UCLouvain KU LEUVEN

6D Pose Estimation of Uncooperative Spacecraft for Proximity Operations: a Neural Radiance Fields Perspective on Data-Driven Methods

Antoine Legrand

Thesis submitted in partial fulfillment
of the requirements for the joint degree of:
PhD in Engineering Sciences (UCLouvain)
&
Doctor of Engineering Science (PhD): Electrical Engineering (KU Leuven)

Dissertation committee:

Prof. Christophe De Vleeschouwer (Advisor)	UCLouvain, Belgium
Prof. Renaud Detry (Advisor)	KU Leuven, Belgium
Prof. Benoît Macq (Secretary, Public)	UCLouvain, Belgium
Prof. Dirk Vandepitte (Secretary, Prelim.)	KU Leuven, Belgium
Prof. Yang Gao	HKUST, Hong-Kong SAR
Dr. Jonathan Denies	Aerospacelab, Belgium
Prof. Laurent Francis (Chairman, Public)	UCLouvain, Belgium
Prof. Em. Jean Berlamont (Chairman, Prelim.)	KU Leuven, Belgium

Version of June 2026.

Non quia difficilia sunt non audemus, sed quia non audemus difficilia sunt.
"It is not because things are difficult that we do not dare,
but because we do not dare that they are difficult."
— Seneca

Abstract

On-orbit servicing and active debris removal missions depend on rendezvous and proximity operations in which a chaser spacecraft must maneuver in close proximity to a target or dock with it. Autonomous operation around an uncooperative target requires monocular vision-based estimation of the target's 6D pose, *i.e.*, position and orientation, relative to the chaser. Recent data-driven approaches show strong promise but suffer from several limitations that hamper their adoption. First, their scope of application is constrained to perfectly known targets because training hinges on an explicit geometry and appearance model such as a CAD model to produce a synthetic training set. Second, their dependence on synthetic data weakens out-of-domain generalization and creates a risk of performance degradation when models encounter real on-orbit imagery. Third, data-driven methods rely on opaque components, which lower transparency and affect confidence in safety-critical operations.

Motivated by these limitations, this thesis addresses three research questions. (i) How can pose estimation be achieved when no explicit target model is available? (ii) How can we move beyond image-wise transforms to strengthen domain generalization? (iii) How can we visualize the target features exploited by pose estimators to understand their decision process and increase confidence? This work adopts a unified perspective based on novel view synthesis through Neural Radiance Fields (NeRFs). Novel view synthesis implements the inverse mapping of pose estimation by rendering an image from a specified camera viewpoint rather than inferring pose from an image. Within this perspective, NeRFs (i) replace explicit target models for dataset creation, (ii) enable appearance-augmented yet geometry-consistent training, and (iii) allow gradient-based inspection of the 3D cues that support pose estimation.

This thesis advances vision-based spacecraft pose estimation through three core contributions. First, we introduce a NeRF-based 3D reconstruction method tailored to on-orbit imagery through additional degrees of freedom that compensate for illumination variability and pose uncertainty. Second, we develop an image synthesis method that exploits the decoupling of geometry and appearance in NeRFs to generate geometrically consistent images under diverse viewpoints and appearance conditions. This enables the learning of a pose estimation network without requiring access to a target CAD model and improves generalization beyond the training domain. Third, we introduce a visualization technique that learns a NeRF via gradients back-propagated through a pose estimator to reveal the target 3D cues on which the estimator relies. Together, these contributions (i) expand applicability to targets without explicit models, (ii) improve robustness under domain shift at deployment, and (iii) increase transparency to support adoption in future proximity operations.

Keywords: 6D Pose Estimation, Image Synthesis, Neural Radiance Fields, Novel View Synthesis, Out-Of-Domain Generalization, Spacecraft Pose Estimation

Résumé

Les missions de maintenance en orbite et de retrait actif de débris spatiaux reposent sur des opérations de proximité et rendez-vous spatiaux au cours desquelles un véhicule chasseur doit manœuvrer à proximité immédiate d'une cible ou s'y arrimer. L'approche autonome d'une cible non coopérative exige l'estimation monoculaire fondée sur la vision de la pose 6D de la cible, *i.e.*, position et orientation, relativement au chasseur. Les approches récentes apprises sur les données sont prometteuses mais souffrent de plusieurs limitations qui freinent leur adoption. Premièrement, leur champ d'application est limité aux cibles parfaitement connues, car l'entraînement repose sur un modèle explicite de géométrie et d'apparence, tel qu'un modèle CAO, pour produire un jeu d'apprentissage synthétique. Deuxièmement, leur dépendance aux données synthétiques affaiblit la généralisation hors domaine et crée un risque de dégradation des performances lorsque les modèles sont confrontés à des images réelles en orbite. Troisièmement, les méthodes apprises s'appuient sur des composants opaques, ce qui réduit la transparence et affecte la confiance dans des opérations critiques pour la sécurité.

Motivée par ces limitations, cette thèse considère trois questions de recherche. (i) Comment réaliser l'estimation de pose en l'absence de modèle explicite de la cible ? (ii) Comment aller au-delà des transformations au niveau de l'image afin de renforcer la généralisation entre domaines ? (iii) Comment visualiser les caractéristiques de la cible exploitées par les estimateurs de pose afin de comprendre leur processus de décision et d'accroître la confiance ? Ce travail adopte une perspective unifiée fondée sur la synthèse de vues nouvelles au moyen de Neural Radiance Fields (NeRFs). La synthèse de vues nouvelles met en œuvre l'application inverse de l'estimation de pose en rendant une image à partir d'un point de

vue caméra spécifié plutôt qu'en déduisant la pose à partir d'une image. Dans cette perspective, les NeRFs (i) remplacent les modèles explicites de la cible pour la création de jeux de données, (ii) permettent un entraînement augmenté en apparence tout en respectant la cohérence géométrique, et (iii) autorisent une inspection fondée sur les gradients des indices 3D qui soutiennent l'estimation de pose.

Cette thèse fait progresser l'estimation de pose de véhicules spatiaux fondée sur la vision au travers de trois contributions essentielles. Premièrement, nous introduisons une méthode de reconstruction 3D fondée sur les NeRFs, adaptée aux images en orbite grâce à des degrés de liberté supplémentaires qui compensent la variabilité de l'illumination et l'incertitude de pose. Deuxièmement, nous développons une méthode de synthèse d'images qui exploite le découplage géométrie-apparence des NeRFs pour générer des images géométriquement cohérentes sous des points de vue et des conditions d'apparence variés. Cela permet l'apprentissage d'un réseau d'estimation de pose sans nécessiter l'accès au modèle CAO de la cible et améliore la généralisation au-delà du domaine d'entraînement. Troisièmement, nous introduisons une technique de visualisation qui apprend un NeRF via des gradients rétro-propagés à travers un estimateur de pose afin de révéler les indices 3D de la cible sur lesquels l'estimateur s'appuie. Ensemble, ces contributions (i) élargissent l'applicabilité aux cibles dépourvues de modèles explicites, (ii) améliorent la robustesse face aux changements de domaine lors du déploiement, et (iii) accroissent la transparence afin de favoriser l'adoption lors de futures opérations de proximité.

Samenvatting

On-orbit onderhoud en actieve verwijdering van ruimtepuin zijn afhankelijk van rendezvous- en nabijheidsoperaties waarbij een achtervolgend ruimtevaartuig dicht bij een doelwit moet manoeuvreren of ermee moet koppelen. Een autonome werking rond een niet-coöperatief doelwit vereist een monoculaire, op visie gebaseerde schatting van de 6D-pose, positie en oriëntatie, van het doelwit ten opzichte van het achtervolgende ruimtevaartuig. Recente data-gedreven benaderingen zijn veelbelovend maar hebben verschillende beperkingen die dagelijks gebruik bemoeilijken. Ten eerste is hun toepassingsgebied beperkt tot perfect op voorhand gekende doelen, omdat de training berust op een expliciet geometrie- en textuurmodel, zoals een CAD-model, om een synthetische trainingsset te produceren. Ten tweede verzwakt de afhankelijkheid van synthetische data de veralgemening van de prestaties buiten het trainingsdomein en ontstaat er een risico op prestatie-achteruitgang wanneer modellen geconfronteerd worden met echte beelden in een baan om de aarde. Ten derde steunen data-gedreven methoden op ondoorzichtige componenten, wat de transparantie vermindert en het vertrouwen aantast bij veiligheid-kritische operaties.

Gemotiveerd door deze beperkingen behandelt dit proefschrift drie onderzoeksvragen. (i) Hoe kan poseschatting worden gerealiseerd wanneer er geen expliciet model van het doelwit beschikbaar is? (ii) Hoe kunnen we verder gaan dan transformaties op beeldniveau om de domeingeneralisatie te versterken? (iii) Hoe kunnen we de eigenschappen van het doelwit visualiseren die door poseschatters worden benut om hun beslissingsproces te begrijpen en het vertrouwen te vergroten? Dit werk hanteert een eenduidig perspectief gebaseerd op 'synthese van nieuwe aanzichten' via Neural Radiance Fields (NeRFs). Synthese van nieuwe aanzichten implementeert de omgekeerde transformatie van poseschatting door een beeld te renderen

vanuit een opgegeven camerapose in plaats van een pose uit een beeld af te leiden. Vanuit dit perspectief vervangen NeRFs (i) expliciete doelmodellen voor datasetcreatie, (ii) maken zij textuur-verrijkte maar geometrisch consistente training mogelijk, en (iii) laten zij gradiënt-gebaseerde inspectie toe van de 3D-eigenschappen die poseschatting ondersteunen.

Dit proefschrift brengt de op visie gebaseerde poseschatting van ruimtevaartuigen vooruit via drie kernbijdragen. Ten eerste introduceren wij een op NeRFs gebaseerde 3D-reconstructiemethode die is toegespitst op beelden genomen in een baan om de aarde, aan de hand van extra vrijheidsgraden die variabiliteit in belichting en pose-onzekerheid compenseren. Ten tweede ontwikkelen wij een beeldsynthesemethode die het ontkoppelen van geometrie en uiterlijk in NeRFs benut om geometrisch consistente beelden te genereren onder uiteenlopende gezichtspunten en textuurcondities. Dit maakt het mogelijk een poseschatnetwerk te trainen zonder toegang tot een CAD-model van het doel en verbetert de generalisatie buiten het trainingsdomein. Ten derde introduceren wij een visualisatietechniek die een NeRF aanleert, via gradiënten die terug-gepropageerd worden door een poseschatter, om de 3D-eigenschappen van het doelwit te onthullen waarop de schatter vertrouwt. Gezamenlijk vergroten deze bijdragen (i) de toepasbaarheid op doelen zonder expliciete modellen, (ii) verbeteren zij de robuustheid bij domeinverschuiving tijdens de missie, en (iii) verhogen zij de transparantie ter ondersteuning van gebruik bij toekomstige nabijheidsoperaties.

Remerciements

Ce manuscrit tourne une page, celle d'une décennie passée à l'Université Catholique de Louvain. Alors que je n'étais qu'un jeune étudiant (encore innocent) sur le point de passer l'examen d'admission, je ne me serais certainement pas imaginé défendre une thèse de doctorat en intelligence artificielle, encore moins dans le cadre d'une cotutelle avec la KU Leuven. Quant à la mener au sein d'une entreprise aussi improbable qu'Aerospacelab, n'en parlons pas. Et pourtant… Le chemin fut long, semé d'embûches (ou de Bush, je ne suis plus sûr). Mais, telle *Titine* continuant à tailler la route, lentement mais avec style (enfin, un style), en bravant aussi bien les affres du temps (« *c'est de la rouille porteuse, t'inquiète* ») que certains dogmes, cette thèse a donc continué sa route jusqu'à son terme. Avant d'en finir, il me faut remercier toutes celles et ceux qui, par leur soutien, ont largement contribué à donner à cette aventure une saveur bien particulière.

En premier lieu, j'adresse mes remerciements à ma famille, pour son soutien et pour les valeurs qu'elle m'a transmises. Je pense aussi à mes amis qui, souvent sans s'en rendre compte, ont contribué à cette thèse et dont l'appui a été essentiel.

