Improving satellite pose estimation across domain gap with generative adversarial networks

Alessandro Lotti[1,a *]

[1]Department of Industrial Engineering, Alma Mater Studiorum Università di Bologna
Via Fontanelle 40, 47121, Forlì, Italy

[a]alessandro.lotti4@unibo.it

Keywords: Pose Estimation, Computer Vision, Vision-Based Navigation, Domain Gap, Generative Adversarial Networks

Abstract. Pose estimation from a monocular camera is a critical technology for in-orbit servicing missions. However, collecting large image datasets in space for training neural networks is impractical, resulting in the use of synthetic images. Unfortunately, these often fail to accurately replicate real image features, leading to a significant domain gap. This work explores the use of generative adversarial networks as a solution for bridging this gap from the data level by making synthetic images more closely resemble real ones. A generative model is trained on a small subset of unpaired synthetic and real pictures from the SPEED+ dataset. The entire synthetic dataset is then augmented using the generator, and employed to train a regression model, based on the MetaFormer architecture, which locates a set of landmarks. By comparing the model's pose estimation accuracy on real images with and without generator preprocessing, it is observed that the augmentation effectively reduces the median pose estimation error by a factor 1.4 to 5. This compelling result validates the efficacy of these tools and justifies further research in their utilization.

Introduction

Accurately navigating an active probe around a target spacecraft is crucial for many space missions, such as satellite servicing and debris removal. In this context, the ability to operate autonomously is essential to ensure that the servicer can make timely maneuvers and respond quickly to changing conditions, especially to avoid collisions. Therefore, the chaser shall be able to measure and control its state relative to the client throughout the rendezvous.

In this scenario, monocular cameras have emerged as an appealing sensor solution because of their low power consumption and small form factor. To estimate the relative chaser-to-target pose (i.e. position and attitude), a set of 2D landmarks extracted from the image is matched with the 3D model of the spacecraft using Perspective-n-Point solvers [1], as illustrated in Fig. 1. Image processing (IP) is a critical aspect of this software routine and recently many authors proposed to leverage neural networks (NNs) for feature extraction. Unfortunately, these models require large-scale datasets for training which are highly impractical to collect and label in orbit with accurate pose information. To overcome this challenge, researchers have turned to synthetic datasets generated through 3D computer graphics software [2,3,4]. On one side, these allow for rapid generation of thousands of images but, on the other side, they struggle to accurately reproduce the visual characteristics and large diversity of spaceborne pictures. This discrepancy between synthetic and real-world images is known as domain gap, and is a major obstacle to the adoption of NNs for vision-based navigation. Indeed, NNs tend to develop over-reliance on synthetic features, leading to inaccurate predictions when applied to real-world scenarios.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 license. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI. Published under license by Materials Research Forum LLC.

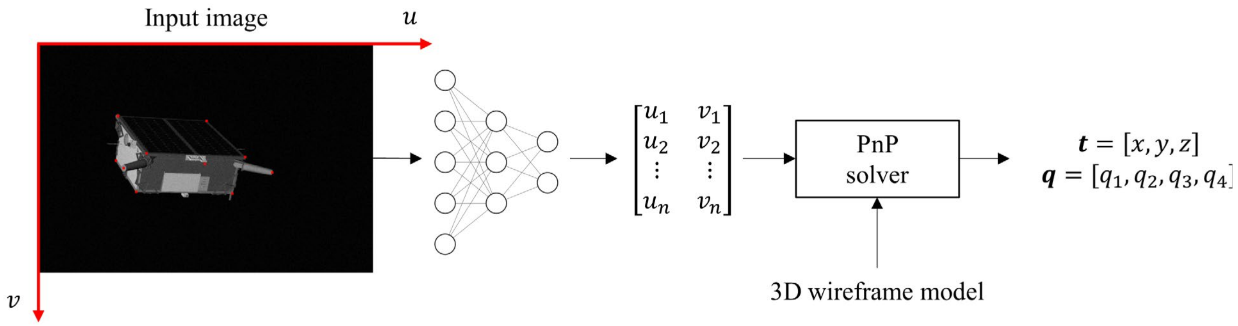
Input image
u
v
$\begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \\ \vdots & \vdots \\ u_n & v_n \end{bmatrix}$
PnP solver
3D wireframe model
$\boldsymbol{t} = [x, y, z]$
$\boldsymbol{q} = [q_1, q_2, q_3, q_4]$

Fig. 1 - Schematization of the pose estimation process.

While domain gap is actively researched within the computer vision community (see [5] for a throughout survey), few works addressed this issue in the framework of satellite pose estimation. According to the results of the ESA sponsored "Satellite Pose Estimation Competition 2021" [6], the use of adversarial training proved to be an effective technique. Although training the NN to also deceive a discriminator can help to learn domain invariant features, this method demands a large number of real pictures. Obtaining such images, however, can be costly and require specialized robotic facilities.

