A Lightweight Model for Satellite Pose Estimation

Pierluigi Carcagnì(✉), Marco Leo, Paolo Spagnolo, Pier Luigi Mazzeo, and Cosimo Distante

CNR-ISASI, Ecotekne Campus via Monteroni snc, 73100 Lecce, Italy
pierluigi.carcagni@cnr.it

Abstract. In this work, a study on computer vision techniques for automating rendezvous manoeuvres in space has been carried out. A lightweight algorithm pipeline for achieving the 6 degrees of freedom (DOF) object pose estimation, i.e. relative position and attitude, of a spacecraft in a non-cooperative context using a monocular camera has been studied. In particular, the considered lite architecture has been never exploited for space operations and it allows to be compliant with operational constraints, in terms of payload and power, of small satellite platforms. Experiments were performed on a benchmark Satellite Pose Estimation Dataset of synthetic and real spacecraft imageries specifically introduced for the challenging task of the 6DOF object pose estimation in space. Extensive comparisons with existing approaches are provided both in terms of reliability/accuracy and in terms of model size that ineluctably affect resource requirements for deployment on space vehicles.

Keywords: Spacecraft pose estimation · 6DOF pose · Deep learning · Monocular vision · Space imagery

1 Introduction

Estimating the 6 degrees of freedom (6DOF) pose of space-borne objects (e.g., satellites, spacecraft, orbital debris) is a crucial step in many space operations such as docking, non-cooperative proximity tasks (e.g., debris removal), and inter-spacecraft communications (e.g., establishing quantum links). It has unique challenges that are not commonly encountered in the terrestrial setting. Due to the importance of the problem, the Advanced Concepts Team (ACT) at ESA recently held a benchmark competition called Kelvins Pose Estimation Challenge (KPEC)[1].

The 6DOF pose estimation of objects in images is a traditional computer vision task. Methods based on template matching [21] were initially used to this aim. Unfortunately, when an object undergoes occlusions or drastic illumination changes they become unreliable. To fix this issues local feature/keypoint matching approaches were introduced [11]. The extraction of the fiducial points for

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

S. Sclaroff et al. (Eds.): ICIAP 2022, LNCS 13231, pp. 3–14, 2022.
https://doi.org/10.1007/978-3-031-06427-2_1

the purpose of calculating the correspondences is carried out using handcrafted descriptors such as Harris corners or Canny edges, lines e.g. by Hough transform, or scale invariants such as SIFT, SURF and ORB [13]. The object poses are subsequently obtained by solving a Perspective-n-Point (PnP) problem [10]. The above mentioned methods based on handcrafted descriptors, tend to produce low-quality outputs in difficult conditions (weak or absent texture of the surfaces, strong lights, etc.) typical in an operational context in space.

To overcome these drawbacks, recent advances in pose estimation techniques, for terrestrial applications, have been based on deep learning (DL) algorithms instead. In general, these algorithms bypass the classic pipeline based on handcrafted features, and instead try, through the use of an appropriate deep convolutional neural network (DCNN), to learn, in an end-to-end way, the non-linear transformation between the two-dimensional space of the input image and the six-dimensional exposure space of the network output. Learning is carried out through appropriate supervised training.

Deep learning strategies have demonstrated robust behaviour under difficult operating conditions in terms of scene illumination and object surface texture. However, for efficient operations in terms of processed fps (frames per second) it is necessary to have appropriate hardware with power constraints as opposed to the classic approaches based on handcrafted descriptors.

