LSPnet: A 2D Localization-oriented Spacecraft Pose Estimation Neural Network

Albert Garcia[1] Mohamed Adel Musallam[1] Vincent Gaudilliere[1] Enjie Ghorbel[1]
Kassem Al Ismaeil[1] Marcos Perez[2] Djamila Aouada[1]

[1] Interdisciplinary Center for Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg
{albert.garcia, mohamed.ali, vincent.gaudilliere, enjie.ghorbel
kassem.alismaeil, djamila.aouada}@uni.lu

[2] LMO
m.perez@lmo.space

Abstract

Being capable of estimating the pose of uncooperative objects in space has been proposed as a key asset for enabling safe close-proximity operations such as space rendezvous, in-orbit servicing and active debris removal. Usual approaches for pose estimation involve classical computer vision-based solutions or the application of Deep Learning (DL) techniques. This work explores a novel DL-based methodology, using Convolutional Neural Networks (CNNs), for estimating the pose of uncooperative spacecrafts. Contrary to other approaches, the proposed CNN directly regresses poses without needing any prior 3D information. Moreover, bounding boxes of the spacecraft in the image are predicted in a simple, yet efficient manner. The performed experiments show how this work competes with the state-of-the-art in uncooperative spacecraft pose estimation, including works which require 3D information as well as works which predict bounding boxes through sophisticated CNNs.

1. Introduction

In recent years, more and more space mission scenarios have involved close-proximity operations with uncooperative space objects such as space debris (*e.g.* active debris removal), out-of-order satellites (in-orbit servicing) or even comets and asteroids (space exploration).

In these scenarios, a chaser spacecraft seeks to approach then to capture or to dock at a target orbiting space object, that is uncooperative. In other words, the latter is not communicating any information to the former, either actively (*e.g.* by radio communication) or passively (*e.g.* by featuring a fiducial marker) [14]. In this context, most space relative navigation methods first require to estimate the relative position and attitude, referred to as pose, between both spacecrafts, then to track the relative pose over time using previous estimations [3]. To enable autonomous close-proximity operations in space with uncooperative targets, robust and efficient on-board pose initialization solutions are required. To this end, several vision-based works propose to use active sensors such as Light Detection and Ranging (LIDAR) [13, 10]. Despite their demonstrated efficiency, such sensors remain heavy and high-power consuming. On the contrary, relying on a single monocular camera has the advantage of complying with strict power and mass

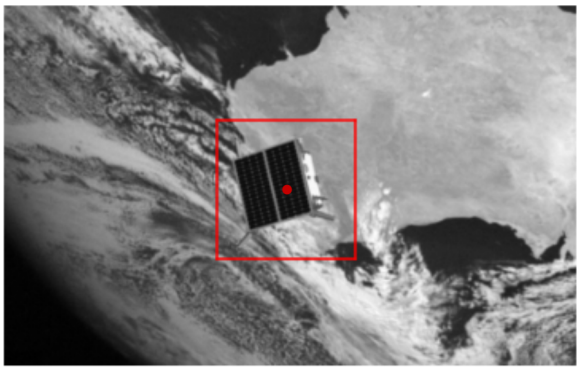

Figure 1. Example of a spacecraft bounding box detection by LSPnet. The bounding box is constructed based on the central red dot yielded by LSPnet.

This work was funded by the Luxembourg National Research Fund (FNR), under the project reference BRIDGES2020/IS/14755859/MEET-A/Aouada, and by LMO (https://www.lmo.space).

Article to be published at AI4Space 2021 IEEE Conference on Computer Vision and Pattern Recognition Workshops

requirements relating to space missions, while ensuring a low level of system complexity [3]. However, computing the pose of a known uncooperative object from a single monocular camera is a challenging task. First of all, the target object in the field of view of the camera can be depicted in a wide range of different scales, depending on the target's size and on the distance to the chaser spacecraft. Secondly, in a large number of cases, it is necessary to deal with cluttered backgrounds introduced by the Earth which can heavily complicate the task of pose estimation. Finally, due to the nature of the input data, a target detection step over the captured image is desired to reliably perform orientation estimation as in the latest Satellite Pose Estimation Challenge (SPEC) [8]. The downside of including a detection step is the increase in the solution complexity as well as the decrease of its computational efficiency.

The solution proposed herein, named *2D Localization-oriented Spacecraft Pose Estimation Neural Network* (LSPnet), deals with the aforementioned challenges while remaining simple and efficient. Our work takes advantage of the Deep Learning (DL)-based advances in Computer Vision by implementing a Convolutional Neural Network (CNN). In contrast to the common approach of implementing an object detection network (such as YOLO [18] or Faster-RCNN [19]) for the detection step, our approach is capable of yielding a simple bounding box in a straightforward manner. Additionally, a 2D-localization process is developed in order to aid the part of the network responsible for 3D position estimation. In other words, LSPnet learns how to estimate the spacecraft position while being driven by an auxiliary network which detects the center of the spacecraft in the image. Finally, combining the predicted position into the center detection network, a region of interest (ROI) crop (see Figure 1) is performed and used as input for orientation estimation, thus yielding the full pose of the uncooperative target spacecraft. Figure 2 presents a high-level overview of LSPnet as well as the connections between its modules.