Generative adversarial networks (GANs) represent a possible approach to address this challenge by narrowing domain gap at the data level. Indeed, they could be employed to enhance synthetic datasets, making them more closely resemble real-world images.

GAN models typically comprise two neural networks, a generator and a discriminator, that are trained together to learn a mapping from a source domain to a target domain. The purpose is to transfer the visual characteristic of real pictures to synthetic images which should become indistinguishable from the real ones. GANs provide a versatile and cost-effective way to address the scarcity of spaceborne images as the same generator could in principle be applied to any synthetic dataset regardless of the depicted target.

This work demonstrates how the regression error on real images can be greatly reduced by simply preprocessing synthetic images through a trained GAN.

Methods

The SPEED+ dataset [7] is adopted for this study. This includes synthetic grayscale images (47966 for training and 11994 for validation) and real pictures of a spacecraft mockup captured in a laboratory environment at a maximum distance of 10 m from the camera. Fig. 2 displays synthetic and real samples, which feature two different illumination conditions, namely *lightbox* (6740) and *sunlamp* (2791). Throughout the workflow, input images are resized from the original 1200x1920 px resolution to 320x512 px to reduce the computational time while preserving the original aspect ratio. This project employs a state-of-the-art image-to-image (I2I) translation algorithm based on the contrastive learning method proposed in [8], which does not require paired samples from the source and target domains during training. The generator network is built using a Residual Network [9] with 9 residual blocks. The I2I model is trained on randomly selected subsets of 1000 synthetic and real images for 200 epochs, with a batch size of 1. The learning rate is linearly decreased from 2e-4 to 0 starting from epoch 100. Training is repeated separately for *sunlamp* and *lightbox* domains. Later, all synthetic images from the SPEED+ train partition are processed through the generator obtaining two sets of synthetic-enhanced pictures.

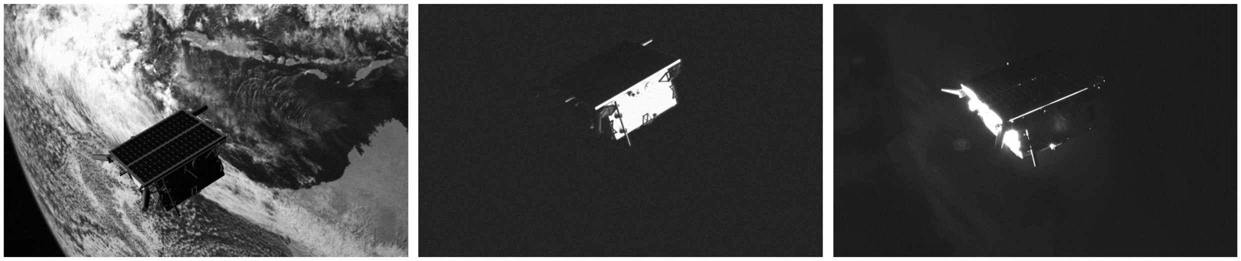

Fig. 2 - Images from the SPEED+ dataset [7] depicting the satellite in similar poses: synthetic (left), lightbox (center), sunlamp (right).

During translation, for each synthetic image a random sample from the 1000 real pictures is selected to provide a context for the I2I translation.

A collection of original and enhanced images is illustrated in Figure 3, showcasing the ability of the generator to capture the appearance of the target domain while preserving most of the content of the original image.