Besides, they require a large amount of manual labels including the 2D keypoints, masks, 6D poses of objects, and other extra labels, which are usually very costly. Many recent 6DOF pose estimation methods exploited 3D object models to generate synthetic images for training because labels come for free. However, due to the domain shift of data distributions between real images and synthetic images, the network trained only on synthetic images fails to capture robust features in real images for 6DOF pose estimation [19]. Another effective pathway could be to combine the strength of deep neural networks and geometric optimisation for example by incorporating a perspective-n-point (PnP) solver in a deep neural architecture [3]. Most 6DOF pose estimation deep networks rely on an encoder-decoder architecture. To handle large scale variations for 6D object pose estimation they can rely on an additional object detection network or they exploit the inherent hierarchical architecture of the encoder network, which extracts features at different scales [6]. Among different deep architectures, the High-Resolution Network (HRNet) [15], initially introduced for human pose estimation [17], has also recently gathered very relevant results also for object detection and semantic segmentation. Differently from existing state-of-the-art frameworks, that first encode the input image as a low-resolution representation through a sub-network that is formed by connecting high-to-low resolution convolutions in series (e.g., ResNet, VGGNet), and then recover the high-resolution representation from the encoded low-resolution representation, HRNet maintains high-resolution representations through the whole process. The benefit is that the resulting representation is semantically richer and spatially more precise.

This scientific fervour and the increased accessibility of space platforms, and space data, recently pushed researchers to investigate the application of machine learning researches in space activities. One of these 'research frontiers' concerns the pose estimation of space objects for autonomous rendezvous or for capturing uncontrolled targets in debris removal operations. Spacecraft (vehicles designed for operation outside the earth's atmosphere) and satellites (objects that orbit a natural body) have two types of systems: payload, which comprises instruments that facilitate the primary purpose of the spacecrafts; and operations systems, which support the payload and allow it to reach, stay, and work in space. Modern space compute systems are moving towards shared/re-configurable, multicore systems which would be capable of running just lite DL models. Besides, another limiting factor for on-board compute is power [9].

As a consequence, not all the DL models developed for terrestrial operations can be exploited in space and then it is important to investigate how to find a good trade-off between model complexity and accuracy/reliability.

In this work, a study on computer vision techniques for automating rendezvous manoeuvres in space has been carried out. In particular, algorithms for estimating the 6DOF object pose estimation, i.e. relative position and attitude, of a spacecraft in a non-cooperative context using a monocular camera have been studied. The term "non-cooperative" implies that the target spacecraft does not have an active communication link or markers such as LEDs or reflectors useful for distance and attitude estimation. To this aim, in this work, a 6DOF system suited to be exploited on a Satellite Platform, in 50-kg micro-satellite class, has been developed. The paper introduces and assesses, for the specific challenging task of the 6DOf object pose estimation in space, a lite architecture inspired by the Lite-HRNet [18] recently introduced for human pose estimation. By our knowledge, this is the first attempt to exploit the lite-HRNet for space operations. Experiments were performed on the Spacecraft PosE Estimation Dataset (SPEED) [14], the first publicly available machine learning set of synthetic and real spacecraft imageries. Extensive comparisons with existing approaches are provided both in terms of reliability/accuracy and in terms of model size that ineluctably affect resource requirements for deployment on space vehicles [5].

The rest of the paper is organized as follows: Sect. 2 describes the problem, the proposed algorithmic pipeline introduced to address it and the dataset of space images used for experimental tests. Subsequently, Sect. 3 accurately describes experimental results and it provides a deep discussion about the advisability of the proposed trade-off between accuracy and computational requirement of resources with respect to existing works in the literature. Finally, Sect. 4 concludes the paper.

2 Methodology and Data

In this work, a PnP based pose estimation approach, exploiting keypoints extracted by means of deep learning techniques, has been employed [16]. There are extensive applications based on this approach in the literature, including

action recognition, human-computer interaction, intelligent photo editing, pedestrian monitoring, etc. [7]. In particular, it is possible to divide the problem into two main categories: top-down methods and bottom-up methods. According to the top-down paradigm, the object of which to estimate the pose is first detected and then its pose is estimated by exploiting only the informative content of the region into the bounding box surrounding the identified object. The bottom-up paradigm directly regresses the positions of the keypoints belonging to the same object, or it detects all the keypoints in the scene and subsequently it groups the keypoints on the same object. Although the top-down paradigm is more expensive from a computational point of view, since in addition to the extraction of keypoints it requires a preliminary phase of detection of the object of interest, it is more accurate than the bottom-up paradigm [4].