The remainder of the paper is organized as follows: Section 2 reviews the related literature on spacecraft pose estimation. Section 3 formulates the problem and the proposed approach. Section 4 describes and discusses the experimental evaluation. Lastly, Section 5 concludes the paper.

2. Related work

Spacecraft pose estimation from a single monocular camera has extensively drawn techniques from Computer Vision literature. Thus, we provide an overview of the different approaches that appear in space applications.

2.1. Model-based approaches

Many pose estimation methods rely on a 3D model of the target spacecraft. One of the most proposed approaches

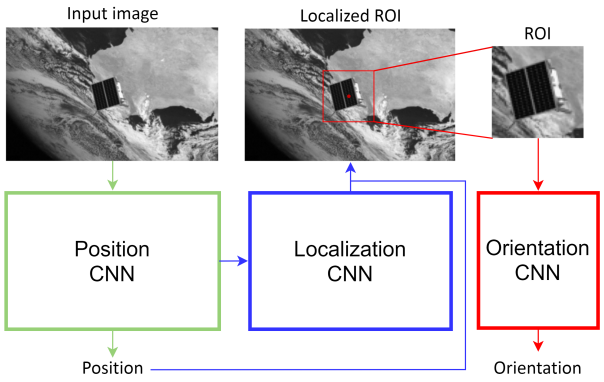

Figure 2. High-level architecture overview of the proposed LSPnet. Given an input image with a target spacecraft, LSPnet yields its position and orientation as well as a bounding box. Highlighted in green are the parts of LSPnet aimed towards position estimation, in red the ones aimed towards orientation estimation while in blue are the parts which support both tasks.

consists in matching the 2D input image with a 3D wireframe of the spacecraft. To do so, visual features are extracted from the image then matched to their corresponding elements in the wireframe. In these cases, the final pose is obtained by solving the Perspective-n-Point (PnP) problem. The features used to solve PnP vary between interest points (keypoints), corners, edges and depth maps. Classical works use handcrafted Computer Vision filters to detect these features such as Canny and Sobel filters [11, 5]. State-of-the-art in handcrafted Computer Vision filtering extracts a range of different features which are afterwards fused into a dense feature representation [2].

In addition to keypoint-based solutions which aim to predict the pose by performing 2D-3D matching, there are other approaches that also rely on the spacecraft 3D model. Another common model-based approach is to minimize the *projection* error defined as the misalignment between the spacecraft in the image and the projected 3D model by the predicted pose. These approaches require an initial pose which is afterwards refined through a *projection* error minimization process. For instance, the work of [6] first initializes a rough pose by means of feature matching and then refines it through a multi-dimensional Newton-Raphson algorithm used to minimize a projection error. Similarly, the work of [21] fuses a weak gradient elimination technique to detect finer features and estimates the pose based on a Newton-Raphson projection minimization fashion.

2.2. Appearance-based approaches

In comparison to feature-based methods, some approaches rely on directly exploiting the *appearance* of the spacecraft in the image. To the best of our knowledge, the only appearance-based method using a monoc-

ular camera for spacecraft pose estimation is the work of [22]. This method performs Principal Component Analysis (PCA) over the spacecraft present in a query image in order to match it to a dataset of stored images with their corresponding pose ground truths. By performing PCA, they drastically reduce the dimensions of the dataset. Despite this, the proposed method still requires to compare the query image to each entry of the stored dataset, thus making it not scalable [3].

2.3. Deep Learning-based approaches

In recent years, there has been a clear trend to rely on DL techniques in order to perform spacecraft pose estimation. The latest SPEC challenge [8] informs of a clear dominance in DL-based solutions among the participant teams. Following this trend, several works aim to directly regress the pose of the spacecraft through CNNs such as the Spacecraft Pose Network (SPN) presented in [20], the network proposed in [17] or the off-the-shelf GoogLeNet CNN [25] implemented in [16]. The recent work of [24] implements a double VGG architecture [23] to directly regress translation and rotation over a synthetic dataset as well as over a laboratory-acquired dataset simulating an on-orbit assembly operation. Deep Learning techniques offer great robustness against different lighting scenarios as well as robustness against cluttered backgrounds [9]. Other works combine Deep Learning with classical approaches, *e.g.* Deep Learning keypoints regression combined with PnP solving. The works of [4, 15, 1] all perform a first step of zooming in into a ROI yielded by an Object detection neural network. Afterwards, [4, 1] regress a set of manually selected keypoints while [15] regresses the corners of the target spacecraft in an ordered manner to avoid additional matching computations. Keypoint-based pose estimation solutions are generally robust and accurate provided that high quality 2D-3D correspondences can be obtained beforehand. Variations in lighting conditions as well as occluded keypoints can heavily impact pose accuracy. Fortunately, Deep Learning-based techniques have proven to efficiently handle these scenarios thanks to their generalization capabilities [9].