Une part importante de ces années, je l'ai partagée avec un groupe formidable composé d'*Anciens* (*Alexandre*, *Anne-So*, *Niels*, *Abolfazl*, *Gabriel*, *Maxime*, *Vladimir*, *Baptiste*, *Victor*, *Antoine*, *Jean*, *Florine*, *Benoit*, *Gilles* et *Mathieu*) et de *petits jeunes* (*Sarra*, *Elias*, *Ali*, *Thomas*, *Hippolyte*), ainsi que d'un membre honoraire (*Dany*). Au-delà des discussions techniques et des séminaires qui ont nourri ce travail, la douce folie du groupe l'a rendu bien plus léger. Entre les *Vendrediplodocus*, les *CCBs* et les pérégrinations d'un certain panda, je ne regrette rien : ce groupe était vraiment formidable. Plus largement, je suis reconnaissant à l'UCLouvain et à la KU Leuven pour l'environnement de recherche dont j'ai pu bénéficier. Je remercie en

particulier le groupe de *Benoit* (*Karim*, *Dani*, *Maxime*, *Tiffanie*, *Gauthier* et *Benoît*), ainsi que le personnel administratif (*Isabelle*, *Nathalie*, *Patricia* et *Anne*), sans oublier *John* et *Laurent* pour ces cinq dernières années passées en ICTEAM.

J'ai eu la chance de réaliser cette thèse au sein d'Aerospacelab. Je tiens à remercier l'ensemble des collègues avec lesquels j'ai pu collaborer, et plus spécialement *Jonathan*, *Mikko* et *Benoît*. La supervision assurée par *Jonathan* et *Mikko*, ainsi que leurs conseils, techniques autant qu'humains, ont fortement contribué à mon évolution. Je suis également reconnaissant à *Jonathan* pour la relecture minutieuse de cette thèse et pour sa participation au jury. Enfin, un merci particulier à *Benoît* pour avoir cru en ce projet un peu fou et en cette collaboration avec le monde académique, et surtout pour avoir fondé cette perle rare de l'industrie européenne qu'est Aerospacelab. J'ai hâte de voir ce que l'avenir nous réserve.

Cette thèse est le fruit d'une collaboration enrichissante mais encore trop peu commune. À ce titre, je remercie le SPW Économie, Emploi, Recherche pour son soutien, et tout particulièrement les docteurs *Fabian Lapierre* et *Fleur Roland* pour le suivi du projet et leur engagement en faveur des partenariats entre universités et entreprises.

Je remercie les membres de mon jury pour leur apport à cette thèse. Aux professeurs *Laurent Francis* et *Jean Berlamont*, j'adresse ma reconnaissance pour avoir assuré la présidence des défenses et pour leur aide face à quelques péripéties administratives. Je suis également reconnaissant à la professeure *Yang Gao* pour sa participation, pour ses remarques ayant clarifié plusieurs points du manuscrit, ainsi que pour nos échanges lors des deux dernières éditions d'ISpaRo. Je tiens aussi à exprimer ma gratitude au professeur *Dirk Vandepitte* pour avoir accepté de participer à ce jury sur un sujet éloigné de son champ habituel de recherche, et pour ses commentaires particulièrement détaillés. Enfin, je souhaite remercier le professeur *Benoit Macq* pour sa relecture, pour m'avoir initié à la recherche lors de mon mémoire, et pour la bonne humeur partagée en ELEN.

Je remercie enfin mes promoteurs *Christophe* et *Renaud* pour le suivi technique, leur contribution à la rédaction de nos articles et leur soutien constant. Une gratitude particulière va à *Christophe* dont l'implication, la disponibilité et l'optimisme, alliés à un pragmatisme et une franchise rares, ont été décisifs pour tenir sur la durée. Les mots ne suffisent pas pour exprimer à quel point il a contribué à faire de moi un meilleur ingénieur, un meilleur chercheur, et, peut-être aussi, une meilleure personne.

Bref, merci à vous tous et bonne lecture!

Table of Contents

I Foundations

List of Figures

List of Tables

List of Acronyms

Space Operations

ADR	Active Debris Removal	2
ESA	European Space Agency	52
GNC	Guidance Navigation & Control	4
IOD	In-Orbit Demonstration	33
LEO	Low Earth Orbit	4
MLI	Multi-Layer Insulation	23
RPO	Rendezvous & Proximity Operation	2
OOS	On-Orbit Servicing	2

Computer Vision & Neural Networks

3D-GS	3D Gaussian Splatting	131
CNN	Convolutional Neural Network	18
MLP	MultiLayer Perceptron	47
MTL	Multi-Task Learning	24
NeRF	Neural Radiance Field	7
NVS	Novel View Synthesis	10
ODR	Online Domain Refinement	37
OOD	Out-Of-Domain	5
PnP	Perspective-n-Point	19
SLAM	Simultaneous Localization and Mapping	67
ViT	Vision Transformer	18

Computing Platforms

Image Quality Metrics

Miscellaneous

Author's publication list

This thesis is derived from the publications listed below, which together form the basis of the dissertation. Some of these publications have been integrated into the manuscript, while others, although part of the project or of my training, are not discussed in details.

Core Papers

The following papers form the core of this thesis. Chapters 5 to 7 are therefore directly adapted from these papers.

- [1] **A. Legrand**, R. Detry, & C. De Vleeschouwer (2026).
  NeRF-based Spacecraft Reconstruction from Monocular Imagery Under Illumination Variability and Pose Uncertainty.
  (under review).
- [2] **A. Legrand**, R. Detry, & C. De Vleeschouwer (2026),
  CAD-Free Learning of Spacecraft Pose Estimators via NeRF-Based Augmentation.
  (under review).
- [3] **A. Legrand**, R. Detry, & C. De Vleeschouwer (2025),
  NeRF-based Visualization of 3D Cues Supporting Data-Driven Spacecraft Pose Estimation.
  in Proceedings of the 2nd International Conference on Space Robotics (IEEE ISpaRo25), Dec. 2025.

Related Papers

The following papers address spacecraft pose estimation. They consist in preliminary works that motivated the contributions brought by the three core papers. The two first were integrated in Chapter 2 while the two later have been extended to form the second core paper (see Chapter 6).

- [4] **A. Legrand**, R. Detry & C. De Vleeschouwer (2022)
  End-to-end neural estimation of spacecraft pose with intermediate detection of keypoints,
  in Proceedings of the IEEE/CVF European Conference on Computer Vision Workshops(ECCVW), AI4Space Workshop, Oct. 2022.
  (Best Presentation Award).

- [5] **A. Legrand**, R. Detry & C. De Vleeschouwer (2025).
  Domain generalization for in-orbit 6d pose estimation,
  in AIAA Journal of Aerospace Information Systems, Sept. 2025.

- [6] **A. Legrand**, R. Detry & C. De Vleeschouwer (2024). *Leveraging neural radiance fields for pose estimation of an unknown space object during proximity operations*,
  in IEEE International Conference on Space Robotics 2024 (ISpaRo24), Workshop on Advances in Orbital Robotics: In Orbit Manipulation, Servicing, and Assembly, available at *https://arxiv.org/pdf/2405.12728*, Jun. 2024.
  (Best Poster Award).

- [7] **A. Legrand**, R. Detry & C. De Vleeschouwer (2024).
  Domain generalization for 6D pose estimation through NeRF-based image synthesis,
  available at *https://arxiv.org/pdf/2407.10762*, Jul. 2024

Unrelated Paper

The following paper was written as part of the UCLouvain doctoral formation. It consists in an extension of my Master Thesis.

- [8] **A. Legrand**, B. Macq & C. De Vleeschouwer (2022).
  Forward error correction applied to JPEG-XS codestreams,
  in Proceedings of the 29th International Conference on Image Processing (IEEE ICIP22), Oct. 2022.

Statement on the Use of Generative AI

This thesis was partially written with the help of generative AI assistance tools (MS Copilot).

- Generative AI was used for mere language assistance and reformulation purposes.
- Generative AI was used for generating programming code during the last phase of this thesis. The generated code corresponds to functions or scripts that are simple but time-consuming to implement. GenAI programming mainly consisted in speeding up data management (changing image format, image resolution, dataset structure, label format conversion or other minor data issues), generate SLURM scripts to manage the numerous experiments launched during the thesis, extracting key data from log files and make visually appealing graphs. The coding of the core parts of this thesis is my own, built on top of the *K*-Planes and SPNv2 open-source implementations.
- Generative AI was used to generate visuals. Most figures were generated using `TikZ`. As writing `TikZ` code is often cumbersome, MS Copilot was used to generate the code of those figures given an exact definition of my expectations.

The content of this thesis is therefore my own (unless otherwise specified) and generative AI has only been used in accordance with the UCLouvain and KU Leuven guidelines and appropriate references have been added. I have reviewed and edited the content as needed and I take full responsibility for the content of the thesis.

1

Introduction

SATELLITES are supporting many applications affecting our whole society. Earth observation satellites provide data for weather [9] and climate monitoring [10], agricultural optimization [11], disaster management [12, 13] and contribute to our security [14]. Satellite communications provide globally accessible connectivity that strengthens our economy and security through resilient data links [15]. Global Navigation Satellite Systems (GNSSs) enable modern aerial, naval and terrestrial navigation systems [16]. As components of our daily life, economy, and defense, satellites have become critical assets with a broad societal impact. The increasing demand for satellite-based services has challenged traditional development models and enabled the emergence of the *New Space* sector. Through vertical integration, reduced development and manufacturing costs, shorter design cycles, and lower launch prices, *New Space* actors are deploying large, cost-effective constellations that drive an exponential growth in the number of satellites orbiting Earth.

1.1 Motivation

The sharp rise in orbital traffic has made collision avoidance and debris mitigation central challenges for space operations. In-orbit collisions have already caused satellite inoperability (*Cerise* [17]), damage to a crewed spacecraft (*Shenzhou-20* [18]), satellite destruction and formation of a large

amount of space debris (*Iridium 33-Cosmos 2251* [19]). Debris has also been produced by anti-satellite (ASAT) tests [20, 21]. This growing debris population increases the risk of the Kessler cascading collision syndrome [22].

Although debris-mitigation regulations exist [23–25], they cannot prevent collisions involving end-of-life satellites or other non-maneuverable debris. Consequently, On-Orbit Servicing (OOS) [26, 27] and Active Debris Removal (ADR) [28, 29] missions are receiving increasing attention to reduce debris-related risks. OOS aims to repair, refuel or inspect satellites to extend their operational life [30, 31], while ADR focuses on debris capture and removal [32, 33]. Both rely on Rendezvous & Proximity Operations (RPOs), in which a chaser approaches and potentially docks with a target. These operations have traditionally been tele-operated, but limitations such as communication latency, limited bandwidth, ground-station coverage, and human reliability motivate a shift toward autonomy.

A key component of autonomous close-proximity operations is the perception system, whose primary function is to estimate the target's relative pose, that is, its position and orientation with respect to the chaser [34]. The complexity of this estimation depends on whether the rendezvous is *cooperative* or *uncooperative*. In cooperative scenarios, the target includes aids such as fiducial markers or communication links, whereas in uncooperative cases the chaser must rely solely on on-board sensing.

Multiple sensing modalities exist [35], including Light Detection and Ranging (LiDAR) systems [36, 37], infrared sensors [38, 39], Time-of-Flight (ToF) sensors [40], event-based cameras [41], and stereo cameras [42]. Among these, monocular cameras remain the most practical option thanks to their low cost, mass, volume, and power consumption [43–45], despite the intrinsic difficulty of estimating relative pose from a single image. Data-driven methods have been proposed to exploit appearance-based geometric target features, but current approaches still face fundamental limitations that hinder their operational adoption. This thesis addresses these limitations and facilitate their operational use.

