HiSST: 4th International Conference on
High-Speed Vehicle Science Technology
22 -26 September 2025, Tours, France

Deep Learning-Based Robust Optical Guidance for Hypersonic Platforms

Adrien Chan-Hon-Tong[1], *Aurélien Plyer*[1], *Baptiste Cadalen*[1], *Laurent Serre*[1]

Abstract

Sensor-based guidance is required for long-range platforms when GNSS can be denied. To bypass the structural limitations of the classical registration-on-reference-image framework, we offer in this paper to encode the appearance of the surrounding of the target (at all resolutions) from a stack of images of the scene into a deep network. This new framework is showed to be relevant on bimodal scene (e.g. when the scene can or can not be snowy) even if it raises question about the loss of epipolar geometry which is much more understood and mastered than gray-box deep networks.

Keywords: large scale guidance, deep learning, hypersonic platform

1. Introduction

Localizing a camera using the current image is as old as computer vision [1]. However, SLAM frameworks (Simultaneous Localization and Mapping) only offer relative localization. To restore absolute localization, one must combine the information provided by the current image with external information such as GNSS, or, anchor points (points for which the absolute 3D position is known) visible in the image. This last idea leads to the framework of registration on a reference image widely used in remote sensing: by using anchors recognized in the current image, PnP algorithms [7] allow restoring the absolute position of the camera (and even the related coordinate system see **Fig. 1a**).

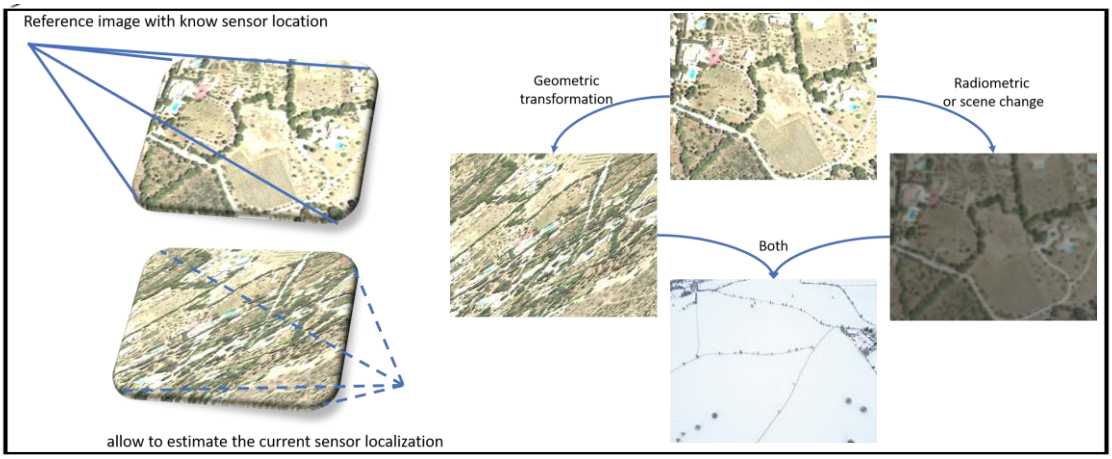

Fig 1. a) On the left, Principe of registration on reference image to recover absolution localization. **b)** On the right, this process may face both geometric transformation and/or physical ones.

However, image matching can be challenging (see **fig. 1b**): one should recognize an area despite appearance change due to different localization of the sensor (geometric transformation) or radiometric/physical change or both. Precisely, the state of the art mostly focuses the geometric accuracy of the matching in mostly controlled setting which is critical for accurate 3D reconstruction.

[1] *ONERA Université Paris-Saclay, firstname.lastname@onera.fr*

However, for recovering absolute position, geometric accuracy is less important than robustness to uncontrolled changes. Yet, the classical registration-on-reference-image framework suffers from a structural drawback: it is heavily dependent on the quality of the reference image and on the similarity between this reference image and the current one.

To bypass this limitation, this paper proposes relying instead on a stack of images of the scene to capture common changes that can arise (e.g., snowy or not) and to implicitly fill in missing information in each individual image (e.g., each image may contain clouds, but the entire scene can be seen across the stack). As manipulating the stack is inconvenient, we propose using a small deep network to directly learn a mapping between the current image and the orientation toward the target even if not visible.