3. Proposed approach

Formally, the problem statement of this work is the prediction of the object's pose, *i.e.* the pose of *O*, relative to the camera frame *C*. In other words, the goal of the presented scenario is to predict the origin of the object's reference frame as well as its axes with respect to the camera's reference frame. This goal is achieved by estimating both the translation vector $t = (x, y, z)$ and the rotation matrix R which transforms the reference frame of *C* into the reference frame of *O*. Both t and R are expressed in the camera basis meaning that the z coordinate of the translation vector expresses the distance to the object *O*. The rotation R, which is expressed as a 3×3 orthogonal real matrix, can also be represented by a quaternion $q = (q_1, q_2, q_3, q_4)$ with unit norm. In doing so, several advantages appear such as eliminating the gimbal lock problem as well as encoding the same rotation using only 4 values instead of 9. The quaternion q encodes the same rotation through a closed-form mathematical formulation of a rotation axis and the angle to apply around the rotation axis. Given an input image I which depicts a target spacecraft obtained through a monocular visual sensor, composed by a single channel (gray) or by three channels (red, green and blue), and given the here designed Deep Learning network LSPnet with optimized weights w, then $LSPnet(I, w) = (t, q)$.

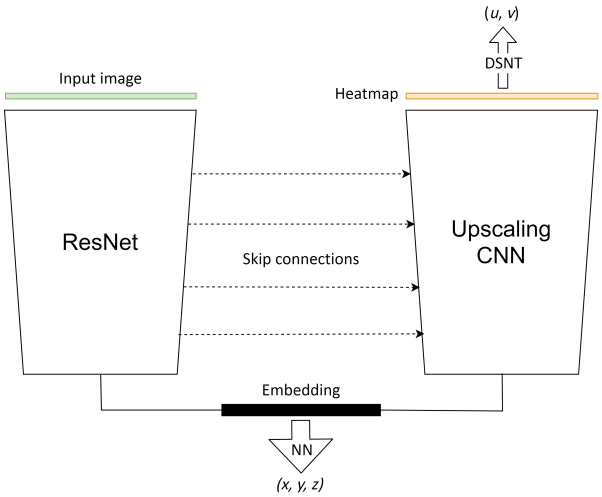

Figure 3. Architecture overview of the Translation module for the prediction of the 3-dimensional translation vector, $t = (x, y, z)$, as well as for the prediction of the pixel coordinates center of O, (u, v).

LSPnet is formed by three interconnected CNNs (named Position, Localization and Orientation) as depicted in Figure 2. When grouping together the Position and Localization CNNs they can be referred to as the Translation module due to their collaborative behavior for optimizing the predicted translation t. Once the Orientation CNN is connected to the Translation module, a full Pose estimation module is formed. The following sections cover the Translation module, the complete Pose module, a connection variation between the two modules and a specifically designed data augmentation technique.

3.1. Translation module

Based on the successful CNN architecture named Unet [26], popularly used for image segmentation, as well as based on the scalable ResNet [7] architecture, famous for its breakthrough success in image recognition, we combine both methodologies into a unified CNN which composes

the Translation module. We cast the goal of the Unet architecture, *i.e.* image segmentation, into pixel coordinates regression through the inclusion of the transformation titled *Differentiable spatial to numerical transform* (DSNT) [12]. DSNT is designed to be a non-trainable Neural Network layer for transforming heatmaps into 2-dimensional coordinates in a differentiable manner, thus avoiding a direct decoupling of the predicted heatmaps from the loss function to be optimized. Combining the 2D spatial information conservation capability of Unet, the scalability of ResNet and the relevance in coordinates regression of DSNT, the final Translation module is formed. Figure 3 offers an overview of the here described architecture for regressing the translation $t = (x, y, z)$ as well as the pixel coordinates (u, v) which localizes the position of the object O in the image. For practical reasons, the ground truth (u, v) values can be derived from projecting the ground truth coordinates x and y from the translation vector into the image plane. In this case the pixel coordinates (u, v) represent the center of the object's reference frame in the image pixel space.

Given the input image I, ResNet extracts an embedding which encodes 2D spatial information related to the target spacecraft. The extracted embedding is transferred through two different paths. First of all, a Neural Network (NN) formed by Fully Connected and ReLu layers is responsible for predicting the translation vector t. The second path connects to an upscaling CNN responsible for transforming the embedding back to the original size of the input I. This CNN makes use of upscaling layers followed by convolutional layers to increase the size of the embedding until reaching the original size. Additionally, following the Unet architecture, intermediate tensors from ResNet are concatenated into intermediate layers of the upscaling CNN. These connections are also known as skip connections. It is surmised that, through the use of these skip connections, different scales of the target spacecraft depicted in the image can be efficiently handled. The reason for this comes from the fact that the skip connections happen at different layers of ResNet, and thus the visual receptive field at each skip connection is different (from small to big scale features being captured). Once the embedding has been upscaled to the original size, it is convoluted to only present one channel and, afterwards, it is normalized. This normalized one-channel image corresponds to a heatmap representing the probabilities in the image for the pixel coordinates (u, v) of the spacecraft's center. DSNT takes as input this normalized heatmap and regresses (u, v). This same DSNT layer is the responsible for inducing the upscaling CNN towards predicting meaningful heatmaps. The upscaling CNN, which focuses on localizing the center of the spacecraft in the image pixel space, induces a localization-oriented training to ResNet when optimizing the embedding for translation estimation.