1.2 Problem Statement

Figure 1.1 illustrates the problem of estimating the pose of an uncooperative spacecraft from monocular images. A camera mounted on a chaser spacecraft provides a stream of images acquired at approximately 10 frames per second while approaching the target at close-range, *e.g.*, at less

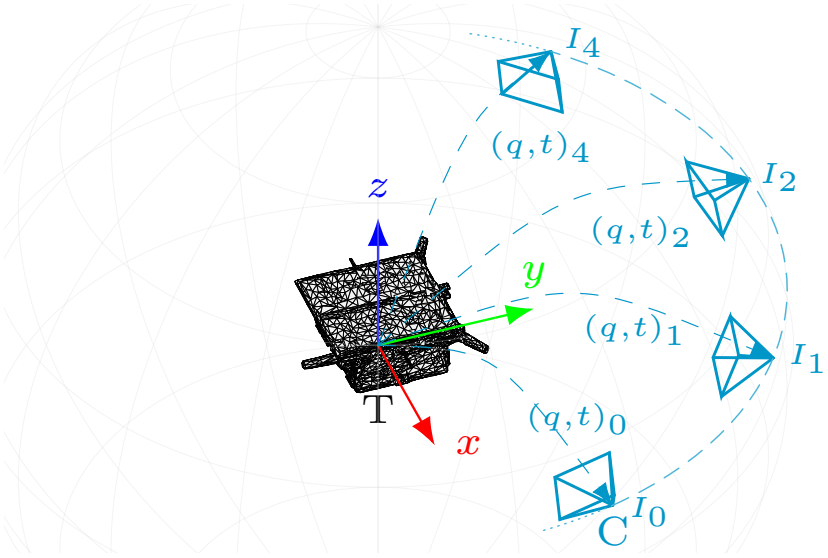


Fig. 1.1 Vision-based uncooperative spacecraft pose estimation: problem overview. During a close-proximity operation, a camera mounted on a chaser spacecraft C takes a stream of monocular pictures of the target T. At each timestep i the chaser must estimate the relative pose, *i.e.*, orientation q_i and position t_i based on the corresponding image I_i.

than a few hundred meters. The target is assumed to be **known** [1], *i.e.*, its Computer-Aided Design (CAD) model is available, **but uncooperative**, *i.e.*, it was not designed to support autonomous RPOs through fiducial markers, or inter-satellite communication capabilities. This uncooperative scenario is the most frequent in both ADR and OOS missions. In the former case, the debris cannot communicate with the chaser, and the presence of fiducial markers on the target cannot be assumed. In the latter case, although some satellites are designed to support RPOs, most of them are not built with fiducial markers nor inter-satellite communication capabilities.

[1] This classification between known and unknown targets in the literature suggests a binary choice; either the target can be perfectly described or no information is available prior to the mission. However, this level of knowledge of the target lies on a spectrum. Indeed, for many uncooperative targets such as space debris or commercial satellites, some information might be available prior to launch. For example, the principal dimensions and overall shape might be inferred from publicly available data, pictures of the target taken before its launch, or images captured by ground-based telescopes. Throughout this thesis, we adopt the less restrictive definition and therefore consider that any spacecraft is unknown as soon as its CAD model is not available.

Our goal is to estimate the 6-Degrees-of-Freedom (DoF) pose (q, t), consisting of the orientation quaternion q and the position t, that aligns the target frame $\{\mathcal{T}\}$ with the camera frame $\{\mathcal{C}\}$ attached to the chaser. Although a sequence $\{I_0, \ldots, I_a\}$ is acquired during the maneuver, the pose is computed *independently at each instant* using only the corresponding image. This thesis focuses exclusively on the navigation component, *i.e.*, relative pose estimation, of the Guidance Navigation & Control (GNC) pipeline required to conduct autonomous RPOs. Guidance and control subsystems, as well as the temporal fusion of the pose estimates through a navigation filter which would refine the single-frame pose estimates by estimating the evolution of chaser-target state over time, are beyond the scope of this thesis.

Image-wise pose estimation is challenging due to the visual and computational constraints encountered in Low Earth Orbit (LEO). Spaceborne imagery is affected by harsh illumination conditions that produce extreme brightness contrasts, frequently causing over- or under-exposure. These effects evolve with the spacecraft's relative motion and are amplified by strong specular reflections from spacecraft surfaces. Hence, the images are often degraded so that single-image pose estimation is inherently difficult. Despite the task complexity, the algorithm must operate in real time on space-grade hardware, which imposes strict constraints on its computational cost.

1.3 State-of-the-art Limitations & Research Questions

Dataset Generation

State-of-the-art spacecraft pose estimation solutions are based on data-driven methods [46–49]. They rely on neural networks trained on a large set of images depicting the target. Due to the complexity of acquiring a large number of images of the target before the mission, this training set consists in synthetic images generated through the target CAD model. This synthetic dataset generation therefore implies that spacecraft pose estimation methods implicitly assume access to a perfectly accurate CAD model of the target.

State-of-the-art limitation #1

Existing data-driven pose estimation techniques require an accurate prior model of a spacecraft's geometry and appearance, such as a CAD model, limiting their applicability when such a model is not available.

While some proximity operations may target perfectly known targets such as companies or space agencies performing servicing of their own satellites, many operations aim at (partially) unknown targets, *i.e.*, targets for which geometry and appearance information is not available. For example, a company could service a satellite whose owner is reluctant to give access to its CAD model or perform the removal of debris for which the shape and appearance may have evolved, *e.g.*, due to the effect of radiation on the painting or a collision with space debris. This fundamental limitation of spacecraft pose estimation methods raises the following research question:

Research Question #1

How can spacecraft pose estimation be enabled when an explicit prior model of the target's geometry and appearance cannot be assumed?

Domain Generalization

Once the neural network has been trained on-ground using synthetic images, it is deployed in orbit on the chaser payload and exposed to real on-orbit imagery. Due to harsh illumination conditions, significant discrepancies arise between synthetic and real images [45]. This domain discrepancy, known as the domain gap, often leads to a drop in performance when the model is applied in orbit.

This domain gap can be mitigated through Out-Of-Domain (OOD) generalization techniques which aim at improving the accuracy of a model on an image domain that is different from the training domain. Most existing works rely on image-wise augmentations that apply 2D transformations independently to each sample in the training set [49–51]. Such augmentations can modify appearance through modifications of the pixel values but they remain confined to operations within the 2D grid of the existing images. By acting only on pixel-level content, image-wise augmentations cannot modify the 3D arrangement of the scene or generate novel viewing conditions.

Their influence is therefore limited to local appearance adjustments, which restricts their capacity to capture the full variability encountered in real on-orbit imagery. Consequently, the diversity they introduce is fundamentally limited by the views already present in the original dataset.

State-of-the-art limitation #2

Image-wise augmentation methods operate strictly within the 2D image grid and therefore cannot generate new viewpoints, geometric configurations, or illumination conditions absent from the original dataset. Their inability to modify the underlying 3D scene structure intrinsically limits the diversity achievable during training and restricts improvements in Out-Of-Domain (OOD) generalization.

Going beyond image-wise augmentations through dataset level augmentation could further improve the OOD generalization capabilities of spacecraft pose estimation networks as it would further increase the training set diversity. This reliance on image-wise augmentation techniques to improve the OOD generalization abilities therefore raises the following research question.

Research Question #2

How can dataset-level information be aggregated to enable augmentations whose diversity surpasses the limits of image-wise transformations and improve Out-Of-Domain generalization?

Lack of transparency

While data-driven spacecraft pose estimation methods outperform previous methods based on hand-crafted features [46], this comes at the cost of a lack of explainability of their prediction. Neural networks are indeed perceived as "black boxes" whose decision process is opaque and deeply entangled in the network's weights. This lack of explainability slows down the adoption of data-driven methods in real missions. Indeed, stakeholders are unlikely to rely on systems they do not understand and for which no strong guarantee on their behavior can be provided.

State-of-the-art limitation #3

Data-driven spacecraft pose estimation models lack transparency in their internal decision mechanisms, limiting their suitability for safety-critical missions where trust and interpretability are essential.

Building that trust would require a deep understanding of how a network processes the input data to carry out its estimation. A first step in that direction is to analyze the visual cues that are exploited by the network to estimate the relative pose. This motivates the third and final research question addressed in this thesis:

Research Question #3

How can we visualize the spacecraft features that a pose estimation network primarily relies on, in order to gain insights into its decision process and increase confidence in its predictions?

1.4 Research Vision & Contributions

Our three research questions are addressed through a unifying perspective grounded in implicit scene representations such as Neural Radiance Fields (NeRFs) [52]. As depicted in Figure 1.2, we formalize the 6D pose estimation of an uncooperative spacecraft as a function $F_\Theta : \mathcal{I} \to \mathrm{SE}(3)$, where $\mathcal{I}$ denotes the image space and $\mathrm{SE}(3)$ denotes the special Euclidean group of three-dimensional rigid motions, *i.e.*, the pose space. We introduce the backward function $M_\Phi : \mathrm{SE}(3) \to \mathcal{I}$ which maps poses back to images. This function, implemented as an implicit scene representation, encodes target geometry and appearance without relying on explicit meshes or textures. The relationship between F_Θ and M_Φ provides the conceptual foundation for addressing the research questions through the following contributions, which are illustrated in Figure 1.3.

Dataset generation without relying on an explicit model

Spacecraft pose estimation methods require an explicit model, such as a CAD model, of the target to generate the large training set used to train the estimation network F_Θ. Our first contribution mitigates this dependency

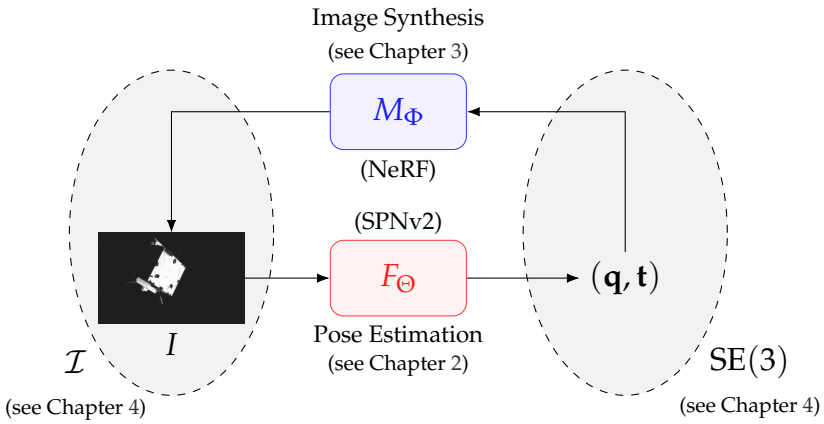


Fig. 1.2 Conceptual relationship between pose estimation (F_Θ) and image synthesis (M_Φ). Given an image $I \in \mathcal{I}$, the mapping F_Θ estimates a pose $(q, t) \in \mathrm{SE}(3)$. The mapping M_Φ implements the backward relation, and generates an image $I \in \mathcal{I}$ given a pose label $(q, t) \in \mathrm{SE}(3)$. This thesis addresses the limitations of F_Θ by exploiting its backward mapping M_Φ. The first part of the thesis therefore introduces the foundations of spacecraft pose estimation (Chapter 2) and novel view synthesis (Chapter 3), and describes the datasets, *i.e.*, image–pose pairs, as well as the evaluation metrics (Chapter 4) that define the image space $\mathcal{I}$ and the pose space $\mathrm{SE}(3)$.

by leveraging an implicit scene representation M_Φ that captures both the target's geometry and appearance from a small set of images depicting the spacecraft. We design an implicit representation tailored to in-orbit imagery, resilient to harsh illumination conditions and to uncertain pose annotations. Once trained, this model enables the synthesis of an arbitrary large dataset for training F_Θ, even in the absence of an explicit target model.

Contribution #1

This thesis develops a method to infer target geometry and appearance from on-board imagery, enabling the training of pose estimation models even when explicit prior models are not available.

Overcoming limitations of image-wise augmentations

The image generation process of the implicit target representation M_Φ learned from a set of images sampled from a limited training set can be randomized. This allows not only the generation of images from novel

viewpoints but also the synthesis of images for which the target appearance is randomized. Such dataset-level augmentation introduces variability that exceeds what image-wise augmentations can provide, ultimately improving the OOD generalization of the pose estimator F_Θ.

Contribution #2

This thesis presents a dataset-level augmentation method that produces more diverse samples than image-wise augmentations, both in terms of viewpoint variation and appearance diversity.

Visualizing target features exploited by the estimator F_Θ

Although an implicit scene representation M_Φ should normally be learned on images, we demonstrate that M_Φ can be trained using pose-guided gradients back-propagated through a frozen pose estimation network F_Θ. This supervision forces the implicit model to encode the visual cues that F_Θ relies upon to minimize its pose-prediction loss.

Contribution #3

This thesis proposes a visualization approach that identifies and highlights the target visual cues exploited by a pose estimation network during pose prediction, thereby enhancing model transparency.

1.5 Thesis Outline

This thesis is organized into three parts.

Part I: Foundations.

As the thesis is built around the relationship between pose estimation and novel view synthesis illustrated in Figure 1.2, this first part presents the background on these tasks, along with the datasets and metrics used for validation.

- **Chapter 2** surveys spacecraft pose estimation, formalizes the estimator F_Θ, and identifies key limitations (CAD model dependence, domain gap, and limited transparency). It also describes SPNv2 [49], the pose estimation network used across the thesis.