This approach is particularly relevant for optical guidance of hypersonic platforms. In such contexts, GNSS can be denied, embedded accelerometers lead to large drifts after several thousand kilometers of travel and stellar-based-localization is unavailable during the decreasing part of the fly. Also, such platforms usually follow terminal trajectories while performing high speeds maneuvers that constrains the images that can be obtained, where the target itself could be outside the cameras field of view, and, besides, at very high altitude, the target is not visible even if in the sensor field of view. Furthermore, using a deep network can increase robustness to the precise spectral band and/or potential distortion created by heat when capturing the scene at very high speeds.

Despite this approach clearly introducing logistical issues (the need to collect a stack of images of the scene, the need to train a specific network for a single target, the lack of well-understood geometric foundations), we provide a case of a bimodal scene (with and without snow) where classical baseline fails while our method mitigates the bimodal issue. This model also performs correctly under large simulated heat noise potentially tackling this new issue introduced by hypersonic platform.

2. Related works

There is a very large literature on SLAM and registration, currently being revisited by the rise of efficient deep network methods for geometric tasks. SIFT+lightglue [5], which combines original SIFT [8] and efficient deep learning descriptors seems to be the current state of the art of image matching, challenged by new approaches performing end-to-end dense matching [3,4,11].
However, SIFT+lightglue focuses on robustness to point of view. So, it may be sensitive to strong changes in the appearance of the scene. End-to-end dense matching methods may be more robust to those changes by implicitly learning the existence of such drift (MatchAnything [4] can even match an image to a symbolic map, for example), but they are today implemented with very expensive transformer layers making them unacceptable for embedded platforms (the web demo of [4] requires 16s per pair of small images). Also, from a functional point of view, SIFT+lightglue and MatchAnything perform registration, not directly the final guidance task. In this sense, appearance-only SLAM like Fab-Map [1] is somehow related to this work. Yet, Fab-Map aims to detect already known areas (loop closing) while we map image appearance to target direction.

Let us point out that our idea of creating an implicit model of a scene from a stack of images is also related to Nerf (Neural Radiance Field) literature [9]. However, here we do not really model the scene but rather the appearance of the target and/or it surrounding at different scales/orientations...
To summarize, our work is inspired by Nerf but applied to guidance. It does not rely on transformer-based dense correlation to perform registration, thus being much faster than MatchAnything. Finally, compared to classical registration techniques whose current state of the art seems to be SIFT+lightglu, our pipeline does not depend on a specific reference image, offering robustness to common changes in the scene and/or specific noise like heat noise.

3. Direct Guidance Learning

The idea of our method is to use the ability given an image to generate the image which would have been seen from another camera position. Given A the parameter of the camera, and a reference image X in which the target position p is known, it is somehow easy and/or mathematically well understood to generate a pair $x, y = G(A, X, p)$ where x is the image corresponding at the appearance of the scene X under camera parameter A, and, y the target position relatively to image x. Precisely, the new image can be generated with a simple homography by assuming that the source image is flat which is acceptable for remote sensing application, even if, using ray tracing with topographic data would be

better. Now inversely, from the image x, it is not trivial to recover mathematically neither the camera parameter A nor the target position y i.e. the process of synthetizing an image G is well known but not inversible. Thus, the offered idea is to directly learn a mapping between x and y using a database generated though function G (mapping A instead of y can be done functionally but as this object is less smooth, it would probably require a larger deep network). So precisely, the idea is to sample a large set of parameters $A_1, \dots, A_R$ and to generate though G (and the stack of reference images) a database of images/target position $x_1, \dots, x_R$ and $y_1, \dots, y_R$, and then, to train the model to produce y_i given x_i. This idea is summarized in **Fig 2**.