2.1 Processing Pipeline

In this paper, a top-down paradigm has been exploited. The starting point was the pipeline introduced in [2] and depicted in Fig. 1. The algorithmic components of the pipeline have been modified in order to make the entire pipeline suitable for use with embedded systems, and then to be compliant with computational constraints for in space operations where power and resources are much more limited than for terrestrial tasks. The pipeline is model-based, i.e. it relies on the availability of a 3D model of the target object and the pose of the onboard camera has to be estimated with respect to the actual model configuration extracted from the acquired images. Hence, the pipeline consists of three main processing modules performing:

1. the detection of the object in the image;
2. the keypoints estimation in the detected bounding-box of the object;
3. the computation of the 2D-3D keypoints correspondences, i.e. between the available 3D points of the 3D model of the target and the estimated 2D ones of the detected object, and final pose estimation by means of a PnP (Perspective-n-Point) based algorithm.

Taking into account the limited computational resources and energy consumption constraints on board of a spacecraft/satellite, the underlying idea of this work is to address the keypoints extraction task of the target vehicle by a 'lite' deep convolutional network in order to reduce the computational complexity and the size of the trained model, in terms of parameters and memory occupation.

In this paper, a Lite-HRNet architecture [18] has been implemented and tested for the keypoints detection task, lowering this way the architectural complexity, compared to HRNet based approach exploited in [2], and therefore achieving the results of making the pipeline compliant with the computational resources available on board of the vehicle.

The Lite-HRNet has been built-up following the same architectural strategies exploited for HRNet. Starting from a high-resolution convolutional subnetwork

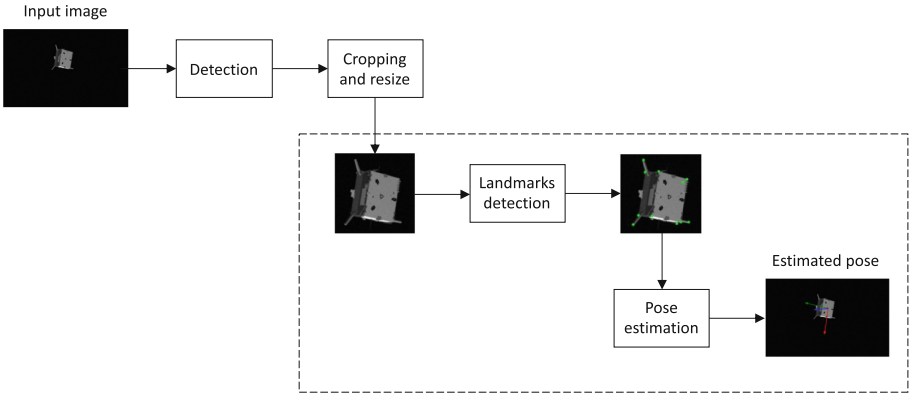

Fig. 1. Pose estimation pipeline.

as the first stage, high-to-low resolution subnetworks are gradually added and connected in parallel, keeping the high-resolution of the initial image through the whole process (Fig. 2).

Complexity is reduced by applying the efficient shuffle block, introduced with ShuffleNet [20], to the HRNet architecture and then using fewer layers and smaller widths. Moreover, the costly point-wise 1×1 convolution operation, heavily used in the original shuffle blocks, has been replaced by a lightweight unit, named conditional channel weighting. This allows, exploiting element-wise weighting operations, the architecture to obtain a linear complexity with respect to the number of channels instead of the quadratic complexity of the 1×1 convolution operations in the original implementation of the shuffle block. Finally, two lightweight functions have been introduced, a cross-resolution weighting function and a spatial weighting function, in order to compute the weight maps from all the channels across resolutions and for each resolution respectively compensating the role played by the point-wise 1×1 convolution operation (Fig. 3).