- **Chapter 3** reviews implicit scene representations with a focus on NeRFs, formalizes the backward function M_Φ, and details its implementation with K-Planes [53], which supports our three contributions.

- **Chapter 4** describes the datasets that are used to validate our proposed contributions along with the pose estimation and image quality metrics.

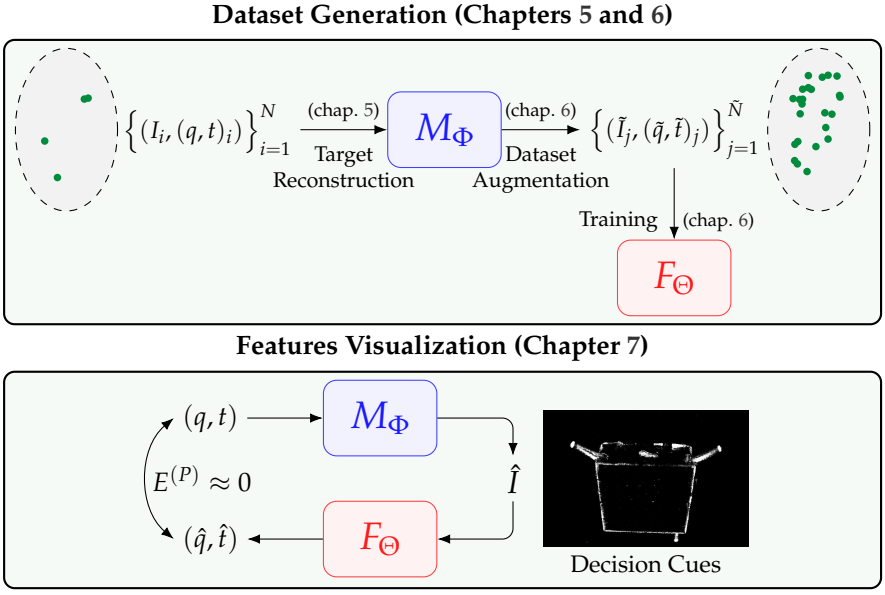


Fig. 1.3 Overview of our contributions under a F_Θ-M_Φ perspective. This thesis exploits a Novel View Synthesis (NVS) tool M_Φ to address the limitations encountered by a pose estimation network F_Θ. **(Top)**: Our first contribution, presented in Chapter 5, consists in a method that enables the learning of a NVS representation M_Φ from a limited annotated image set. This representation M_Φ can then be used to generate an augmented image set that provides better coverage of the image space. Hence, our second contribution is an augmentation method which is described in Chapter 6. By training a pose estimator F_Θ on that augmented set, its OOD generalization capabilities are improved. **(Bottom)**: By training M_Φ using the gradients back-propagated through the pose estimation network F_Θ, M_Φ learns to represent the target features that are primarily exploited by F_Θ. This third contribution, developed in Chapter 7, provides some insights on the decision process of spacecraft pose estimation networks.

Part II: Contributions.

This second part explores how NVS techniques can be used to improve spacecraft pose estimation networks. In this part, we present three methods that use the implicit model M_Φ to (i) reduce the reliance on CAD models, (ii) enhance out-of-distribution generalization, and (iii) increase the interpretability of F_Θ. These 3 contributions are summarized in Figure 1.3.

- **Chapter 5** describes our NeRF-based method for implicitly modeling both the target geometry and appearance from on-orbit imagery. The method is tailored to be resilient to harsh illumination conditions and uncertain pose annotations through additional degrees of freedom that are granted to the original NeRF architecture.
- **Chapter 6** introduces a method that leverages M_Φ for dataset-level augmentation, diversifying viewpoints and appearances beyond image-wise transforms. The augmented data expands the training distribution of F_Θ and improves its OOD generalization.
- **Chapter 7** presents a visualization method based on an implicit novel view representation M_Φ learned using gradients backpropagated through a frozen pose estimator F_Θ. The implicit representation can then be used to reveal the target features that are the most influential in the network's pose-prediction process.

Part III: Perspectives and Conclusions.

This final part looks ahead to future research and synthesizes the thesis' outcomes.

- **Chapter 8** concludes the thesis by summarizing the main answers given to the research questions and reaffirming how this thesis contributed to the field. It highlights the limitation and underlying assumptions of this work and discusses avenues for research in spacecraft pose estimation.

PART I
Foundations

2

Data-driven Pose Estimation of Uncooperative Spacecraft

Spacecraft pose estimation algorithms currently rely on data-driven methods. This chapter therefore provides background on data-driven spacecraft pose estimation. First, Section 2.1 formulates the problem of estimating the pose of an uncooperative spacecraft from monocular imagery and describes how data-driven techniques address this task. Sections 2.2 to 2.4 then outline the key challenges inherent to data-driven approaches and summarize how current methods tackle them: the design of neural network architectures, the training strategy required for robust performance, and considerations related to on-board deployment on an actual mission. These aspects are partly addressed by SPNv2 [49], which serves as the baseline pose estimation method in this thesis and is therefore presented in Section 2.5.

2.1 Problem Statement

As illustrated in Figure 2.1, this thesis focuses on estimating the 6D pose, *i.e.*, position t and orientation q, of a target spacecraft relative to a monocular camera mounted on a chaser spacecraft performing close-range operations [46]. The target spacecraft is assumed to be **known** (its CAD model is available) **but uncooperative**, *i.e.*, it does not contribute to the rendezvous

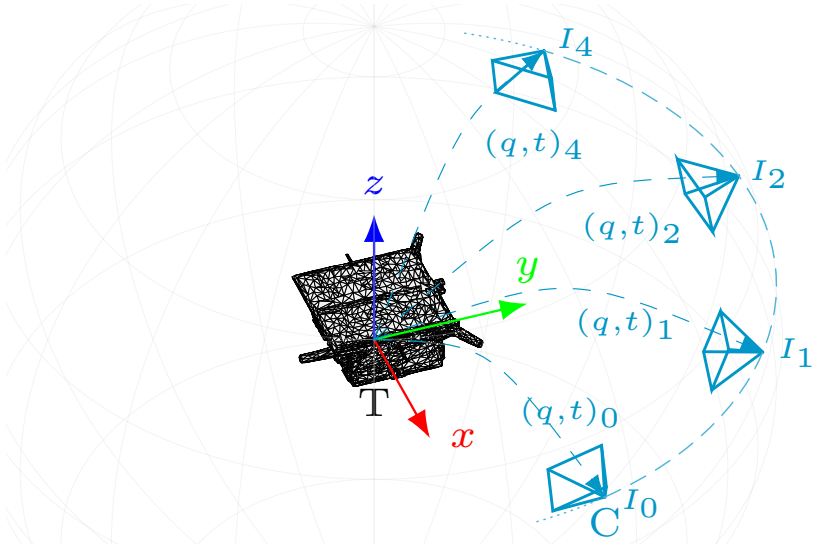


Fig. 2.1 Uncooperative spacecraft pose estimation from a stream of monocular images: problem overview. During a close-proximity operation, a camera mounted on a chaser spacecraft C takes a stream of pictures of the target T. At each timestep i the chaser must estimate the relative pose, *i.e.*, orientation quaternion q_i and position vector t_i based on the corresponding image I_i.

through fiducial markers or inter-satellite communication capabilities. Consequently, the chaser spacecraft must estimate the target pose exclusively from a sequence of images captured by its on-board camera.

In-orbit imagery is generally of low quality. Due to the absence of atmospheric diffusion, spacecraft surfaces directly illuminated by the Sun often appear strongly over-exposed, an effect amplified by specular surfaces that are common with spacecraft. Besides direct solar illumination, the target may also be lit by Earth albedo, but this contribution is orders of magnitude weaker, resulting in strong contrasts between lit and shadowed regions. Furthermore, illumination conditions may vary significantly along the approach. This is due to the relatively short orbital period of LEO, combined with the slow dynamics of rendezvous maneuvers[1]. Consequently, the pose estimation method must be robust to a wide range of illumination

[1]In both OOS and ADR missions, the duration of the rendezvous phase is not strictly constrained. Since the operation involves an autonomous spacecraft operating in proximity to a potentially tumbling target, a cautious and therefore slow approach should be preferred. In addition, we assume that the target tumbles at less than $5°/s$, which ensures that the inter-frame motion remains limited and compatible with vision-based tracking.

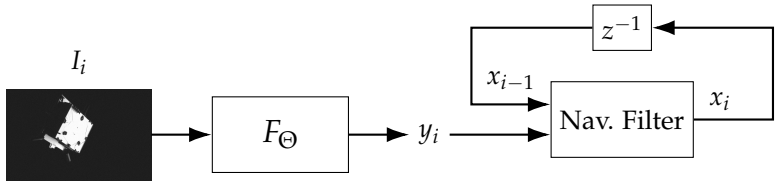


Fig. 2.2 Uncooperative spacecraft pose estimation from a stream of monocular images: methodological approach. A neural network F_Θ estimates the pose on each image independently. The pose estimates are then temporally aggregated through a navigation filter.

conditions. These factors collectively make vision-based pose estimation particularly challenging.

Figure 2.2 presents a high-level overview of the methods commonly used to address spacecraft pose estimation. The input image stream is processed frame by frame. For each image I_i, an estimator F_Θ generates a relative pose measurement y_i. This measurement is then provided to a navigation filter—typically an Extended or Unscented Kalman Filter [54–56]—which fuses the new measurement y_i with the previously estimated state x_{i-1} to obtain the updated relative state x_i. This state is defined by the relative position t, orientation quaternion q, and the corresponding translational and angular velocities v and ω. The estimator F_Θ may be coupled to the filter either loosely, by outputting a direct pose estimate (q_i, t_i), or tightly, by supplying indirect pose-related measurements, such as image coordinates of predefined spacecraft keypoints [57]. The design of an EKF or UKF tailored to spacecraft relative navigation is a subject of active research, with solutions proposed for both tightly-coupled [57–60] and loosely-coupled [57, 61–63] configurations. However, this thesis focuses exclusively on the design of the pose estimator F_Θ, whose role is to derive accurate pose estimates from potentially low-quality images depicting a known but uncooperative target. The temporal aggregation of these estimates through a navigation filter is therefore outside of the scope of this thesis.

Early approaches relied on hand-crafted image features such as keypoints or edges to infer the target pose [64–66]. Although computationally efficient and able to achieve accurate results under nominal lighting conditions, these methods lacked robustness to illumination variations [43]. Following the widespread adoption of deep learning in computer vision [67–71], data-driven approaches have been introduced for spacecraft pose estimation [43, 46].

These methods rely on a neural network F_Θ which predicts the pose

estimate $(\hat{q}, \hat{t})$ from an input image I, *i.e.* $((q, t) = F_\Theta(I))$. Such networks are characterized by their architecture—the arrangement and connectivity of their computational layers—and by a set of learnable parameters Θ optimized from a dataset of image–pose pairs.

The following sections address three central questions that hinder the adoption of data-driven pose estimators. Section 2.2 analyzes architectural tradeoffs of pose estimators while Section 2.3 discusses learning strategies required for robust performance across diverse conditions. Finally, Section 2.4 examines considerations related to the deployment of learned pose estimation networks on-orbit.

2.2 Neural Architectures for Spacecraft Pose Estimation

Neural networks used for visual tasks such as pose estimation typically comprise two sequential components. A **task-agnostic backbone** first extracts image features, which are subsequently processed by a **task-specific head** that predicts the output, *e.g.*, the relative 6D pose. The backbone may be implemented using a Convolutional Neural Network (CNN) [72], a Vision Transformer (ViT) [73], or a hybrid architecture combining both paradigms [74, 75]. While CNNs build image representations through hierarchical convolutional operations that aggregate local patterns, ViTs rely on self-attention [76] over image patches, enabling global, image-wide reasoning. Hybrid architectures integrate both mechanisms to jointly leverage local feature extraction and global contextual modeling.

This section focuses on the task-specific head, which governs the model's behavior for pose estimation, rather than on the backbone architecture. The remainder of the section presents an overview of end-to-end, keypoint-based, and hybrid architectures to summarize their respective tradeoffs.