3.2. Pose module

The Pose module is built on top of the Translation module. Given the architecture as well as the outputs of the Translation module, the Pose module requires several components extracted from the former in order to estimate the quaternion q. First of all, the predicted pixel coordinates (u, v) are taken and further processed in order to find a ROI and zoom over it. This ROI crop, yielded by a straightforward bounding box technique, is used for the estimation of the orientation through a ResNet CNN. The technique used for predicting the bounding box is as follows:

1. The center of the bounding box in the image I corresponds to the predicted pixel coordinates (u, v).

2. The bounding box always takes the shape of a square which should contain the spacecraft (on its entirety if possible).

3. Knowing that the bounding box is a square, the only unknown variable left to estimate is its side length. Taking the predicted distance from the camera to the object, encoded in the z component of the translation vector t in $meters$ unit, a scaling transformation is applied as follows

$$BBL = \frac{K_O}{z} \tag{1}$$

where BBL is the length of the square bounding box in $pixels$ and K_O is a constant parameter in $pixels \times meters$ unit which depends on the spacecraft being processed in the image.

The hyperparameter K_O needs to be fine-tuned depending on each spacecraft object O. This parameter encodes the size of the spacecraft in relation to the camera parameters. The predicted bounding box lets LSPnet zoom in into a ROI which has a significantly higher signal to noise ratio. Finally, the cropped ROI is then rescaled to a fixed size in order to process it through a ResNet CNN responsible for the regression of the orientation in quaternion form. Note how the proposed bounding box technique presents the advantage of not requiring any object detection ground truth labels. It is worth mentioning that after ResNet yields a 4-dimensional vector (in combination with Fully Connected and ReLu layers), it is normalized to impose the predicted quaternion to have unit norm. In addition to the here described methodology for estimating the orientation, a neural network data-flow connection can be added to the Orientation CNN in hopes of improving its accuracy.

3.2.1 Heatmap Concatenation

Prior to performing the ROI crop, *i.e.* when taking the predicted bounding box and zooming on it over the input im-

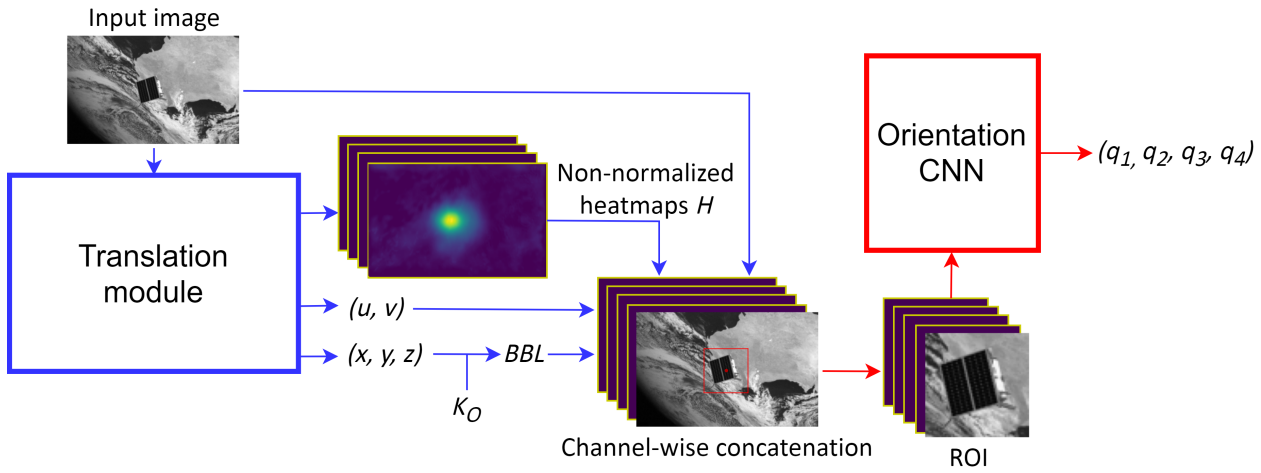

Figure 4. Diagram of the full Pose module pipeline which includes HC. For architectural details of the Translation module refer to Section 3.1 and to Figure 3.

age I, the H non-normalized heatmaps yielded by the Localization CNN (which are afterwards combined into a single normalized heatmap) can be concatenated channel-wise into the image I. For better clarification through the rest of the paper, we label this process as Heatmap Concatenation (HC). This technique offers the Orientation CNN an input tensor with $H + C_I$ channels where C_I is the number of channels of I. By implementing HC we are creating a differentiable data-flow from the Orientation CNN into the whole Translation module. This in turn means that LSPnet can be entirely optimized at the same time in a fully-differentiable manner. An ablation study presented in Section 4.1. offers insights on the impact of including HC on LSPnet. Figure 4 shows a diagram of the connections between the Translation module and the Orientation CNN in order to compose the final Pose module pipeline which includes HC.