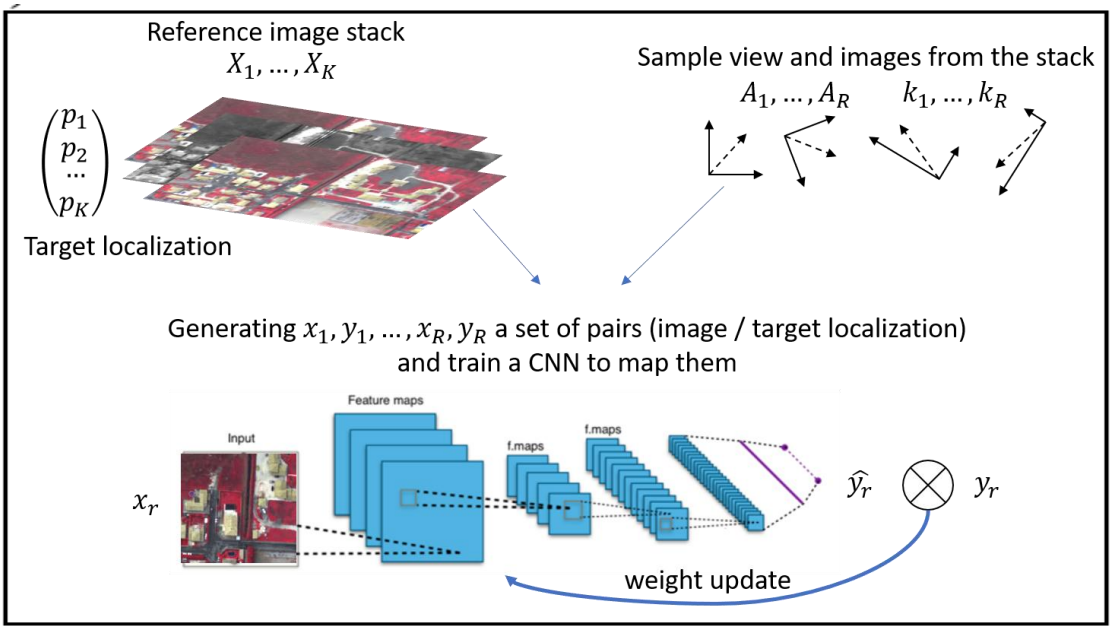

Fig 2. Overview of the offered framework: guidance is cast as the problem of learning the target localization y from an image x on a dataset of pairs (x, y) synthetized from reference image X where target position p is known under a random camera parameter A. At runtime, the network directly predicts a localization from the current image without needing the stack.

4. Experiments

4.1. Global experimental setting

The data source of our experiment is an IGN BD Ortho visible image (which looks like natural images) and some Copernicus Sentinel-2 (shorten in S2) images both visible (bands 2,3,4), or IR (band 8).

We consider 2 sets of experiments. First, we aim to measure performances of our algorithm from computer vision point of view, typically under large diversity of camera positions. For these experiments, we sample a random rotation (around the vertical), a random zoom, and a small uniform rotation around the other two axes to generate each matrix A. Then, we also consider a more representative setting where the points of view are not sampled randomly but along trajectories of a hypersonic platform and with a S2 infrared image with low change but with specificities related to hypersonic platform like heat noise.

4.2. Implementation details

When relying on S2 image, a critical preprocessing is required to recover somehow 8bit images (or at least to avoid a too large range of pixel values). We rely on histogram is equalized at image level for this purpose. There exist many other pre-processing methods like local histogram normalization, TOA normalization, or contrast enhancement. Yet, none of these are universally better as each may have side effect: TOA requires additional information which may not be available during mission preparation, contrast enhancement technics can create spurious textures on flat areas and patch base normalization

may create normalization inconsistency. Thus, despite probably not optimal, simple global histogram normalization has the interest of being straightforward.

In these experiments, S2 images are mostly cloudless (but with large changes in snow), yet the algorithm is designed to resist low cloud cover (obviously, images with strong cloud cover should not be added to the stack of reference images).

The higher we are, the harder it is to be precise on target coordinates in the metric system because pixel resolution decreases. So, we normalize error in pixel coordinates: the task becomes selecting the pixel in which the target belongs (even if not visible when the platform is too high). We thus report both mean square error and number of samples for which the error is less than 10px to avoid the metric being biased by outliers.

For the deep network, we put emphasis on limiting the number of layers: all experiments are done with the first 4 blocks of ConvNext Tiny [6] followed by a task-specific dense layer allowing us to achieve 60FPS on CPU only at inference on 256x256 images. Currently, the same model with the first 4 blocks of EfficientNet B0 [12] has been tested but performs poorly in comparison to ConvNext Tiny despite both networks representing the state of the art of small convolutional deep networks. Probably, EfficientNet would have required more blocks to capture the problem, damaging running time. Let us point out that with this setting, the accuracy of the offered method cannot be higher than 5 pixels (it predicts an 8x8 pixel block containing the target). Yet, this is not really an issue are, we want absolute rather than accurate localization (this absolute localization could be filtered with accelerometer measurements which are locally accurate despite drifting).