2.2.1 End-to-End Architectures for Direct Pose Prediction

A first, straightforward approach, illustrated in Figure 2.3, consists in directly predicting the relative pose from the extracted image features using a single neural network F_Θ. Such end-to-end architectures [46, 48] rely exclusively on neural components to map an input image to a pose estimate. This formulation offers the advantage that **all neural components are optimized jointly using a pose estimation loss**, which penalizes both

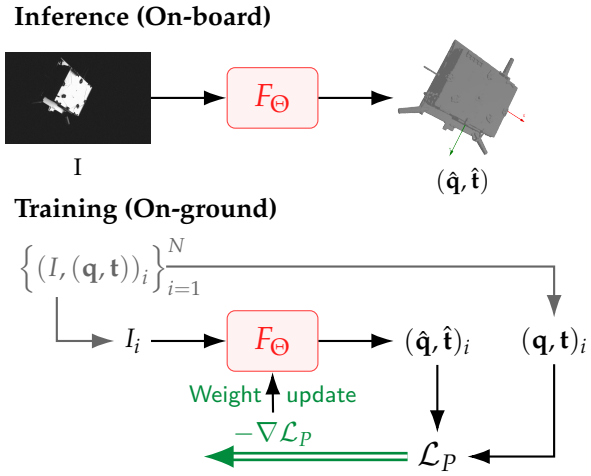


Fig. 2.3 End-to-end pose estimation network architecture $(\hat{q}, \hat{t}) = F_\Theta(I)$. At inference, the network F_Θ directly predicts the image pose $(\hat{q}, \hat{t})$. Before deployment, the network weights Θ are iteratively optimized using the gradients computed through the back-propagation of the pose loss $\mathcal{L}_p$, computed between pairs of predicted and ground-truth poses.

angular and translational errors, without relying on proxy objectives. However, learning the full mapping between any image of the target spacecraft and its corresponding relative pose is highly complex. As a result, most end-to-end spacecraft pose estimation networks are less accurate than other architectures.

2.2.2 Keypoint-Driven Pose Estimation Architectures

Because the direct image-to-pose mapping is difficult to learn, many approaches [47, 77] adopt keypoint-based architectures. As shown in Figure 2.4, the mapping F_Θ is decomposed into two stages. A neural network f_θ first predicts the 2D coordinates $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ of predefined spacecraft keypoints in the image plane. These keypoints usually correspond to salient points on the spacecraft, such as corners or tips of appendages. These predicted coordinates, together with their known 3D positions in the spacecraft reference frame and the camera calibration matrix, are then provided to a **Perspective-n-Point (PnP)** solver [78], which outputs the relative 6D pose.

This design is computationally efficient: the PnP solver has a linear

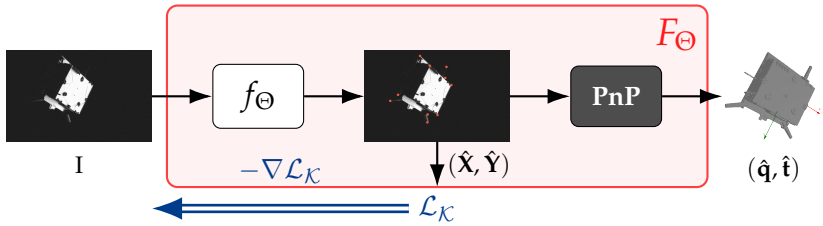


Fig. 2.4 Keypoint-based pose estimation network architecture $(\hat{q}, \hat{t}) = F_\Theta(I)$, where $F_\Theta = \text{PnP} \circ f_\Theta$

computational complexity with respect to the number of keypoints, and the neural network tackles a simpler prediction task. Predicting keypoints primarily requires **recognizing local image patterns**, which CNNs are explicitly engineered to learn. Consequently, keypoint-based architectures generally outperform purely end-to-end networks.

However, their optimization is not aligned with the final goal of pose estimation. Since the PnP solver is not differentiable, back-propagation of the pose loss through the solver is not possible. The network f_θ is therefore trained solely to minimize a keypoint regression loss, making it **optimal for keypoint accuracy, but sub-optimal for pose accuracy**, as optimal keypoint predictions do not necessarily guarantee optimal PnP outputs.

To solve this limitation, Backward-PnP [79] and subsequent works [80, 81] explored novel methods to approximate the gradients that could be computed if the perspective PnP solver were differentiable, e.g., through the implicit function theorem. Although they address the non-differentiability of PnP solvers, preliminary experiments conducted in this thesis showed that they do not improve the pose estimation accuracy compared to keypoint-based supervision.

2.2.3 Hybrid Pose Estimation Architectures

As previously discussed, keypoint-based methods outperform end-to-end ones but remain sub-optimal because the network is not trained using a pose-level objective. To address this limitation, several works [4, 5, 49, 82] propose hybrid architectures that combine both formulations. These models benefit from the **local geometric reasoning enforced by keypoint prediction** while also enabling **training with a pose-aligned loss**. Hybrid architectures can be either **sequential** [4, 5, 82] or **parallel** [49].

In the **sequential** configuration, illustrated in Figure 2.5, a first neural

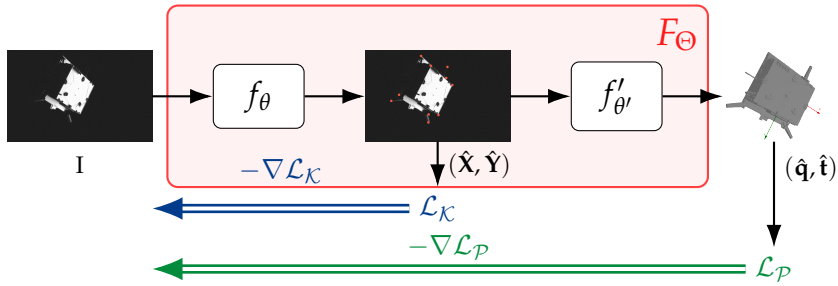


Fig. 2.5 Hybrid sequential pose estimation network architecture $(\hat{q}, \hat{t}) = F_\Theta(I)$, where $F_\Theta = f'_{\theta'} \circ f_\theta$

network f_θ predicts the coordinates of some predefined keypoints. A second network $f'_{\theta'}$ then takes these coordinates as input and regresses the relative pose. Because this second network is differentiable, the pose loss can be back-propagated through both sub-networks.

In addition to the pose loss, the keypoint loss may also be applied to stabilize training. Preliminary works conducted during this thesis [4, 5] validated the relevance of this architectural choice.

In the **parallel** configuration depicted in Figure 2.6, the shared backbone f_Θ extracts image features that are processed by two heads; each head being a network branch dedicated to a given learning objective. One head directly predicts the relative pose, while the other regresses keypoints, which can be combined with a PnP solver to compute an alternative pose estimate. This dual-head supervision encourages the backbone to learn

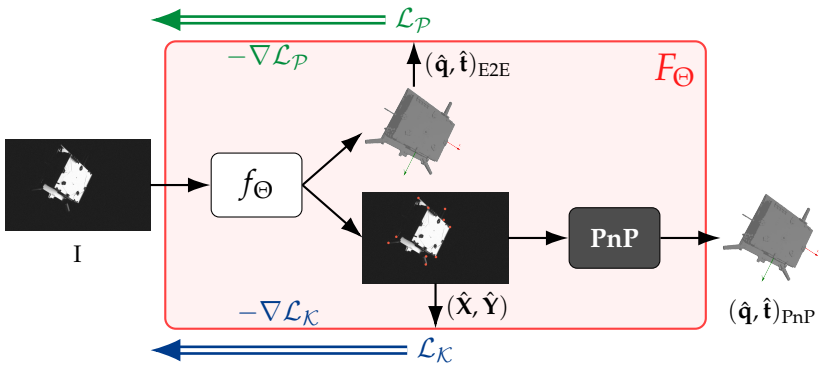


Fig. 2.6 Hybrid parallel pose estimation network architecture $(\hat{q}, \hat{t}) = F_\Theta(I)$

both **pose-relevant global features** (through the direct pose head) and **local geometrical structures** (through the keypoint head), improving generalization [49]. At inference, either of both heads can be used for pose prediction.

2.2.4 Architectural Tradeoffs and Limitations

We have discussed the distinctions between end-to-end, keypoint-based, and hybrid architectures. End-to-end models are conceptually simple and fully optimized via a pose loss, while keypoint-based approaches incorporate explicit geometric structure. Hybrid architectures combine both benefits by enforcing local reasoning through keypoints while retaining end-to-end differentiability.

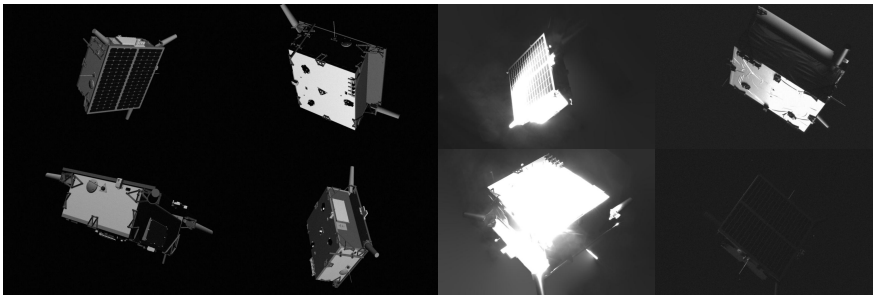
Despite these architectural differences, **all data-driven spacecraft pose estimation networks rely on the same fundamental assumption**. All methods are trained on a dataset depicting the target spacecraft. Because acquiring large numbers of orbital images is extremely challenging, such datasets are necessarily generated using rendering tools based on the target spacecraft CAD model. Therefore, every learning-based methods implicitly assumes that **the CAD model of the target spacecraft is available**.

State-of-the-art limitation #1

Existing data-driven pose estimation techniques require an accurate prior model of a spacecraft's geometry and appearance, such as a CAD model, limiting their applicability when such a model is not available.

2.3 Learning Strategies Under Domain Gap

As discussed in Section 2.2, spacecraft pose estimation networks are trained on synthetic images generated using the target CAD model. However, once deployed on-board, the model must operate on real images captured in orbit. The discrepancies between synthetic and real imagery introduce a significant challenge that must be addressed through an appropriate learning strategy. Section 2.3.1 formalizes this OOD generalization problem, while Sections 2.3.2 and 2.3.3 describe two families of techniques



Training images Test images

Fig. 2.7 Domain gap visualization. **(Left):** SPEED+ [45] synthetic images. **(Right)**: *real* images from SPEED+ [45]. *Real* images exhibit more challenging illumination conditions and texture mismatches compared to the synthetic training images.

designed to improve these generalization capabilities, namely multi-task learning and data augmentations. Finally, Section 2.3.4 highlights the limitations of current approaches and motivates the second contribution of this thesis.

2.3.1 From Domain Gap to Out-of-Domain Generalization

Figure 2.7 illustrates the differences between synthetic images and on-orbit imagery. Synthetic images generally exhibit smooth and homogeneous illumination conditions, whereas real images often contain strong contrasts between lit and shadowed regions. Real images also tend to be over-exposed due to direct solar illumination on specular spacecraft surfaces. Moreover, there are notable texture mismatches: synthetic renderings typically contain simple, low-detail textures, while real spacecraft are more textured. For example, the folds of the Multi-Layer Insulation (MLI) blanket are absent from synthetic images yet contribute significantly to the texture of real imagery. These illumination and texture discrepancies have a substantial impact on the pose estimator's performance on real images [45, 83].

This problem, known as the **domain gap**, arises from the differences between a source domain $\mathcal{S}$, *e.g.*, synthetic images, and a target domain $\mathcal{T}$, *e.g.*, on-orbit images, as illustrated in Figure 2.8. Because the network is trained only on source images, it learns the image-to-pose mapping based on patterns present in $\mathcal{S}$. If these patterns do not appear in $\mathcal{T}$, the model may fail to predict the pose accurately.

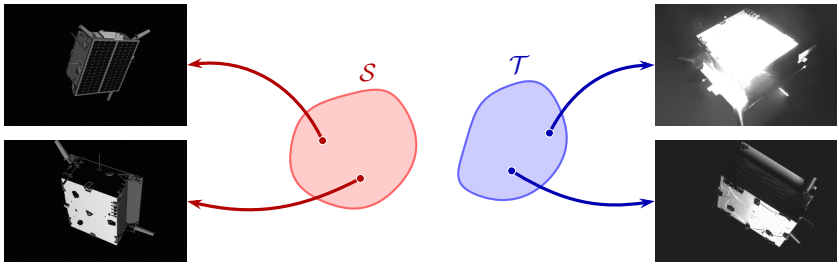


Fig. 2.8 Domain gap as a distribution shift: a neural network trained on a source domain $\mathcal{S}$, *e.g.*, synthetic images, performs poorly on a target domain $\mathcal{T}$, *e.g.*, on-orbit images, because the source and target images follow two different data distributions. As the network is only exposed to source samples during training, it struggles with the target samples.