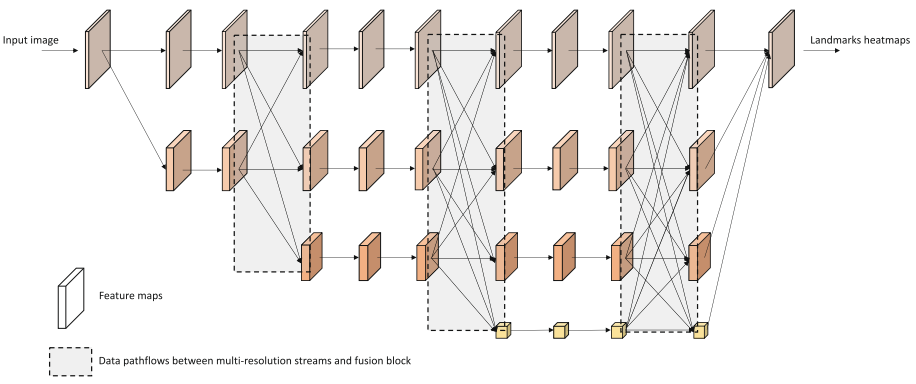

Fig. 2. Starting from an high-resolution convolutional subnetwork as first stage, high-to-low resolution subnetworks are gradually added and connected in parallel maintaining high-resolution through the whole process.

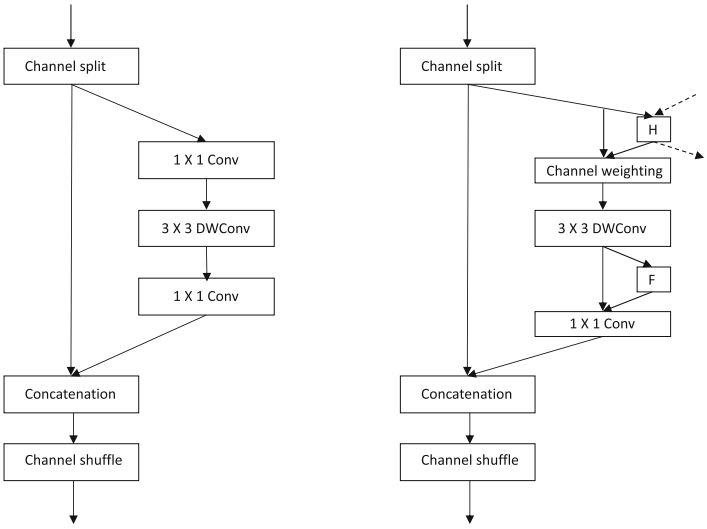

Fig. 3. (a) The shuffle block. (b) Conditional channel weighting block. H denotes the cross-resolution weighting function, F denotes the spatial weighting function. Weights distribution and representations from and to other resolutions are represented by the dotted lines.

In this work, in particular, the Lite-HRNet-18 implementation, where suffix 18 indicates the number of layers, has been exploited. A size of 768×768 pixels (the same of the solution in [2]) has been set for the input window and for the output heatmaps.

2.2 The Dataset

For training and validating the proposed pipeline, the SPEED (Spacecraft Pose Estimation Dataset) dataset [8] was exploited. It consists of 8-bit monochrome images in JPEG format with resolution 1920×1200 pixels. The dataset has three main folders of images: a folder containing 12000 synthetic images for training, a folder with 2998 similar synthetic images for testing and a folder with 300 real images of the Tango satellite mock-up, same format and resolution as the synthetic images. Ground truth data, in terms of 6DOF poses (position and orientation), is provided only for the images in the training set. Some of the training images are shown in Fig. 4. Images are particularly challenging due to the large variations for the satellite in terms of lighting condition, distance from the camera, orientation and background.

The camera model used for rendering the synthetic images is the same one as the actual camera used for capturing the 300 images of the mock-up. The related intrinsics camera parameters are: $resolution = 1920 \times 1200$ pixels, focal

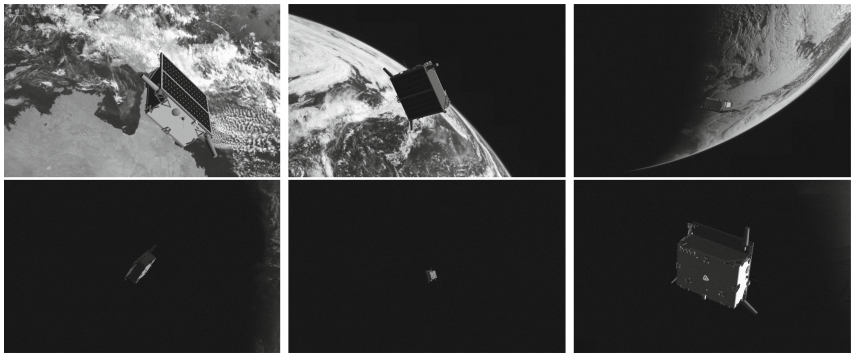

Fig. 4. Some Tango satellite images, at different distances from the camera, sampled from the SPEED dataset.

length $f = 17.6$ mm, Pixel pitch $p = 5.86$ μm/px, Horizontal $FoV = 35.452°$, Vertical $FoV = 22.595°$.