Fine-tuning of the network pretrained on Imagenet is done in several steps: first, the head is aligned on the task; then, a first fine-tuning is performed with SGD and very small learning rate; finally, a classical fine-tuning is performed with advanced optimizer [2]. Let stress that pretraining weights are critically required despite there are designed for ImageNet i.e. mostly to recognize cat & dog from internet. Thus, those weights may pollute the networks with irrelevant features despite being mandatory as we align the network on the task on a very small dataset (corresponding to the slack of images of the scene). There is almost certainly a room for improvement by relying on a S2 data pretrained model.

The baseline considered for comparison is opencv2 SIFT registration on a single reference image with standard Lowe's ratio. Currently we also tested SIFT+lightGlue (pretrained) but it performs similarly as SIFT: this can be explained because first lightGlue is not trained for remote sensing image, and then, because the algorithm should not try to produce a very precise wrapping but rather to deal with very large appearance change, and, for this purpose, pretrained lightGlue descriptors were not more useful than SIFT ones.

4.3. Generic Experiments

This setting is split into 2 sub-experiments. One is with weak change as we consider a single large image from IGN BD Ortho to create views. Then, we consider a setting with strong change using 4 Sentinel2 images (from January 2025 to March 2025) where 2 images contain snow and 2 do not. The first two (1 snow, 1 no-snow) are used for training, and the other two for testing. Thus, in this test, the deep network does neither know the point of view nor the image (and so the fact that there will or will not be snow) and implicitly needs to use the correct reference images when associating current views with the internal model encoded in the network weights.

4.3.1: weak change

This first setting is mostly an experiment designed to ensure algorithms are functional where both methods (baseline and offered one) successfully manage to find efficiently the position of the target as reported in **Table 1**. SIFT baseline achieves even better precision than the offered method in terms of mean square error and both methods produce acceptable predictions in 96\% of the sampled images (most failures are related to images with strong oblique views which are somehow distorted by the absence of topographic data).

Table 1. Performance of target position estimation under weak physical change between reference image and current image

Method	Mse	Frame ratio with less than 10px error
Baseline	1.68px	96.4%
Our	6.58px	96.1%

Still, it is interesting to see that our method has been able to encode the visual appearance of the neighborhood of the target at any resolution as pointed in **Fig 3**.

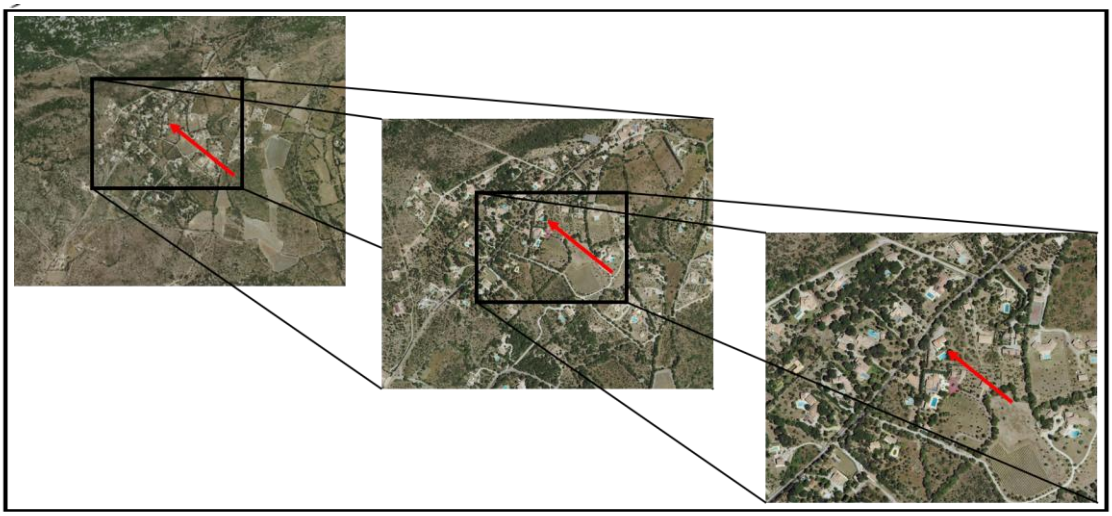

Fig 3. Illustration of the image of this first experiment. Hypothetical target (red arrow) is not really visible in first image, yet, the surrounding is sufficient to know where it is. This explains how our model is able to learn a mapping image-to-target at any resolution.