Two main strategies have been proposed to mitigate this gap: domain adaptation [84] and domain generalization [85]. **Domain adaptation** assumes access to prior information from the target domain—such as labeled or unlabeled images—and exploits it to adapt the model to that specific domain. In contrast, **domain generalization** does not rely on any assumption regarding the target domain. Its goal is to improve performance on arbitrary unseen domains by making the learned representation robust to distribution shifts.

Because acquiring real on-orbit images of the target before the proximity operation is extremely difficult due to the lack of direct observational access, domain adaptation techniques are poorly suited to this problem. Domain generalization methods, however, are particularly relevant, as they aim to train pose estimators that remain robust to a wide variety of illumination and texture variations. The following sections therefore present two types of domain generalization techniques based on Multi-Task Learning (MTL) (section 2.3.2) and data augmentations (section 2.3.3), respectively.

2.3.2 Generalization through Multi-Task Learning

MTL consists in training a neural network not only on its primary task but also on one or several auxiliary tasks, as illustrated in Figure 2.9. These auxiliary tasks may include foreground segmentation [49], face segmentation [5], object detection [49], or depth estimation [51].

The network must therefore learn to satisfy multiple training objectives at the same time. This encourages the model to learn features that are useful across tasks, increasing its expressiveness and robustness. By forcing the

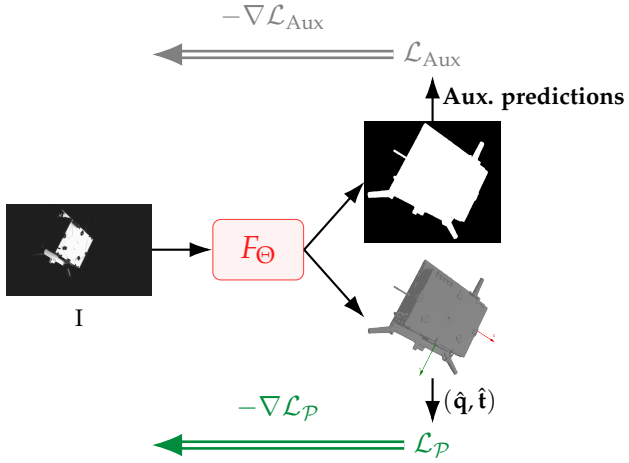


Fig. 2.9 Domain generalization through multi-task learning. In addition to the pose estimation task, the network F_Θ is trained on auxiliary task(s). These tasks may consist in foreground or faces segmentation, target detection, or depth estimation.

network to rely on more general and stable patterns, MTL improves its domain generalization capabilities.

2.3.3 Generalization through Data Augmentations

Figure 2.10 illustrates how data augmentations can enhance the generalization capabilities of a neural network facing domain mismatch. Instead of training the pose estimator directly on the source domain $\mathcal{S}$, the model is trained on an augmented domain $\mathcal{S}'$ obtained by applying augmentations to images sampled from $\mathcal{S}$. The network is thus encouraged to learn more diverse image patterns, which improves its robustness to appearance variations encountered in the target domain $\mathcal{T}$.

Figure 2.11 shows examples of augmentations relevant to spacecraft pose estimation. While standard pixel-intensity perturbations (noise, brightness, contrast) only marginally contribute to generalization, **spatially-localized content removal** and **distribution-level appearance changes** have a much stronger impact. Spatially-localized content removal methods, such as Hide&Seek [86] or synthetic over-exposure [87], prevent the model from relying solely on local cues and encourage the use of more robust global features. Distribution-level augmentations, such as Style Transfer [88] or Fourier-based techniques [89, 90], promote the use of features that are preserved across domains, such as the spacecraft shape.

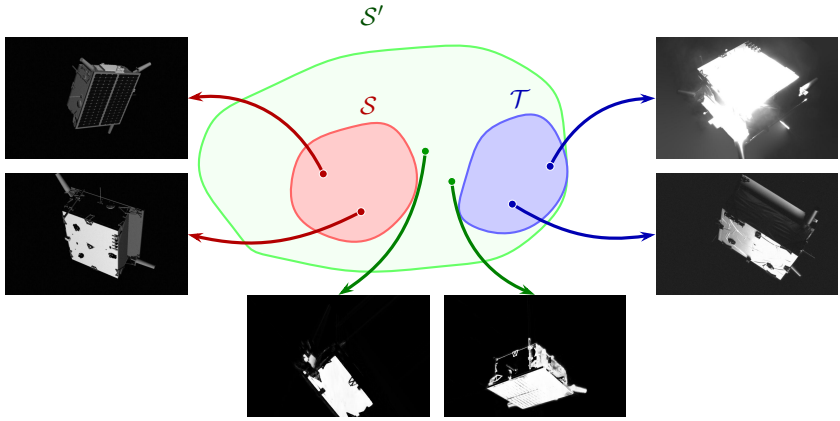


Fig. 2.10 Domain gap mitigation through data augmentations. Instead of training a pose estimation network on a source domain $\mathcal{S}$ that is too distant from a target domain $\mathcal{T}$, the network is trained on a domain $\mathcal{S}'$ obtained through data augmentations of $\mathcal{S}$. Since the training distribution is enlarged through augmentations, the trained neural network does not overfit the source domain and is therefore more resilient to domain mismatches.

To further increase diversity, augmentation techniques can be combined. Figure 2.12 illustrates images obtained by randomly combining three augmentations from the ones mentioned above. These composite augmentations produce distorted but pose-consistent images that preserve the spacecraft geometry while randomizing its appearance. Training on such randomized domains enhances OOD generalization [5, 91].

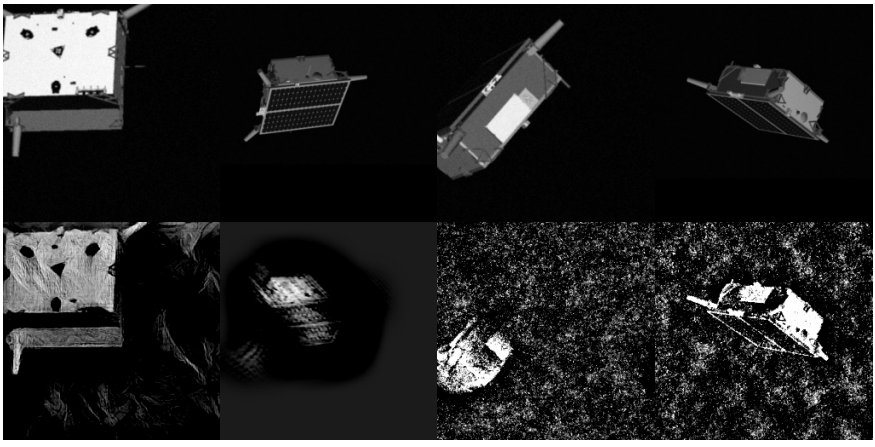


Fig. 2.12 Original (top) and domain-randomized (bottom) SPEED+ [45] images. The semantics are preserved but the texture and illumination significantly differ.

Fig. 2.11 Examples of image augmentations carried out on SPEED+ [45]. Different types of augmentation techniques exist. While some marginally affect the image, others remove some content or may significantly change its appearance. In all cases, the image appearance is modified, but the underlying semantics, *i.e.*, the target geometry, is preserved.

2.3.4 State-of-the-art Limitations

While augmentations improve OOD generalization, they operate by transforming individual images from the initial dataset. As a result, the diversity they introduce is inherently bound by the 2D image grid on which they are applied. Augmentations can alter appearance, but they cannot generate fundamentally new viewpoints, lighting conditions, or global variations that are absent from the base dataset. Therefore, their effectiveness remains directly limited by the diversity of the initial training set.

State-of-the-art limitation #2

Image-wise augmentation methods operate strictly within the 2D image grid and therefore cannot generate new viewpoints, geometric configurations, or illumination conditions absent from the original dataset. Their inability to modify the underlying 3D scene structure intrinsically limits the diversity achievable during training and restricts improvements in Out-Of-Domain (OOD) generalization.

2.4 Model Deployment

In Sections 2.2 and 2.3, we described the network architecture and training methodology required to develop a spacecraft pose estimation model that remains robust under in-orbit imaging conditions. This section addresses the two main challenges associated with the deployment of such a model in an actual mission. First, Section 2.4.1 examines the hardware platforms capable of performing on-board pose inference and introduces the fundamental tradeoffs that should guide hardware selection. Then, Section 2.4.2 discusses how the limited trust in AI-based systems for safety-critical applications affects the adoption of data-driven spacecraft pose estimation methods.

2.4.1 Hardware Platforms

During autonomous rendezvous and proximity operations, the chaser spacecraft must continuously estimate the target's relative pose using on-board computations. This requires executing the neural network inference on an

embedded hardware platform. Several classes of computing platforms can fulfill this role, each introducing distinct design tradeoffs. We first outline the criteria that should guide hardware selection, and then analyze the resulting tradeoffs.

Hardware Selection Criteria

The following criteria guide the evaluation of hardware platforms suitable for executing neural-network-based spacecraft pose estimation on-board a chaser spacecraft.

Throughput Estimating the target's relative pose from images requires processing substantial amounts of data per frame. Since the system must operate on a continuous image stream, inference speed directly determines the achievable measurement rate for the downstream navigation filter. An insufficient throughput would lower this measurement rate and would ultimately degrade the final pose estimation accuracy.

Power Efficiency The power available to the payload is constrained by the spacecraft platform and, ultimately, by mission cost. Indeed, increasing power allocation typically requires larger solar panels and batteries, which raise the spacecraft mass and, consequently, its overall cost. Hardware platforms must therefore provide the required performance within stringent power budgets.

Radiation Tolerance Spaceborne electronics are exposed to significant radiation levels throughout the mission lifetime. Radiation effects range from transient events such as Single-Event Upsets (SEUs)—temporary bit flips that do not damage the device—to permanent failures caused by latch-up or burnout. Long-term degradation due to Total Ionizing Dose (TID) also reduces component reliability. Mitigation strategies include the use of radiation-hardened (space-grade) components, hardware redundancy, and shielding to reduce the radiation level experienced by the device.

Versatility Some computing platforms, such as Central Processing Units (CPUs) and Graphics Processing Units (GPUs), inherently support multiple tasks, while others, such as Field-Programmable Gate Arrays (FPGAs), can be reconfigured to address a different task during the mission. This contrasts with Application-Specific Integrated Circuits (ASICs) which are limited to a fixed, application-specific function. Versatile platforms may

reduce overall mission cost by enabling shared use, *e.g.*, by performing both long and short-range pose estimation or processing data from a secondary payload.

Device Cost Mass-produced devices benefit from economies of scale and may offer attractive unit costs. Conversely, specialized or low-volume components can be prohibitively expensive, which directly affects mission costs.

NRE Cost Power efficiency, device cost, and versatility influence the marginal cost of the mission, but Non-Recurring Engineering (NRE) costs can vary widely across hardware classes Application-specific platforms, such as ASICs, require extensive engineering effort and months of development, significantly increasing upfront investment. These costs must therefore be carefully weighed when selecting a hardware solution.

Analysis of Hardware Trade-offs

Figure 2.13 qualitatively compares three classes of computing platforms using the selection criteria introduced previously.

ASICs are custom-designed chips optimized for a single task. They offer high throughput, excellent power efficiency, and strong radiation tolerance. However, they are costly to design and manufacture, require substantial non-recurring engineering effort, and lack versatility once fabricated.

In contrast, space-grade **CPUs** are inexpensive, radiation-tolerant, and highly versatile. They also entail limited development effort. Their main limitation is their modest computational capability, which makes them poorly suited for demanding tasks such as image-based pose estimation. Nevertheless, by constraining model complexity, several works have demonstrated the feasibility of deploying such methods on space-grade CPUs [49, 92, 93].