3.2.2 Center Data Augmentation

Due to the nature of the bounding box methodology, a specialized data augmentation technique, which we name Center Data Augmentation (CDA), can be implemented in order to significantly increase the data variance offered to the Orientation CNN. The implemented data augmentation technique proposes a new bounding box based on the predicted one as follows

$$u_{aug} = u + \mathcal{N}(0,\ BBL * r) \tag{2}$$

$$v_{aug} = v + \mathcal{N}(0,\ BBL * r) \tag{3}$$

$$BBL_{aug} = BBL \tag{4}$$

where (u_{aug}, v_{aug}) is the center of the augmented bounding box, BBL_{aug} is the augmented bounding box length and r is a fixed hyperparameter such that $r \in \mathbb{R}$ and $r > 0$. The hyperparameter r encodes the dispersion added to the center coordinates in relation to the length of the bounding box. To ensure that the augmented bounding box offers a high signal to noise ratio then small values of r should be selected. An example of the results of CDA is depicted in Figure 5. It is worth remarking how CDA is able to provide challenging samples (due to truncated spacecrafts) as well as high signal to noise ratio samples as seen in Figure 5. To conclude, both HC and CDA are fully compatible and can be combined in hopes of further enhancing LSPnet.

4. Experiments

A series of trainings and experimental setups have been carried out using the Spacecraft Pose Estimation Dataset (SPEED) [20]. This dataset offers a set of $12,000$ grayscale pose-labeled synthetic images of size 1200×1920 portraying a spacecraft in space. The rendered images cover a wide range of different distances from the camera to the satellite (from $5m$ to $40m$ approximately). The SPEED images randomly include a realistically rendered Earth on the background, offering a series of challenging samples to predict. The labels for the test set of SPEED have not been disclosed meaning that any comparison done with results of SPEED will not be based on the same test set. Assuming the test set of SPEED follows the same data distribution as its train set, and in hopes of estimating the performance that LSPnet would obtain over the test set, a fixed random split of $10,000$ and $2,000$ images is performed. All the carried trainings share the following common parameters:

1. Batch size of $N_B = 16$ samples
2. Input images are rescaled to 256×409 pixels (thus maintaining the aspect ratio)

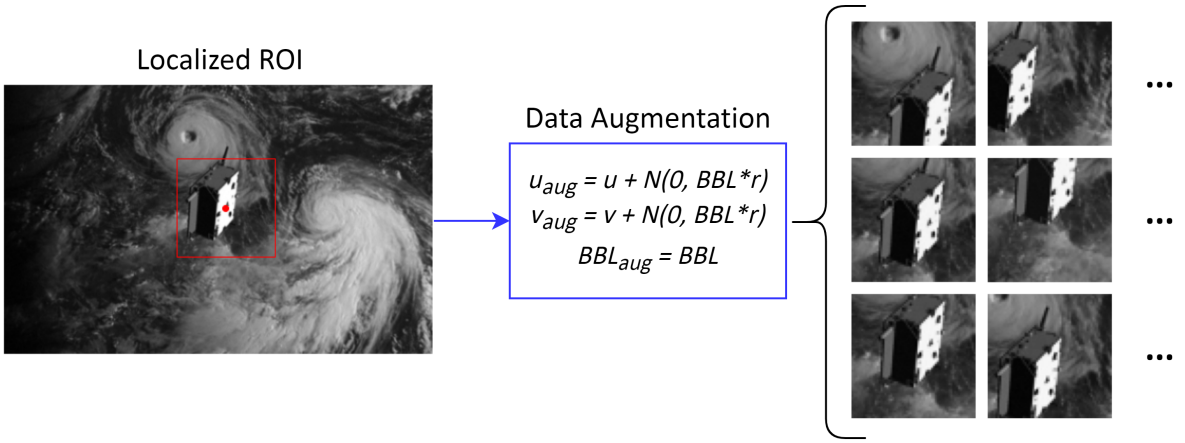

Figure 5. Data augmentation technique, named CDA, for augmenting the bounding boxes extracted from the outputs of the Translation module. The depicted example uses a value of $r = 0.15$ meaning that the standard deviation of each Normal distribution equals to 15% of the bounding box length.

3. Adam optimizer with a learning rate of $1e-4$
4. Learning rate decay of $1/2$ based on reaching a loss plateau
5. Train for as many epochs as needed until convergence
6. When HC is implemented, a total of $H = 64$ non-normalized heatmaps are predicted by the Localization CNN
7. When CDA is implemented, the hyperparameter r is fixed to 0.15

The loss function used to optimize the prediction of t is the Mean Squared Error (MSE) loss which is presented in the following equation

$$L_{position} = MSE(t, \hat{t}) = \frac{1}{N_B}\sum_{i=1}^{N_B}(t_i - \hat{t}_i)^2 \quad (5)$$

where t is the batch of ground truth translation vectors, $\hat{t}$ is the batch of predicted translation vectors, N_B is the batch size and t_i ($\hat{t}_i$ respectively) corresponds to the i-th translation vector within the batch.