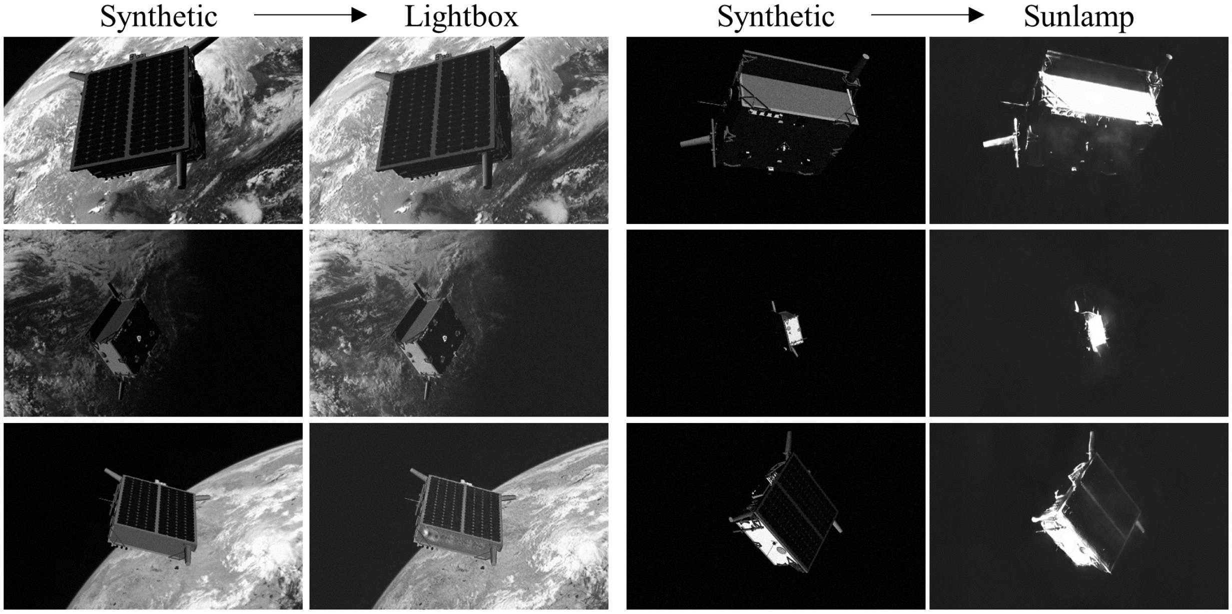

Fig. 3 - Synthetic vs GAN-enhanced images.

To assess the effectiveness of the learned mapping, a pose estimation pipeline according to the scheme illustrated in Fig. 1 has been set up. In this framework a NN is trained to predict the location of 11 landmarks on the enhanced-synthetic datasets and tested on actual *sunlamp* and *lightbox* images. To this end a *ConvFormerS18* model [10] is employed as the backbone of a 2-layer regression head, comprising a depthwise separable convolution, with ReLu activation, followed by a fully connected layer. The ConvFormer belongs to the family of MetaFormer models [11], inspired by Transformers [12], which showcased improved robustness to domain gap over convolutional neural networks [13]. The architecture leverages depthwise separable convolutions as token mixers, which are well-suited for embedded processors. The network is trained using mean absolute error loss for 40 epochs with a batch size of 48 images. The fitting is carried out on Google Colab's Tensor Processing Units with Adam optimizer and cosine decay learning rate, starting from 7.5e-5. The backbone is initialized with Imagenet weights [14] and common data augmentations are applied, these include random image rotation, brightness and contrast adjustments, blurring and gaussian noise.

At inference time, the regressed landmarks are fed into an EPnP solver [15] together with the 3D satellite model and camera parameters.

Results and Discussion

The pose regression error is then evaluated through the following metrics:

$$\boldsymbol{e_t} = |\boldsymbol{t_{BC}} - \boldsymbol{\hat{t}_{BC}}|_\mathbf{2} \tag{1}$$

$$\boldsymbol{\bar{e}_t} = \frac{\boldsymbol{e_t}}{|\boldsymbol{t_{BC}}|_\mathbf{2}} \tag{2}$$

$$\boldsymbol{e_q} = \mathbf{2}\,\boldsymbol{arccos}(|\boldsymbol{q^T}\boldsymbol{\hat{q}}|) \tag{3}$$

$$\boldsymbol{E} = \boldsymbol{\bar{e}_t} + \boldsymbol{e_q} \tag{4}$$

Where $\boldsymbol{t}_{BC}$, $\boldsymbol{q}$ and $\boldsymbol{\hat{t}}_{BC}$, $\boldsymbol{\hat{q}}$ represent respectively the ground truth and estimated position vectors and attitude quaternions aligning the chaser-mounted camera frame (C) and the target body frame (B). Table 1 compares the performance of the NN trained on synthetic enhanced images with that of the model trained without GAN preprocessing. The symbol $\langle \quad \rangle$ denotes the median operator. A solution is defined to be of high quality (HQ) if both the normalized position and rotation errors fall below specific thresholds. For the rotation error, the threshold is 5° when the satellite is within 5 meters, and 10° when it is farther away. Meanwhile, the limit for normalized position error is fixed at 0.1. Notably, GAN preprocessing allowed to decrease the median pose error of a factor 5 and 1.4 on *sunlamp* and *lightbox* respectively. The enhancement is further evidenced by the increase in the percentage of HQ solutions.