4.3.2: strong changes

For the second setting, 2 images (1 with snow, 1 without snow) are used for training the offered algorithm, and the same for testing. However, the baseline is restricted to selecting a single reference image, making it hard to register on the opposite test image. This leads to less than 24% of the test images being correctly processed (in many cases, SIFT matching does not even find 4 good matches for estimating the homography matrix). Inversely, the offered method manages to process correctly more than half of the images distributed across the two modes (snowy and non-snowy). Currently, performance of our method on training images is much higher, highlighting the fact that performance may increase significantly with a larger reference image stack (only two here).

These results, reported in **Table 2**, highlight the fact that relying on a single reference image is not a good idea when strong changes can arise between the reference image and the current one, while encoding the scene with our method on a stack of reference images can mitigate the issue. This is the main result of this paper: classical image matching technics can suffer under strong physical change of the scene. Besides, focusing mainly on geometrical accuracy makes low sense as sensor-based guidance is mostly relevant for providing absolute rather then accurate localization (which can be recovered by filtering with inertial sensor).

Table 2. Performance of target position estimation under strongly bimodal (snow vs no-snow) distribution of reference and testing images.

Method	Mse	Frame ratio with less than 10px error
Baseline	53.03px	23.6%
Our	**42.64px**	**51.3%**

In order to illustrate why SIFT performs poorly, **Fig 4** displays the same crop of two S2 one snowy, one normal (centered on a hypothetical target). One can see how the appearance are different even without any geometrical changes. On this already-registered pair, SIFT extracts around 700 points per image but manage to match only 20 of them. Adding only a little geometric deformation or sub-sampling frequently makes the number of matches going under 4.

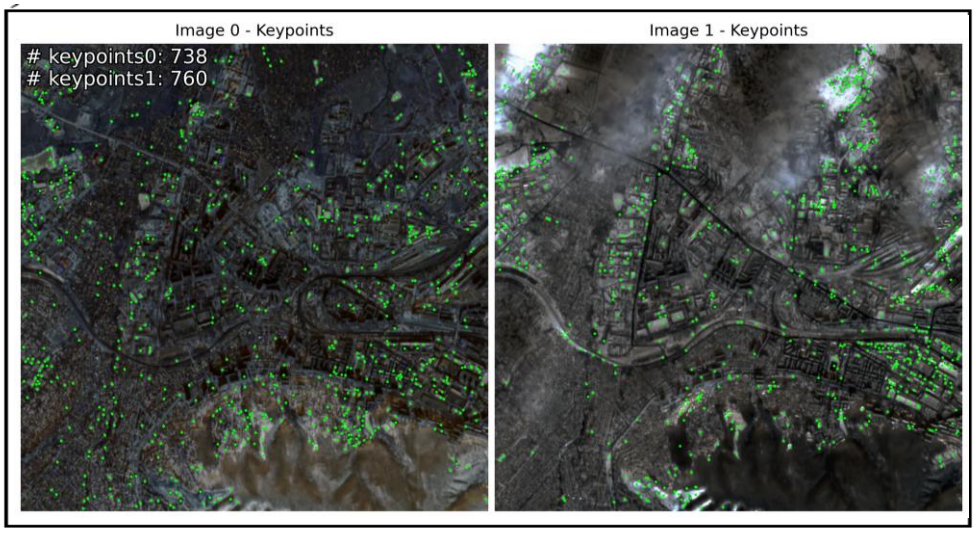

Fig 4. Illustration of SIFT failure: key point are not on the same location on this area captured at 2 different dates due to important physical change (snow) producing strong visual discrepancy (in particular under global histogram normalization).

4.4. Hypersonic platform specific Experiments

As the images seen along a trajectory of a hypersonic platform way exhibit some specificity (specific distribution, noise…), we also offer to evaluate performance not on individual images but on videos related to trajectories of the platform: instead of sampling views A_r randomly, we simulate trajectories of a camera in the head of a hypersonic platform performing somewhat representative maneuvers (speed/resolution may not be representative).

A generic model of a hypersonic platform with lifting has been used to generate a trajectory during re-entry associated with terminal guidance maneuvers. One common maneuver used that would also challenge our method in a significant way is to induce a spinning motion around its longitudinal axis to generate a helicoidal motion. A simplified version of this maneuver imposing a periodic spinning motion is implemented to generate a terminal trajectory a very high speeds in **Fig 5 a)**, where the position of the hypersonic platform is shown by the green square, and the field of view of the camera is the red quadrilateral on the ground, which is constrained by the maneuvers and the trajectory inclination.