3 Experimental Results

Algorithmic pipeline implementation and testing were carried out using the Pytorch framework, and the Python language, in Ubuntu Linux 20.04 environment on a machine equipped with: Intel i7 processor; 64 GB of RAM; NVIDIA TITATN RTX GPU card with 24 GB of RAM.

In the first experimental phase the 3D model of the Tango satellite mock-up has been estimated (3D model and landmarks annotations are not provided with SPEED dataset but only 6DOF ground truth for the training set). This estimation has been carried out by picking up 9 close-up images from the training folder. In each image, 11 keypoints were manually selected: they correspond to some of the strongest visual characteristics in the images which, moreover, are not occluded by other surfaces of the vehicle space.

Figure 5 shows the configurations of the 11 selected keypoints on 3 training images.

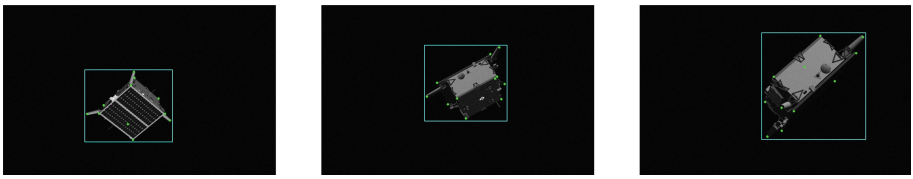

Fig. 5. Three configurations of the selected keypoints for the initial estimation of the 3D model of the Tango satellite mock-up.

Starting from the aforementioned 2D positions of the 11 keypoints in the 9 images and knowing the camera model parameters, the 3D structure of the Tango satellite has been reconstructed by multi-view triangulation.

In particular, multi-view triangulation was performed by minimizing an objective function

$$\sum_{i,j} \left\| p_{i,j} - \pi_{T_j^*}(\mathbf{x}_i) \right\|_2^2 \tag{1}$$

where $p_{i;j}$ denotes the 2D coordinates of the i-th landmark, obtained from the j-th image, and x_i the corresponding 3D landmark. T_j^* is the ground truth pose provided for the image j and π_T is the projective transformation (6DOF pose and intrinsic camera parameters are known) for the x_i 3D landmark onto the image plane. From Eq. 1 the 3D positions of the i selected landmarks were estimated in each image j.

Exploiting the 6DOF ground truth data in the dataset and knowing the estimated 3D model of the satellite, the ground truth 2D positions of landmarks for all training images were obtained by projecting x_i to the image plane by π_{T^*}. Finally, a bounding box was chosen in each image so that the 11 landmarks lie in it.

The 2D positions of the landmarks and the corresponding bounding boxes were then exploited to validate the proposed pipeline for 6DOF pose estimation.

The validation task was carried out by exploiting all available annotated images (i.e. the images in the training folder of the SPEED dataset) and a K-Fold Cross-Validation approach, with a number of folds equal to 6. In the cross-validation, 5 folds were used for training the Lite-HRNet. The input of the net were the patches obtained by cropping the original training image, around the available bounding boxes surrounding the landmarks and the corresponding 2D positions of the 11 selected landmarks. They were then resized to a common dimension of 768×768 pixels. In particular, the network was trained in order to regress 11 heatmaps, of the same size of the input patch, corresponding to the 11 selected landmarks. Ground truth heatmaps were generated as 2D normal distributions with mean equal to the ground truth 2D position of each landmark and standard deviation of 1 pixel.

The remaining fold was then used for validating the capability of the net to automatically estimate landmarks positions in unseen patches. This process was repeated 6 times changing the fold used for validation among the available ones.