Table 1 - *Pose estimation errors obtained by training the NN with and without GAN preprocessing.*

GAN preprocessing	Lightbox				Sunlamp			
	$\langle e_t \rangle [m]$	$\langle e_q \rangle [deg]$	$\langle E \rangle$	HQ	$\langle e_t \rangle [m]$	$\langle e_q \rangle [deg]$	$\langle E \rangle$	HQ
With	**0.369**	**11.2**	**0.288**	**38.4%**	**0.131**	**5.08**	**0.116**	**67.6%**
Without	0.488	15.8	0.403	30.6%	0.580	25.5	0.597	15.8%

Conclusions and Future Work

Overall, the results of this preliminary study demonstrate that GAN preprocessing is effective in reducing the domain gap at the data level, with the added benefit of requiring only a small fraction of the dataset size for training. However, further investigations are needed to explain the discrepancy in the advantages achieved on the two domains, as illustrated in Table 1. Additionally, the reliability of the MetaFormer architecture for the IP step has also been demonstrated by its ability to achieve competitive pose estimation errors despite having only around 24 million parameters.

As a next step, the potential of fusing *lightbox* and *sunlamp* characteristics in a single generator will be investigated to produce even more realistic images that capture the visual features of both domains. The ability of the GAN to translate images depicting different targets from those featured in the training data, without compromising their content, will also be explored. Furthermore, the benefits of combining GAN preprocessing with other domain generalization methods, such as the extensive augmentations and multi-task learning solutions proposed in the literature [16], will be assessed.

References

[1] B. Chen, J. Cao, A. Parra, T.-J. Chin, Satellite Pose Estimation with Deep Landmark Regression and Nonlinear Pose Refinement, Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, IEEE, New York, 2019, pp. 2816–2824. https://doi.org/10.1109/ICCVW.2019.00343

[2] S., Sharma, S. D'Amico, Pose Estimation for Non-Cooperative Rendezvous Using Neural Networks, 2019 AAS/AIAA Astrodynamics Specialist Conference, American Astronautical Society Paper 19-350, Springfield, VA, 2019, pp. 1–20.

[3] P. F. Proenca, Y. Gao, Deep Learning for Spacecraft Pose Estimation from Photorealistic Rendering, 2020 IEEE International Conference on Robotics and Automation, IEEE, New York, 2020, pp. 6007–6013. https://doi.org/10.1109/ICRA40945.2020.9197244

[4] A. Lotti, D. Modenini, P. Tortora, M. Saponara, and M. A. Perino, Deep Learning for Real-Time Satellite Pose Estimation on Tensor Processing Units, Journal of Spacecraft and Rockets, 2023, pp. 1–5. https://doi.org/10.2514/1.A35496

[5] J. Wang, C. Lan, C. Liu, Y. Ouyang,T. Qin, W. Lu, Y. Chen, W. Zeng, P.S. Yu, Generalizing to Unseen Domains: A Survey on Domain Generalization, IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2022.3178128

[6] T. H. Park, M. Märtens, M. Jawaid, Z. Wang, B. Chen, T.-J. Chin, D. Izzo, S. D'Amico, Satellite Pose Estimation Competition 2021: Results and Analyses, Acta Astronautica, vol. 204, 2023, pp. 640–665. https://doi.org/10.1016/j.actaastro.2023.01.002

[7] T. H. Park, M. Martens, G. Lecuyer, D. Izzo, and S. D'Amico, SPEED+: Next-Generation Dataset for Spacecraft Pose Estimation across Domain Gap, IEEE Aerospace Conference (AERO), Big Sky, MT, USA, 2022, pp. 1–15. https://doi.org/10.1109/AERO53065.2022.9843439

[8] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, Contrastive Learning for Unpaired Image-to-Image Translation, European Conference on Computer Vision (ECCV), Springer, 2020, pp. 319–345.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, Perceptual Losses for Real-Time Style Transfer and Super-Resolution, European Conference on Computer Vision (ECCV), Springer, 2016, pp 694-711.

[10] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, X. Wang, MetaFormer Baselines for Vision. arXiv, 2022. https://doi.org/10.48550/arXiv.2210.13452

[11] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, MetaFormer Is Actually What You Need for Vision, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp 10819-10829.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems, 2017, pp 5998–6008.

[13] C. Zhang, M. Zhang, S. Zhang, D. Jin, Q. Zhou, Z. Cai, H. Zhao, X. Liu, Z. Liu, Delving Deep into the Generalization of Vision Transformers under Distribution Shifts, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp 7277-7286.

[14] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems, Vol. 25, No. 2, 2012, pp. 1106–1114. https://doi.org/10.1145/3065386

[15] V. Lepetit, F. Moreno-Noguer, and P. Fua, EPnP: An Accurate O(n) Solution to the PnP Problem, International Journal of Computer Vision, Vol. 81, No. 2, 2009, pp. 155–166. https://doi.org/10.1007/s11263-008-0152-6

[16] T.H. Park, S. D'Amico, Robust Multi-Task Learning and Online Refinement for Space-craft Pose Estimation across Domain Gap, 2022. ArXiv: abs/2203.04275