark, R., Wang, S., Wen, H., Markham, A., and Trigoni, N., *VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem*, 2017. arXiv: 1701.08376 [cs.CV].

[40] Wang, S., Clark, R., Wen, H., and Trigoni, N., "DeepVO: Towards End-to-end Visual Odometry with Deep Recurrent Convolutional Neural Networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, May 2017.

[41] Hochreiter, S. and Schmidhuber, J., "Long Short-term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] Kechagias-Stamatis, O., Aouf, N., Dubanchet, V., and Richardson, M., "Deeplo: Multi-projection deep lidar odometry for space orbital robotics rendezvous relative navigation," *Acta Astronautica*, vol. 177, pp. 270–285, 2020, ISSN: 0094-5765.

[43] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016.

[44] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016.

[45] Redmon, J. and Farhadi, A., "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017.

[46] ——, *YOLOv3: An Incremental Improvement*, 2018. arXiv: 1804.02767 [cs.CV].

[47] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R., *Improving neural networks by preventing co-adaptation of feature detectors*, 2012. arXiv: 1207.0580 [cs.NE].

[48] Poujol, J., Aguilera, C. A., Danos, E., *et al.*, "A Visible-Thermal Fusion Based Monocular Visual Odometry," in *Advances in Intelligent Systems and Computing*, Springer International Publishing, Dec. 2015, pp. 517–528.

[49] Ba, J. L., Kiros, J. R., and Hinton, G. E., *Layer Normalization*, 2016. arXiv: 1607.06450 [stat.ML].

[50] Krueger, D., Maharaj, T., Kramár, J., *et al.*, *Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations*, 2017. arXiv: 1606.01305 [cs.NE].

[51] Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*. MIT Press, 2016, pp. 12–14, 78, 185–191, 286–291, 298–302, 526–531.

[52] Deng, J., Dong, W., Socher, R., *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[53] Kendall, A., Gal, Y., and Cipolla, R., "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[54] Kendall, A., Grimes, M., and Cipolla, R., "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2938–2946.

[55] Zhou, Y., Barnes, C., Lu, J., Yang, J., and Li, H., *On the continuity of rotation representations in neural networks*, 2020. arXiv: 1812.07035 [cs.LG].

[56] Kendall, A. and Cipolla, R., "Geometric Loss Functions for Camera Pose Regression with Deep Learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6555–6564.

[57] Kingma, D. P. and Ba, J., *Adam: A Method for Stochastic Optimization*, 2014. arXiv: 1412.6980 [cs.LG].

[58] Park, T. H., Sharma, S., and D'Amico, S., *Towards Robust Learning-Based Pose Estimation of Noncooperative Spacecraft*, 2019. arXiv: 1909.00392 [cs.CV].

Duarte Rondao is a postdoctoral research fellow in computer vision for space rendezvous in the Robotics and Machine Intelligence group at City, University of London. He obtained his PhD in 2021 from Cranfield University, and has 7 years of experience in the space sector, having worked on two different satellite missions: the European Student Earth Orbiter microsatellite (successfully launched in December 2018); and the ECOSat-III nanosatellite of the Centre for Aerospace Research at the University of Victoria, Canada, the successor to the group's previous Canadian Satellite Design Challenge winning design.

Prof Nabil Aouf received his PhD from McGill University in 2002 at the Electrical and Computer Engineering Department. Currently, he is Professor of Autonomous Systems and Machine Intelligence at City University of London. He is the Director of the Systems, Autonomy and Control (SAC) Centre and the co-Director of the London Space Institute (LSI) at City University of London. He also leads the Robotics, Autonomy and Machine Intelligence (RAMI) group and works very closely with industries that have a strong heritage in autonomous systems and space research. He has authored over 180 high calibre publications in his domains of interest. His research interests are aerospace and defence systems, information fusion and vision systems, guidance and navigation, control, and autonomy of systems. He is an Associate Editor of 4 journals including IEEE Transactions of Intelligent Vehicles.

Prof Mark A. Richardson has a BSc with First Class Honours in Physics from Imperial College London and is an Associate of the Royal College of Science. He has an MSc with Distinction in Applied Optics from Imperial College London, a Diploma of Imperial College, and PhD in Infrared Physics from Cranfield University. He is currently the Pro-Vice-Chancellor of Cranfield Defence and Security. He has written over 300 papers on Infrared Signature Simulation & Modelling and EO&IR Countermeasures. He is the editor and principal author of a book on battlefield surveillance technology and has frequently acted as a consultant and defence analyst to both the UK Ministry of Defence and commercial industry.