Considering a fixed line of sight of the camera in the head of the hypersonic platform, it would therefore generate a stack of images, similar to **Fig 5 b)**, for a given time period of the spinning motion. To generate a sufficient number of stack images, each trajectory has the same target but with weak setting changes, such as the period of the helicoidal motion, initial position and velocities, and small random variation of the hypersonic attitude during the trajectory. Given the very high speed and the sampling rate of the camera, the resolution between each image can be very different, as seen in **Fig 5 b)** where the area of each quadrilateral gets smaller and smaller.

We simulate 100 trajectories (around the same scene/target under weak change setting like in *4.3.1* with an infrared S2 image), 90 for training and 10 for testing. All images from all training trajectories are used for training the network, like for other experiments: views are considered independent during training (but not testing) yet capturing the fact they belong to trajectories and not uniform sampling. **Fig 6** displays output of the algorithm along a testing trajectory.

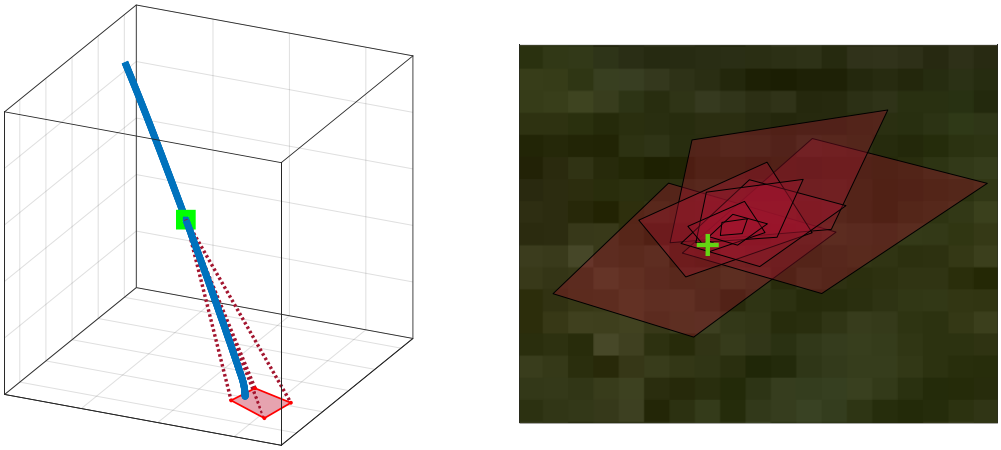

Fig 5. **a)** Re-entry trajectory of a generic hypersonic platform with a spinning motion. **b)** Examples of a stack of images seen by the camera during the periodic spinning maneuvers. Each red quadrilateral represents an image taken by the camera.

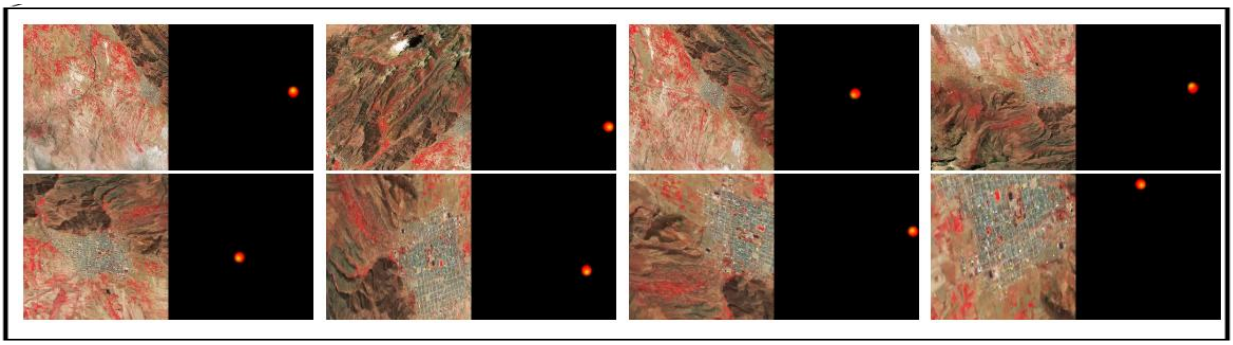

Fig 6. Outputs along a trajectory: all 8 images represent an image and an output mask (red dot is the location of the target, yellow is the pixel-wise predicted likelihood of being the target location). One can again notice the ground resolution difference between first and final image, yet the algorithm can coarsely predict the location of the target in all those situations.