FPGAs occupy an intermediate position. These reconfigurable devices consist of programmable logic and routing resources that can be tailored to implement highly parallel, application-specific hardware architectures. They enable significantly higher throughput than CPUs while offering greater flexibility than ASICs. Space-grade FPGAs are available at reasonable cost and with acceptable power consumption. Their main drawback is the development effort required, which is higher than for CPU-based implementations but still considerably lower than for ASICs. Owing to this balance of performance, flexibility, and cost, FPGAs have been the preferred

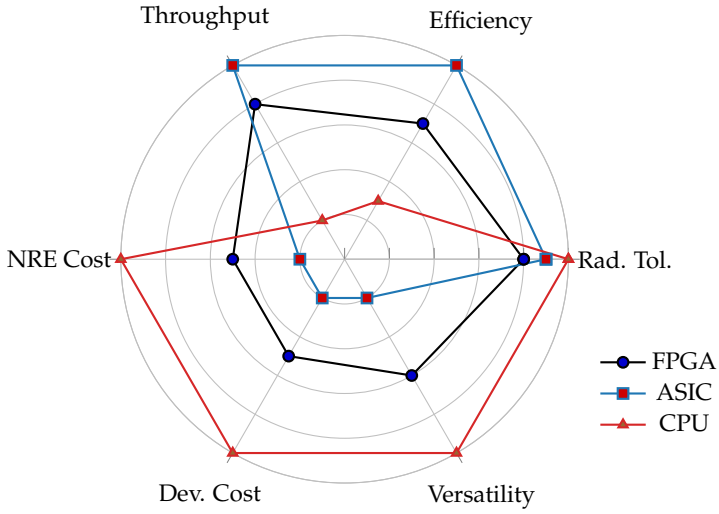


Fig. 2.13 On-board computing platforms: design tradeoffs between ASICs, CPUs and FPGAs. Further is better.

hardware platform considered for implementing spacecraft pose estimation systems [94–96].

For decades, ASICs, CPUs, and FPGAs formed the traditional *Triumvirate* of aerospace computing hardware. In recent years, however, their dominance has been challenged by the rapid rise of **GPUs** in computer vision. This evolution has led to the emergence of **Low-Power GPUs (LP-GPUs)** designed specifically for embedded systems, complementing the already widespread server-grade GPUs used during model training. Figure 2.14 compares FPGAs, server-grade GPUs and LP-GPUs and qualitatively assesses their respective suitability for on-board neural-network-based pose estimation.

Server-grade GPUs deliver exceptional computational performance and are already used to train the networks considered in this work. Deploying inference on such platforms requires limited engineering effort and provides high versatility. However, these benefits come at the cost of a limited power efficiency, high device price, and the absence of radiation tolerance.

Low-Power GPUs offer substantially better power efficiency and lower cost than server-grade GPUs, though this comes with a reduction in computational capability. Although they are not inherently radiation-hardened, recent studies indicate that commercial LP-GPU devices exhibit resilience

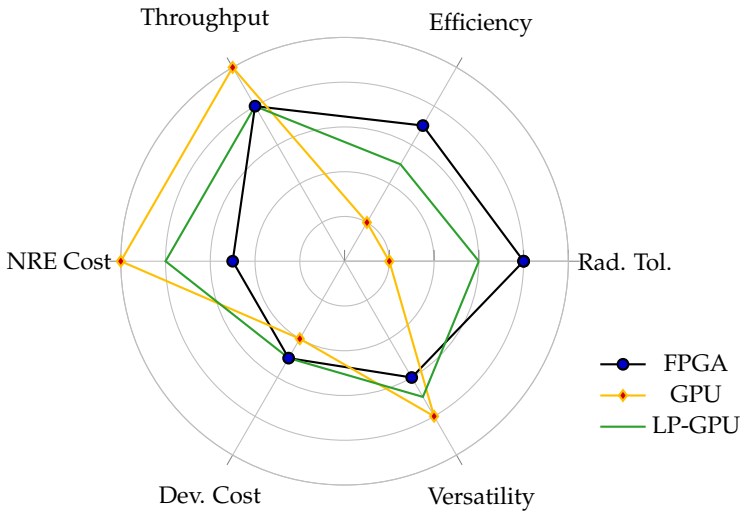


Fig. 2.14 On-board computing platforms: design tradeoffs between FPGAs, GPUs, and Low-Power (LP) GPUs. Further is better.

to radiation levels encountered in Low Earth Orbit, even without shielding [97, 98]. Their combination of computing capability, power efficiency, affordability, and reasonable radiation tolerance makes them potential candidates for running pose estimation networks on-orbit—albeit with a higher associated risk than FPGA-based solutions.

In summary, the comparison between CPUs, FPGAs, and low-power GPUs shows that multiple embedded platforms can reliably support the real-time inference required for spacecraft pose estimation. Although each option involves distinct trade-offs, all provide sufficient computational capability for on-orbit deployment. Hardware availability therefore does not constitute a limiting factor for implementing data-driven pose estimation methods in operational missions.

2.4.2 Human Trust in Data-Driven Systems

In the previous sections, we emphasized that spacecraft pose estimation algorithms have reached a level of maturity that would technically allow their integration within spacecraft payloads [49, 92]. Despite this progress, their adoption in the space industry remains slow, with most companies still prioritizing cooperative approaches based on fiducial mark-

ers. Although economic factors partly contribute to this trend—particularly given that the servicing market is still emerging and largely driven by government-led In-Orbit Demonstration (IOD) missions [32, 33, 99, 100]—a remaining semi-technical barrier continues to hinder the transition toward data-driven, uncooperative pose estimation methods.

A central challenge is the limited trust in data-driven systems, especially in aerospace applications where reliability and safety are paramount. To deploy a data-driven pose estimation model in orbit, stakeholders require strong guarantees that performance will translate from development environments to operational conditions.

Like most learning-based systems, these models are evaluated primarily through benchmarks that rely on imagery generated on the ground. However, producing high-fidelity, realistic datasets is both technically complex and costly. Hardware-In-the-Loop (HIL) facilities capable of replicating the relevant conditions—6-DoF robotic manipulators, high-quality cameras, and illumination sources that emulate both direct solar radiation and Earth albedo—demand substantial financial investment [101]. Their operation further requires specialized personnel, which increases overall cost. Moreover, such facilities can only emulate a limited set of scenarios, meaning that validation remains inherently constrained and cannot fully guarantee in-orbit performance. Consequently, ground-based testing alone is insufficient to build the level of trust required for deployment.

Beyond performance assessments on benchmarks, trust also depends on understanding how data-driven systems arrive at their predictions. Engineers prefer to understand how a system works before placing confidence in it. However, most data-driven models remain poorly understood. Despite their increasing adoption, they are frequently perceived as "black boxes," sometimes even by the very teams who develop them. Gaining access to their internal mechanisms and clarifying the logic behind their outputs could significantly enhance trust, reduce perceived risks, and ultimately accelerate their acceptance and deployment within actual missions.

State-of-the-art limitation #3

Data-driven spacecraft pose estimation models lack transparency in their internal decision mechanisms, limiting their suitability for safety-critical missions where trust and interpretability are essential.

2.5 A Spacecraft Pose Estimation Baseline: SPNv2

In this thesis, we introduce three novel contributions to address the limitations identified in the previous sections. We evaluate their impact using the *open-source* spacecraft pose estimation framework SPNv2 [49]. SPNv2 is a CNN-based method specifically designed to estimate the relative pose of an uncooperative spacecraft under domain gap.

The rest of this section therefore presents SPNv2. First, Section 2.5.1 positions SPNv2 with respect to the state of the art, with an emphasis on the design choices made by its authors. Then, Section 2.5.2 describes its neural architecture, while Section 2.5.3 presents the techniques deployed to improve its generalization capabilities. Finally, Section 2.5.4 discusses its limitations.

2.5.1 From Handcrafted Features to SPNs

As explained in Section 2.1, early spacecraft pose estimation approaches relied on handcrafted image features such as keypoints or edges to infer the target pose [64–66]. Although computationally efficient and able to achieve accurate results under nominal lighting conditions, these methods lacked robustness to illumination variations. Based on this observation, one of the teams working on vision-based spacecraft pose estimation through handcrafted features later published a seminal work on data-driven spacecraft pose estimation [43, 46]. Their work relied on a Spacecraft Pose Network (SPN), which is a CNN designed to estimate the relative target attitude and 2D bounding box. The relative position was then computed using the relative attitude and bounding box predictions, along with the target 3D wireframe model.

Although SPN represented a significant improvement compared to handcrafted pose estimation methods, it performs poorly on real images exhibiting harsh illumination conditions. For this reason, the same team from Stanford later released SPNv2. As further explained in Section 2.5.2, SPNv2 relies on a hybrid neural architecture to estimate the target pose through both keypoint-based and end-to-end branches. Through multi-task learning and 2D image augmentations, SPNv2 significantly outperforms SPN in terms of OOD generalization, as developed in Section 2.5.3.

The same authors later developed SPNv3 [102] to achieve better generalization performance. Unfortunately, this later work is not yet open source. Hence, because of its hybrid design, demonstrated robustness under domain gap, and full open-source availability, SPNv2 is a widely adopted spacecraft pose estimation baseline. Consequently, we adopt SPNv2 as the baseline pose estimation network, denoted F_Θ, to validate our contributions. The remainder of this section provides an overview of this baseline pose estimation network.

2.5.2 Architecture

Figure 2.15 illustrates the architecture of SPNv2 [49]. The input image I of resolution $W \times H$ is first processed by an EfficientNet backbone [103] composed of successive convolutional stages. Each stage produces feature maps of progressively decreasing spatial resolution, *i.e.*, $F^{(i+1)}_{W/2^i \times H/2^i}$ for $i = 2, 3, 4, 5, 6$. As depth increases, the resolution decreases while the extracted features become increasingly abstract.

This resolution–abstraction tradeoff motivates the use of a Bi-directional Feature Pyramid Network (BiFPN) [104], which fuses information across scales. High-resolution feature maps are therefore refined through the abstract representations extracted by deeper layers. The BiFPN outputs five refined feature maps, $P^{(3)}_{W/4 \times H/4} \rightarrow P^{(7)}_{W/64 \times H/64}$, each sharing the same resolution as its corresponding raw input.

The lower-resolution maps, $P^{(4)}$ to $P^{(7)}$, are fed into an EfficientPose head [105] that regresses the relative pose $(\hat{q}, \hat{t})^{(\text{reg})}$. The highest-resolution map, $P^{(3)}$, is provided to a second head that predicts the locations of K predefined keypoints via heatmaps. These keypoint coordinates $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ are then processed by a PnP solver [78] to estimate the relative pose $(\hat{q}, \hat{t})^{(\text{heat})}$. The same map $P^{(3)}$ is also used by a third head to generate the foreground segmentation mask $\hat{m}$.

SPNv2 therefore adopts a hybrid architecture in which the pose is predicted both end-to-end and via a keypoint-based pipeline, thereby combining the strengths of both paradigms (see Section 2.2.3). At inference, the network outputs the relative translation $\hat{t}^{\text{reg}}$ from the EfficientPose head, combined with the orientation quaternion $\hat{q}^{\text{heat}}$ predicted by the keypoint-based branch.

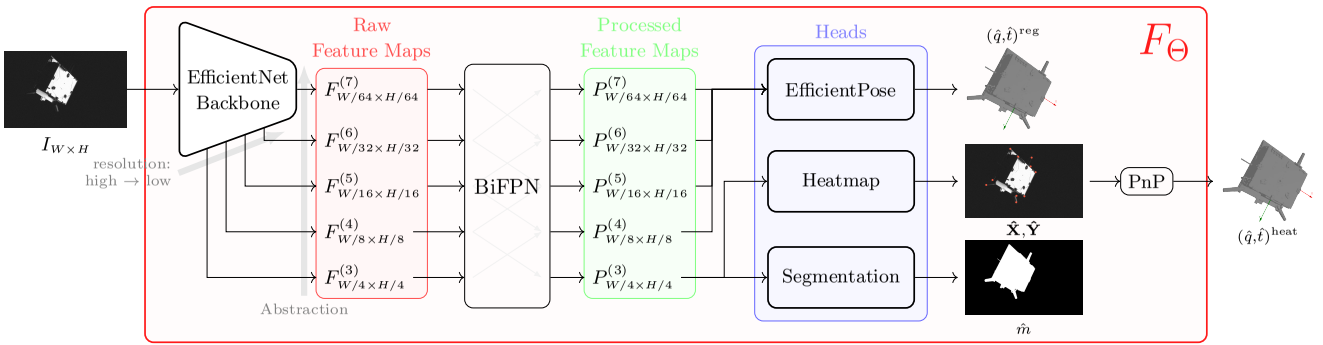


Fig. 2.15 SPNv2 [49] architectural overview: the network first extracts from the input image I multi-resolution feature maps $F^{(i+1)}_{W/2^i\times H/2^i}$ (with $i = 2,3,4,5,6$) through an EfficientNet backbone. These maps are then refined through a Bi-directional Feature Pyramid Network (BiFPN) to fuse the multi-scale information. Finally, the refined features $P^{(i+1)}_{W/2^i\times H/2^i}$ are processed by multiple task-specific heads. An EfficientPose head directly regresses the relative spacecraft pose $(\hat{q},\hat{t})^{\text{reg}}$. The second head predicts heatmaps corresponding to the location of K pre-defined keypoints $(\hat{\mathbf{X}},\hat{\mathbf{Y}})$. The corresponding keypoint coordinates are then given to a PnP solver to compute the relative pose $(\hat{q},\hat{t})^{\text{heat}}$. Finally, the third head outputs the foreground segmentation mask $\hat{m}$.