Regarding the optimization of (u, v), the loss function selected is the one proposed in the work of DSNT [12] which is formulated as follows

$$L_{euc}(c, \hat{c}) = ||c - \hat{c}||_2 \quad (6)$$

$$L_{reg}(\hat{h}, \hat{c}) = D(\hat{h}||\mathcal{N}(\hat{c}, \sigma^2 * I_2)) \quad (7)$$

$$L_{center} = \frac{1}{N_B}\sum_{i=1}^{N_B} L_{euc}(c_i, \hat{c}_i) + \lambda L_{reg}(\hat{h}_i, \hat{c}_i) \quad (8)$$

where c is the batch of ground truth centers (u, v), $\hat{c}$ is the batch of predicted centers $(\hat{u}, \hat{v})$, $||\cdot||_2$ is the 2-norm, $\hat{h}$ is the batch of predicted normalized heatmaps, $D(\cdot||\cdot)$ is a divergence measure, I_2 is the 2×2 identity matrix, N_B is the batch size and c_i ($\hat{c}_i$, $\hat{h}_i$ respectively) corresponds to the i-th element within the batch. Based on experimental findings of [12], the hyperparameters σ^2 and λ have been fixed to 1 and the divergence measure $D(\cdot||\cdot)$ has been selected to be the Jensen-Shannon divergence. Combining both Equations (5) and (8), the final loss for the Translation module is formed

$$L_{translation} = L_{position} + L_{center} \quad (9)$$

Following common practice for optimizing quaternion estimation, the following loss has been implemented

$$L_{rotation} = \frac{1}{N_B}\sum_{i=1}^{N_B} 2 * arccos(| < q_i, \hat{q}_i > |) \quad (10)$$

where N_B is the batch size, $< \cdot, \cdot >$ represents the dot product, $|\cdot|$ is the absolute value function and q_i ($\hat{q}_i$ respectively) corresponds to the i-th quaternion within the batch. Finally the complete loss for optimizing LSPnet entirely, *i.e.* the Pose module, is formulated as

$$L_{pose} = L_{translation} + L_{rotation} \quad (11)$$

A set of ablation studies have been performed in hopes of finding a highly performing LSPnet architecture specification. The following sections cover the performed ablation studies as well as a final comparison with state-of-the-art over the SPEED dataset. The error metrics used throughout the following sections are formalized here

$$E_c = \frac{1}{N}\sum_{i=1}^{N}|t_i^c - \hat{t}_i^c|, \ with \ c \in \{x, y, z\} \quad (12)$$

$$E_t = \frac{1}{N}\sum_{i=1}^{N}|t_i - \hat{t}_i|_2 \quad (13)$$

Table 1. LSPnet ablation study covering Orientation CNN initialization as well as HC and CDA implementation.

Orientation CNN init.	HC	CDA	E_t	E_q (deg)
Random	✗	✗	$\mathbf{0.519 \pm 1.047}$	36.13 ± 41.23
ImageNet	✗	✗	$\mathbf{0.519 \pm 1.047}$	22.36 ± 37.33
Random	✓	✗	0.588 ± 1.187	33.22 ± 38.78
ImageNet	✓	✗	0.602 ± 1.136	37.64 ± 41.11
ImageNet	✗	✓	$\mathbf{0.519 \pm 1.047}$	$\mathbf{15.70 \pm 23.61}$
Random	✓	✓	0.596 ± 1.106	33.05 ± 36.38

Table 2. Ablation study on Localization CNN enhancing the translation estimation task when connected to Position CNN.

Localization CNN	E_x	E_y	E_z	E_t
✗	0.0571	0.0573	0.519	0.539
✓	**0.0551**	**0.0558**	**0.498**	**0.519**

Table 3. Position CNN initialization study.

Initialization	E_x	E_y	E_z	E_t
Random	0.0746	0.0816	0.666	0.694
ImageNet	**0.0551**	**0.0558**	**0.498**	**0.519**

$$E_q = \frac{1}{N}\sum_{i=1}^{N} 2 * arccos(| < q_i, \hat{q}_i > |) \qquad (14)$$

where N is the size of the dataset being evaluated and t_i^c ($\hat{t_i^c}$ respectively) corresponds to the coordinate value c (x, y or z) of the i-th translation vector within the dataset. If not specified otherwise, E_c and E_t are expressed in meters (m) while E_q is expressed in radians (rad).

4.1. Ablation studies

During all the ablation studies performed, both CNN architectures for the Position CNN and the Orientation CNN have been fixed to ResNet18. Based on the following findings, the best performing configuration is selected and scaled up to ResNet50 before comparing to state-of-the-art.

It has been surmised that the Localization CNN aids the Position CNN in the process of translation estimation. To shed light on this claim, Table 2 presents translation results when using Position CNN alone in comparison to including the Localization CNN. Note that both architectures have been initialized with ImageNet weights to ensure fair comparison. A slight decrease in translation error points to the idea that Localization CNN aids, to some extent, the translation estimation task.