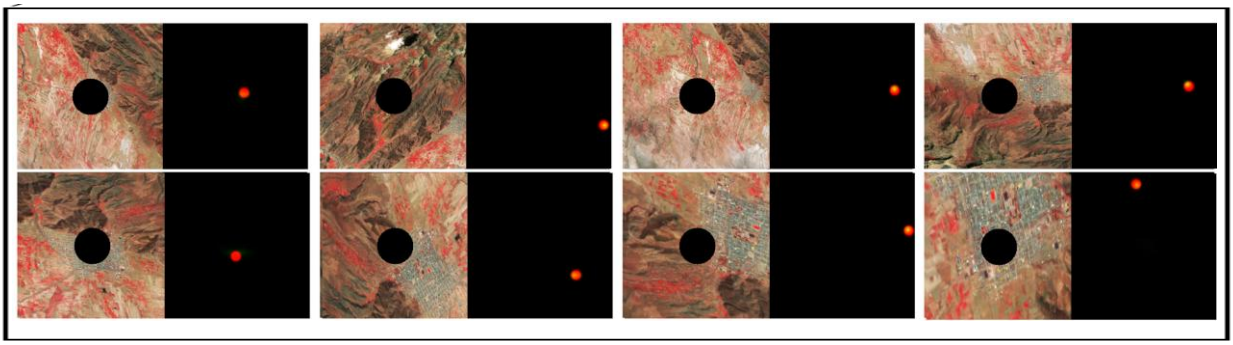

Fig 7. Same as *Fig 6* but with an occultation miming thermal protection of the optical windows.

Then, we perform the same experiment with 2 noises directly related to hypersonic platform: first a thermal protection and a heat noise. For the protection, we add (both in training and testing) a small black circle in the center of the image corresponding to the hypothetical thermal protection of the optical window (see **Fig 7**). Finally, we perform a third time this experiment with a strong simulated heat noise (increasing pixel value up to saturation of a circular area whose radius increase with time). Again, the algorithm achieves a good processing of the test set videos (see **Fig 8**). Let recall that the noise or the thermal protection should be added at train time to be correctly handled at test time (so the 3 figures 6-8 correspond to 3 different sets of weights for our CNN, yet achieving good result in

each respective cases). More quantitatively, predicted target is correct within 10px on at least 66% of the frames for all 10 testing trajectories of all three setting (despite number of frames correctly processed is logically higher for the clean images) in particular even despite the strong visual impact of the simulated heat noise.

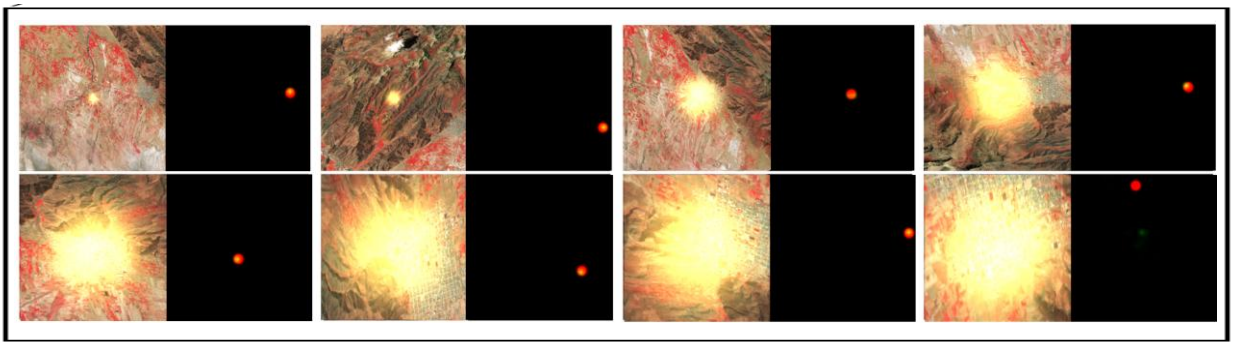

Fig 8. Same as *Fig 6* but with a simulated heat noise.