The object poses in all the patches extracted for the 6 validation folds (12.000 images) were then obtained by solving a Perspective-n-Point (PnP) problem [10] exploiting the 2D-3D correspondences between the 11 predicted landmarks and the 3D structure model of the satellite mock-up.

Lite-HRNet was trained by scratch, for each of the 6 validation steps, employing the ADAM optimizer with starting learning rate = 0.001 (dropped by a 0.1 factor at the 120th and 170th epochs respectively), momentum = 0.9 and weight decay = 0.0001 parameters. A total number of 180 training epochs has been chosen.

A rotation and a translation error were finally computed. In particular, indicating with q^* and q the rotation quaternion ground truth and the estimated one, the rotation error E_r is defined as:

$$E_R = 2 \cdot cos^{-1}|q \cdot q^*|, \tag{2}$$

and the translation error is defined as:

$$E_T = \frac{\|t - t^*\|_2}{\|t\|_2}, \tag{3}$$

for an overall error:

$$E = E_R + E_T. \tag{4}$$

The experienced mean errors on validation folds were $\overline{E}_R = 0.0302$, $\overline{E}_T = 0.0075$ and $\overline{E} = 0.377$ respectively. The $\overline{E}$ error is plotted in Fig. 6 versus the distance between the camera and the target. In particular, the figure reports the mean total error $\overline{E}$ on the y-axis, whereas the x-axis indicates distance ranges from the camera, expressed in meters, of the corresponding detected TANGO vehicle. The green bars correspond are related to the proposed pipeline whereas the orange ones correspond to the error achieved by the pipeline in [2], where the non-lite version of the HRNet was exploited. For a fair comparison, the pipeline in [2] was applied in the identical operating conditions as for the proposed algorithmic pipeline, i.e. without the final iterative refinement, in order to not add external bias in the evaluation of the benefit in using the lightweight version of the HRNet. As expected, the error decreased while the distance increased. It is worth noting that for short distances errors for the two pipelines are comparable. Of course, the gap between the performance of the lite version of HRNet increased at long distances since, on small targets, the landmark positioning failed.

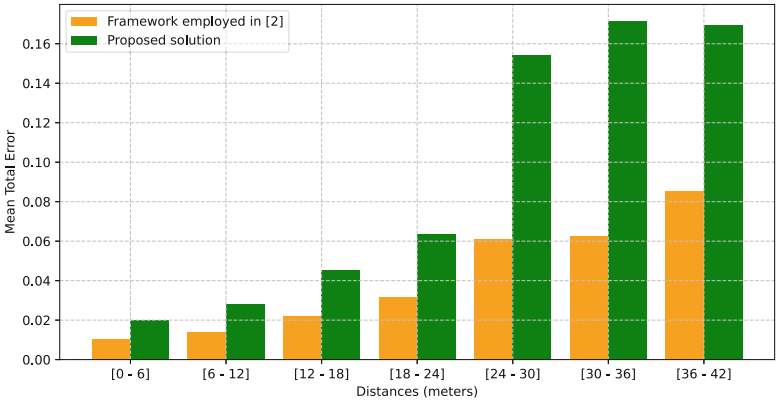

Fig. 6. Mean total error vs Distance of the target from the camera.

The numerical comparison between the proposed lite pipeline and the original one is reported in Table 1. The overall rotational and translation errors increased with a total error increasing from 0.0185 to 0.377. The rightmost column indicates the model size, in terms of number of network parameters (M), for both pipelines: it is worth noting that the original model had a size of 28.5 M whereas the proposed one has a size of 1.1 M (more than 25 times less).

Table 1. Errors and complexity comparison.

Approach	Mean rot. error	Mean trans. error	Mean total error	Network #Params
[2]	0.0141	0.0044	0.0185	28.5 M
Proposed	0.0302	0.0075	0.0377	1.1 M

Another relevant aspect to take into account is the average prediction time. The average time to predict the pose of the target vehicle by using the proposed pipeline is about 5 ms (i.e. 5.10 ms) whereas it is about 12 ms (i.e. 11.94 ms) by using the pipeline in [2].