Once Localization CNN has been found to aid in optimizing the predictions of t, a brief comparison between ImageNet initialization and Random initialization of the Position ResNet weights has been performed and can be found in Table 3. Having fixed the initialization of the Position CNN to ImageNet weights, due to its highly positive impact, a complete ablation study has been carried covering the following architectural and training decisions: (1) Orientation CNN initialization, (2) implementation of HC and (3) implementation of CDA. Table 1 presents all the results obtained during this complete ablation study. Note that when HC is not implemented the Translation module used is the best one obtained among Table 2 and Table 3. Also note that not all the possible combinations have been tested. This is due to the findings that have been extracted throughout the ablation study process which are the following,

- Similarly to Position CNN initialization, Orientation CNN ImageNet initialization significantly improves rotation estimation when not including HC, *i.e.* when the input is only composed by the ROI crop.

- It has been found that when HC is implemented, and thus the input of the Orientation CNN is a tensor with $H + C_I$ channels, ImageNet initialization negatively impacts the orientation performance. This can be justified due to the drastically different data distributions from ImageNet with respect to the non-normalized heatmaps predicted by the Localization CNN. In this sense a random initialization is more fitting to the specialized data distribution introduced by such heatmaps.

- Based on the previous findings, the remaining configurations worth testing are ImageNet initialization with CDA as well as Random initialization with both HC and CDA.

- It is also found that when training the full Pose module at the same time, meaning that HC is implemented to enable an end-to-end differentiable training, the translation error slightly increases. This is due to the fact that in this scenario the Translation module has to be optimized to both predict the translation vector t as well as transfer orientation-meaningful heatmaps to the Orientation CNN. When no HC is implemented the Translation module is completely decoupled from orientation estimation and thus it is solely trained for position-related tasks.

Table 4. SPEC results and comparison.

Rank	Team	E_t	E_q (deg)	PnP
1	UniAdelaide [4]	0.032 ± 0.095	0.41 ± 1.50	Yes
2	EPFL cvlab	0.073 ± 0.587	0.91 ± 1.29	Yes
3	pedro fairspace [17]	0.145 ± 0.239	2.49 ± 3.02	No
-	SLAB Baseline [15]	0.209 ± 1.133	2.62 ± 2.90	Yes
⋮	⋮	⋮	⋮	⋮
7	Gabrie1A	0.318 ± 0.323	12.03 ± 12.87	No
-	LSPnet (Ours)	0.456 ± 1.010	13.96 ± 20.13	No
8	stainsby	0.714 ± 1.012	17.75 ± 22.01	No
9	VSI_Feeney	0.734 ± 1.273	23.42 ± 33.57	No
10	jblumenkamp	2.656 ± 2.149	35.92 ± 49.72	Yes

Table 5. SPN and LSPnet comparison results.

Model	E_x	E_y	E_z	E_q (deg)
SPN [20]	0.055	0.046	0.78	**8.43**
LSPnet (Ours)	**0.048**	**0.045**	**0.44**	13.96

- Lastly, when randomly initializing the Orientation CNN, only implementing HC improves orientation results at the expense of slightly worsening the translation results. Furthermore, when implementing both HC and CDA the orientation results very slightly improve (and the translation error slightly increases).

Overall, the best performing configuration found does not implement HC, implements CDA and initializes the Orientation CNN with ImageNet weights. Such combination means that the translation and orientation tasks are better optimized by LSPnet when decoupled. It is worth noting how CDA greatly improves orientation estimation.

4.2. State-of-the-art comparison

Given all the architectural decisions taken based on the presented ablation studies, the best performing LSPnet is chosen and compared to the SPEED state-of-the-art. A first comparison is performed with the SPN network proposed in the same SPEED work [20]. Table 5 presents the obtained comparisons. LSPnet significantly surpasses SPN in depth estimation (z coordinate of the translation vector). SPN, on the other hand, is capable of predicting more accurately the rotation quaternion q. For fair comparison, it is needed to highlight the fact that SPN relies on 3D information, performs Object detection through an off-the-shelf Object detection Deep Learning model and refines the predicted pose. Conversely, LSPnet achieves competitive results without requiring any of the aforementioned characteristics. The final comparison is done thanks to the latest SPEC challenge [8] which also relied on the SPEED dataset. For this reason, the results presented in SPEC are fit to be referenced in order to localize LSPnet into the uncooperative spacecraft state-of-the-art for pose estimation. The SPEC challenge involved nearly 50 teams working towards pose estimation on the SPEED dataset for 5 months. Table 4 shows the obtained results by LSPnet in the context of the top 10 ranking teams of SPEC. LSPnet can be *ranked* at the top 8 position. According to SPEC, a total of seven teams reconstructed the 3D model of the spacecraft to further use it in a keypoint-based solution (*e.g.* in combination with a PnP solver). SPEC also found that a recurring technique across the teams is the detection of the spacecraft in the image through Object detection Deep Learning models (such as YOLO) or through image segmentation. Moreover, pose refinement steps can also be found among the teams (*e.g.* [4]). LSPnet is capable of *ranking* top 8 while not relying on any 3D information, not refining the pose and implementing a simple yet efficient Object detection technique (easily augmented by CDA). Note that even though we are directly comparing LSPnet results with SPEED state-of-the-art results, our results are based on a different test set. This means that all the comparisons done through this section should be taken as indications of how LSPnet performs with respect to the state-of-the-art.