For comparison, we also perform those experiment (guidance under simulated heat noise) with the SIFT baseline, and surprisingly, it still manages to extract sufficient matching points between the degraded IR image and a single composite optical reference image, compiled from multiple Google images (see **Fig 9**). The red cross denotes the accurate position of the targeted point. Green lines illustrate the corresponding point pairs identified by SIFT, while the blue line represents the detected footprint of the IR image within the reference image, which appears non-square due to trajectory inclination. The blue point and green circle indicate the estimated location of the target within the IR image onboard the missile. As evident, SIFT matching significantly aids in target identification, even in cases of severe degradation. However, it is noteworthy that this performance is largely attributed to the highly structured environment depicted in the image, particularly the presence of nearby mountains and it relies heavily on a limited number of correspondences between the flight image and the reference image highlighting a potential instability on harder cases.

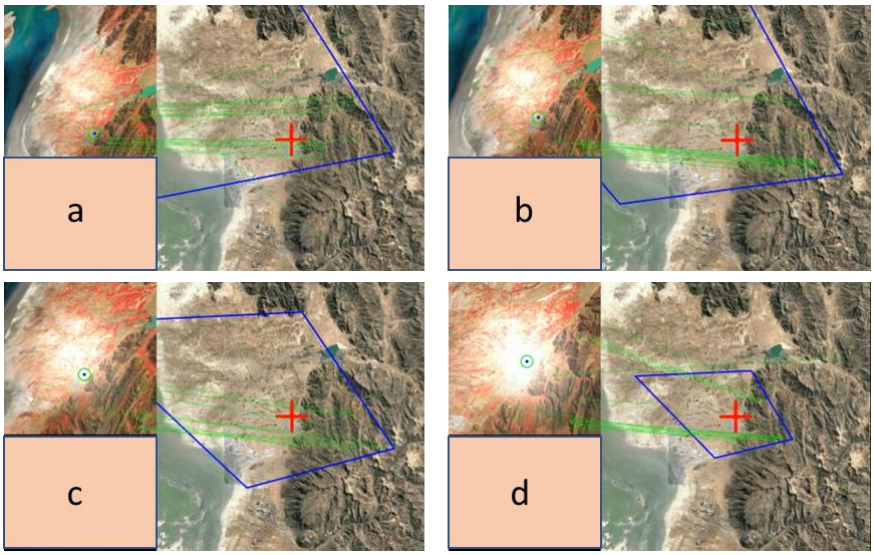

Fig 9. Some results of SIFT baseline under heat noise (reference built from Google Earth sources).

5. Conclusion

In this paper, we point out the limits of the registration-on-a-single-reference-image framework for sensor-based guidance and offer replacing it by directly learning a mapping between image and target localization using small deep convolutional networks on a stack of reference images.

Despite successes in these preliminary experiments, it is obvious that this framework has many critical drawbacks compared to the baseline. In particular, given the purpose of this algorithm, a simple statistical evaluation on a test set (and removal of well-understood geometric routines) may raise many questions. Further research will be needed to strengthening these results and evaluating at larger scale the relevancy of deep-learning-based guidance for such critical platforms.

We also acknowledge that thermal effect should be evaluated further in future works. Various aero-thermo-optical (ATO) effects (see [13]) have not been considered including: Optical index variations through the supersonic shock and within the shock layer, Turbulence-induced resolution degradation, Thermomechanical effects within the window, Self-emission of the window during heating…

Acknowledgment We thank ESA and IGN for publicly releasing S2 image and the BD Ortho. We also thank GoogleEarth from which only small crops has been used. NVIDIA Llama Nemotron has been used for improving English writing but all scientific material is generative AI free.

References

[1] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. he International journal of robotics research, 27(6):647–665, 2008

[2] Aaron Defazio, Xingyu Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. The road less scheduled. Advances in Neural Information Processing Systems, 37:9974–10007, 2024

[3] Johan Edstedt, Qiyu Sun, Georg Bökman, Marten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19790–19800, 2024

[4] Xingyi He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. arXiv preprint arXiv:2501.07556, 2025

[5] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17627–17638, 2023

[6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11976–11986, 2022

[7] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. Nature, 293(5828):133–135, 1981

[8] David G Lowe. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60:91–110, 2004

[9] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021

[10] Randall C Smith and Peter Cheeseman. On the representation and estimation of spatial uncertainty. The international journal of Robotics Research, 5(4):56–68, 1986

[11] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8922–8931, 2021

[12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019

[13] WANG HUI, SHOUQIAN CHEN, WANG ZHANG, FANYANG DANG, LINJU, XIANMEI XU and ZHIGANG FAN: Evaluating imaging quality of optical dome affect by aero-optical transmission effect and aero-thermal radiation effect. Optics Express, Vol 28, No5/ March 2020