Since, as largely stated in the introductory section, the computational load and the hardware resources are crucial points for space applications, the above numerical outcomes demonstrated that the introduce pipeline is particularly suited for the rendezvous process since it relies on a model with much fewer parameters allowing shorter computational time per image. To this aim, it is worth noting that a typical rendezvous process can be divided into several phases, including phasing of close-range rendezvous, final approaching, and docking. Relative navigation and control are mainly used in the close-range rendezvous phase, and orbit and attitude combined six-degree of freedom (6DOF) control is used in the final approaching phase [12]. Under this perspective, it is worth noting that, for close-range operations (up to about 20 m), the errors in pose estimations obtained with the Lite-HRNet are comparable with those gathered by state-of-the-art approaches relying on deep learning-based models.

4 Conclusions

This work has proven how the Lite-HRNet can be effectively exploited for 6DOF pose estimation for in space rendez-vous maneuvers. It represents the first attempt of using this recent full-resolution architecture for in space operations. Experimental validations on a benchmark dataset demonstrated that, for close range operations (up to 20 m), the errors in pose estimations are comparable with those gathered by state-of-the-art approaches relying on deep learning based models much more complex allowing, also, a half time of processing per image. Future works will deal with building a top-down pipeline that, relying on temporal/spatial filtering tricks, can alleviate also the computational load of the object detection step. Besides, a tool for generating synthetic datasets of

photorealistic GAN-Generated scenes [1] will be introduced in order to train the pipeline for any target.

Acknowledgement. This work was supported in part by the Ministry of Education, University and Research under the grant PM3 AER01_01181 Modular Multi-Mission Platform.

References

1. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
2. Chen, B., Cao, J., Parra, A., Chin, T.J.: Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
3. Chen, B., Parra, A., Cao, J., Li, N., Chin, T.J.: End-to-end learnable geometric vision by backpropagating PNP optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8100–8109 (2020)
4. Cheng, Y., Wang, B., Yang, B., Tan, R.T.: Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7649–7659 (2021)
5. Hu, X., Chu, L., Pei, J., Liu, W., Bian, J.: Model complexity of deep learning: A survey. arXiv preprint arXiv:2103.05127 (2021)
6. Hu, Y., Speierer, S., Jakob, W., Fua, P., Salzmann, M.: Wide-depth-range 6d object pose estimation in space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15870–15879 (2021)
7. Jeon, M.H., Kim, A.: Prima6d: rotational primitive reconstruction for enhanced and robust 6d pose estimation. IEEE Robot. Autom. Lett. **5**(3), 4955–4962 (2020)
8. Kisantal, M., Sharma, S., Park, T.H., Izzo, D., Märtens, M., D'Amico, S.: Satellite pose estimation challenge: Dataset, competition design, and results. IEEE Trans. Aerosp. Electron. Syst. **56**(5), 4083–4098 (2020)
9. Kothari, V., Liberis, E., Lane, N.D.: The final frontier: deep learning in space. In: Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications, pp. 45–49 (2020)
10. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: an accurate o (n) solution to the pnp problem. Int. J. Comput. Vision **81**(2), 155 (2009)
11. Li, Z., Wang, G., Ji, X.: Cdpn: coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7678–7687 (2019)
12. Luo, Y., Zhang, J., Tang, G.: Survey of orbital dynamics and control of space rendezvous. Chin. J. Aeronaut. **27**(1), 1–11 (2014)
13. Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J.: Image matching from handcrafted to deep features: A survey. Int. J. Comput. Vision **129**(1), 23–79 (2021)
14. Sharma, S., D'Amico, S.: Neural network-based pose estimation for noncooperative spacecraft rendezvous. IEEE Trans. Aerosp. Electron. Syst. **56**(6), 4638–4658 (2020)
15. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)

16. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 699–715 (2018)
17. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV) (2018)
18. Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J.: Lite-hrnet: a lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10440–10450 (2021)
19. Zhang, S., Zhao, W., Guan, Z., Peng, X., Peng, J.: Keypoint-graph-driven learning framework for object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1065–1073 (2021)
20. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
21. Zhu, M., et al.: Single image 3d object detection and pose estimation for grasping. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 3936–3943. IEEE (2014)