5. Conclusions

The goal of the here presented work is to provide the space literature with a simpler yet still effective solution for pose estimation of uncooperative spacecrafts which does not require prior 3D information nor involves pose refinement. Additionally, the proposed model is capable of generating bounding boxes without relying on a complex Object detection model and without needing bounding boxes labels (only requiring translation ground truths). It is shown how LSPnet achieves comparable results with respect to SPEED state-of-the-art. Extensions of this work may target spacecraft generalization as well as pose tracking.

References

[1] 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. 3

[2] Vincenzo Capuano, Shahrouz Ryan Alimo, Andrew Q Ho, and Soon-Jo Chung. Robust features extraction for on-board monocular-based spacecraft pose acquisition. In *AIAA Scitech 2019 Forum*, page 2005, 2019. 2

[3] 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. 1, 2, 3

[4] 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. 3, 8

[5] Xiaodong Du, Bin Liang, Wenfu Xu, and Yue Qiu. Pose measurement of large non-cooperative satellite based on collaborative cameras. *Acta Astronautica*, 68(11-12):2047–2065, 2011. 2

[6] Simone D'Amico, Mathias Benn, and John L Jørgensen. Pose estimation of an uncooperative spacecraft from actual space imagery. *International Journal of Space Science and Engineering 5*, 2(2):171–189, 2014. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE CVPR, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[8] M. Kisantal, S. Sharma, T.H. Park, D. Izzo, M. Märtens, and S. D'Amico. Satellite pose estimation challenge: Dataset, competition design and results. *IEEE Transactions on Aerospace and Electronic Systems*, 2020. 2, 3, 8

[9] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society CVPR, 2004.*, volume 2, pages II–104. IEEE, 2004. 3

[10] Mohamed Adel Musallam, Kassem Al Ismaeil, Oyebade Oyedotun, Marcos Damian Perez, Michel Poucet, and Djamila Aouada. Spark: Spacecraft recognition leveraging knowledge of space environment, 2021. 1

[11] Bo J Naasz, Richard D Burns, Steven Z Queen, John Van Eepoel, Joel Hannah, and Eugene Skelton. The hst sm4 relative navigation sensor system: overview and preliminary testing results from the flight robotics lab. *The Journal of the Astronautical Sciences*, 57(1-2):457–483, 2009. 2

[12] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *CoRR*, abs/1801.07372, 2018. 4, 6

[13] Roberto Opromolla, Giancarmine Fasano, Giancarlo Rufino, and Michele Grassi. Uncooperative pose estimation with a lidar-based system. *Acta Astronautica*, 110:287–297, 2015. Dynamics and Control of Space Systems. 1

[14] Roberto Opromolla, Giancarmine Fasano, Giancarlo Rufino, and Michele Grassi. A review of cooperative and uncooperative spacecraft pose determination techniques for close-proximity operations. *Progress in Aerospace Sciences*, 93:53 – 72, 2017. 1

[15] Tae Ha Park, Sumant Sharma, and Simone D'Amico. Towards robust learning-based pose estimation of noncooperative spacecraft. *arXiv preprint arXiv:1909.00392*, 2019. 3, 8

[16] Thaweerath Phisannupawong, Patcharin Kamsing, and et al. Vision-based spacecraft pose estimation via a deep convolutional neural network for noncooperative docking operations. *Aerospace*, 7(9):126, 2020. 3

[17] 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. 3, 8

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE CVPR*, pages 779–788, 2016. 2

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2

[20] Sumant Sharma and Simone D'Amico. Neural network-based pose estimation for noncooperative spacecraft rendezvous. *IEEE Trans. Aerosp. Electron. Syst.*, 56(6):4638–4658, 2020. 3, 5, 8

[21] Sumant Sharma and Simone D'Amico. Reduced-dynamics pose estimation for non-cooperative spacecraft rendezvous using monocular vision. In *38th AAS Guidance and Control Conference, Breckenridge, Colorado*, 2017. 2

[22] Jian-Feng Shi, Steve Ulrich, and Stephane Ruel. Spacecraft pose estimation using principal component analysis and a monocular camera. In *AIAA Guidance, Navigation, and Control Conference*, page 1034, 2017. 3

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[24] Shubham Sonawani, Ryan Alimo, Renaud Detry, Daniel Jeong, Andrew Hess, and Heni Ben Amor. Assistive relative pose estimation for on-orbit assembly using convolutional neural networks. *arXiv preprint arXiv:2001.10673*, 2020. 3

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE CVPR*, pages 1–9, 2015. 3

[26] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257, 2019. 3