2.5.3 Domain Generalization Strategy

To enhance its domain generalization capabilities, SPNv2 relies on the two mechanisms introduced in Section 2.3, namely multi-task learning and extensive image augmentations. SPNv2's multi-task learning strategy first exploits the complementary pose predictions produced by the EfficientPose head and the heatmap-based branch. Each head serves as an auxiliary task for the other, encouraging the network to learn more robust and transferable features. This effect is further reinforced by the segmentation head, which imposes an additional auxiliary task.

SPNv2 further enhances robustness through a diverse set of photometric and noise-based image augmentations. To promote invariance to occlusions and appearance changes, SPNv2 employs occlusion-based augmentations such as random erasing [106] and sun flares [87], as well as texture randomization through neural style transfer [88]. Through these image-level augmentations, SPNv2 improves its OOD generalization performance.

Beyond these domain generalization mechanisms, SPNv2 also includes a domain adaptation component via an Online Domain Refinement (ODR) strategy, in which the model is finetuned using a small set of unannotated images from the target domains. Because this procedure requires direct access to the target domain, it serves domain adaptation rather than OOD generalization and is therefore not used in this thesis.

2.5.4 Strengths & Limitations

Among state-of-the-art spacecraft pose estimation methods, SPNv2 is the only open-source implementation that combines a carefully engineered architecture with an effective domain generalization strategy. In addition, SPNv2 is designed for low inference complexity, enabling deployment on space-grade CPUs.

Despite these advantages, SPNv2 exhibits the same limitations shared by current state-of-the-art approaches. These limitations are illustrated in Figure 2.16.

1. It requires a large training dataset ($\mathcal{O}(50,000)$ images) generated using the target CAD model, implicitly assuming the availability of such a model.
2. Its domain generalization strategy relies on image-level augmentations, so the induced randomization remains confined to the image plane.

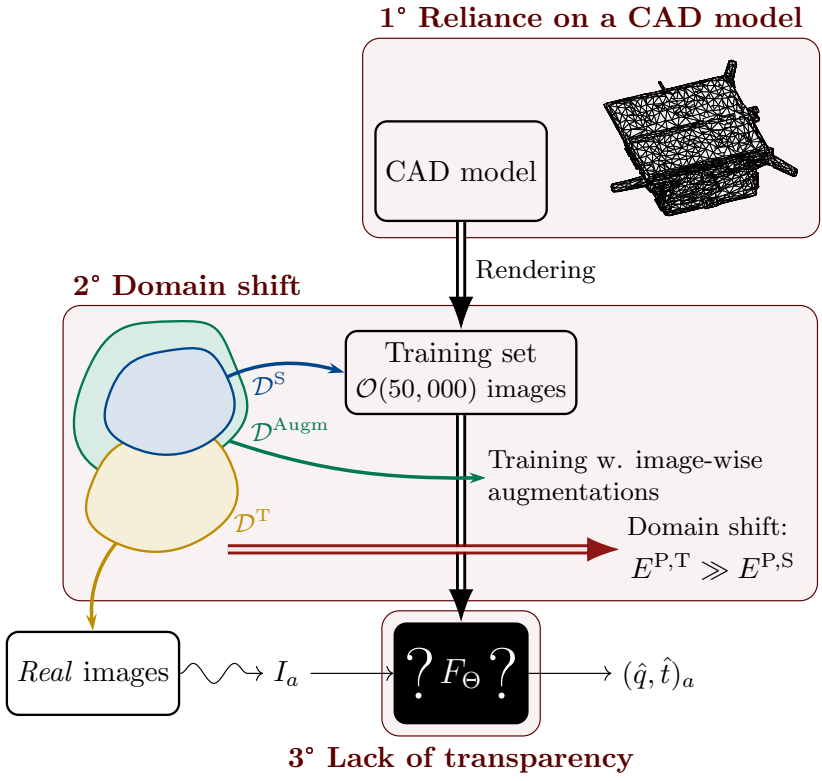


Fig. 2.16 Data-driven spacecraft pose estimation methods suffer from three main limitations. (i) They rely on an explicit target model to synthesize the large training set required to train the network F_Θ. (ii) Due to the distribution shift between synthetic (source) and real (target) domains, the network poorly generalizes even when 2D image augmentations are applied during training. (iii) Data-driven pose estimation methods lack transparency, as their decisions cannot be understood. These limitations hinder the adoption of data-driven spacecraft pose estimation methods.

3. As with all data-driven models, its decision process is opaque, thereby limiting the trust that can be placed in its predictions.

These limitations align with the three research axes addressed in this thesis, (i) namely the reliance on CAD models, (ii) Out-Of-Domain generalization, and (iii) model interpretability. Accordingly, we demonstrate on SPNv2 how our proposed methods effectively address these shortcomings.

2.6 Final Remarks

This chapter provided the foundations of spacecraft pose estimation required to understand the contributions developed in this thesis. As explained in Section 2.1, we focus on the estimation of the target spacecraft from a single frame at a time through a neural network. Data-driven methods were adopted in spacecraft pose estimation because of their ability to handle harsh illumination conditions better than handcrafted image processing approaches.

Through the analysis of three key aspects of data-driven methods, we identified the main limitations that hamper the adoption of data-driven spacecraft pose estimation methods in future missions. These limitations are highlighted in Figure 2.16. First, in Section 2.2, we presented the different types of neural architectures and showed that they nevertheless exhibit the same fundamental limitation, as they rely on an explicit CAD model to train the neural network. Then, in Section 2.3, we described how multi-task learning and image augmentations improve the OOD generalization capabilities of an estimator. We also highlighted that these image augmentations are limited and therefore restrict the OOD improvement. Finally, in Section 2.4, we highlighted that data-driven methods lack transparency and therefore cannot be trusted.

In this thesis, we address these three limitations through the use of NVS methods. To evaluate how our NVS-based contributions improve spacecraft pose estimation methods, we rely on a state-of-the-art spacecraft pose estimation model, SPNv2 [49]. As discussed in Section 2.5, SPNv2, despite being state of the art, exhibits the three limitations addressed in this thesis and is therefore of great interest to assess how our NVS-based contributions alleviate these limitations. In the next chapter, we present the basics of Novel View Synthesis (NVS) methods.

3

NeRF-based Novel View Synthesis

NOVEL VIEW SYNTHESIS (NVS) can be interpreted as the inverse problem of pose estimation. While a pose estimation network F_Θ aims at predicting a pose (q,t) given an image I, *i.e.*, $(q,t) = F_\Theta(I)$, a novel view synthesis tool M_Φ aims at rendering an image I under some viewpoint (q,t). This intrinsic relationship between novel view synthesis and pose estimation is at the heart of this thesis as we exploit this inverse function M_Φ to address three key limitations that undermine the adoption of spacecraft pose estimation networks in real missions. This chapter therefore presents background material on novel view synthesis with an emphasis on NeRFs which are adopted in this thesis to implement M_Φ.

This chapter is organized as follows. First, Section 3.1 formalizes the novel view synthesis problem and provides a high-level view of the method employed to solve the problem. Then, Section 3.2 describes the different parametrizations proposed to implement the M_Φ function and highlights how one of these parametrization, *i.e.*, the NeRF, is well suited for our application. Finally, Section 3.3 provides an in-depth description of the image synthesis process performed by a NeRF and therefore introduces the mathematical notations used throughout the thesis.

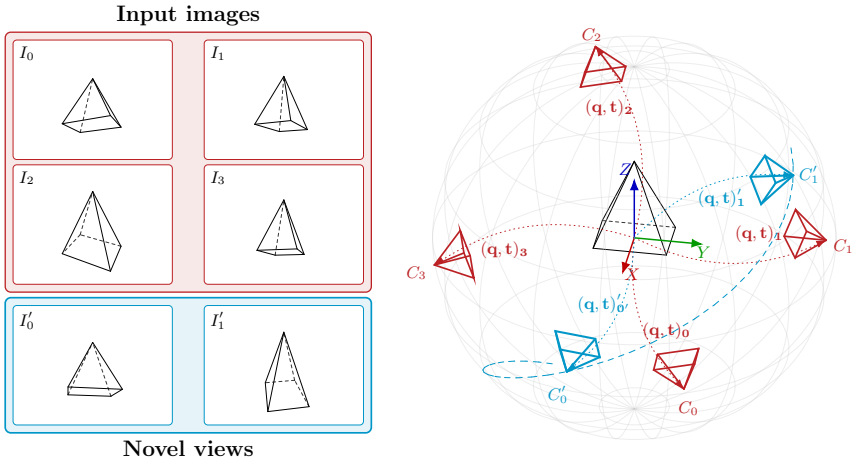


Fig. 3.1 Novel view synthesis: problem overview. Given a set of images $\{I_i\}_{i=0}^{N}$ taken by cameras $\{C_i\}_{i=0}^{N}$ under viewpoints $\{(q,t)_i\}_{i=0}^{N}$ relative to the scene's coordinates, we want to generate the image I'_j that would be taken by a virtual camera C'_j under any novel viewpoint $(q,t)_j$.

3.1 Novel View Synthesis

Figure 3.1 illustrates the novel view synthesis problem. Given a set of images $\{I_i\}_{i=0}^{N}$ depicting a scene, *e.g.*, a pyramid placed at the world's origin, we want to generate images $\left\{I'_j\right\}_{j=0}^{N}$ of that scene under novel, arbitrary viewpoints $\left\{(q,t)_j\right\}_{j=0}^{N}$. The scene is assumed to be static, *i.e.*, immutable over time, so that the scene's geometry does not evolve over time. It is also assumed that the pose $(q,t)_i$ of each camera C_i relative to the world's origin is known.

For this purpose, novel view synthesis methods rely on a learnable representation M_Φ, where M denotes the mapping between a relative pose (q,t) from the $\mathrm{SE}(3)$ space to an image I and Φ denotes its learnable components. The set of captured images $\{I_i\}_{i=0}^{N}$ along with their respective viewpoints $\{(q,t)_i\}_{i=0}^{N}$ is used to learn M_Φ, so that, for any novel viewpoint $(q,t)_i$, the corresponding image can be generated through M_Φ, *i.e.*, $I'_j = M_\Phi\left((q,t)_j\right)$.

As further discussed in Section 3.2, multiple formulations of M_Φ have

been proposed in the literature based on different parametrization of the scene. However, they all rely on updating the representation weights Φ using the gradients back-propagated through M_Φ of a photometric loss computed between the image generated by the network I' and the actual observed image I. The photometric loss consists in the pixel-wise Mean Squared Error (MSE) computed between the observed and generated images, *i.e.*,

$$\text{MSE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \left(I_{ij} - I'_{ij} \right)^2, \tag{3.1}$$

where W and H denote the image width and height, respectively.

Figure 3.2 depicts the overall trend of the photometric loss during the training of a novel view synthesis representation, along with the corresponding image and error maps. It highlights that the use of the photometric loss encourages M_Φ to learn during the early phase of the training the low-frequency content of the scene, *i.e.*, the overall shape of the scene's components. The high-frequency content is then progressively learned so as to improve the sharpness of the learned scene.

3.2 Scene Parametrization

As mentioned in the previous section, different formulations can be used to define the representation M_Φ. Two main types of parametrization exist and depend on whether the scene's description is *explicit*, *i.e.*, the scene is approximated through a set of discrete primitives or *implicit*, *i.e.*, the scene is represented as continuous functions mapping coordinates to color and density.

3.2.1 Explicit Representations

An explicit representation describes the scene's geometry and appearance through a set of primitives. The learned representations belong to three main families that offer different tradeoffs. Earlier works relied on **Point**-based or **Voxel**-based representations while more recent works adopted **shapes**-based representations.

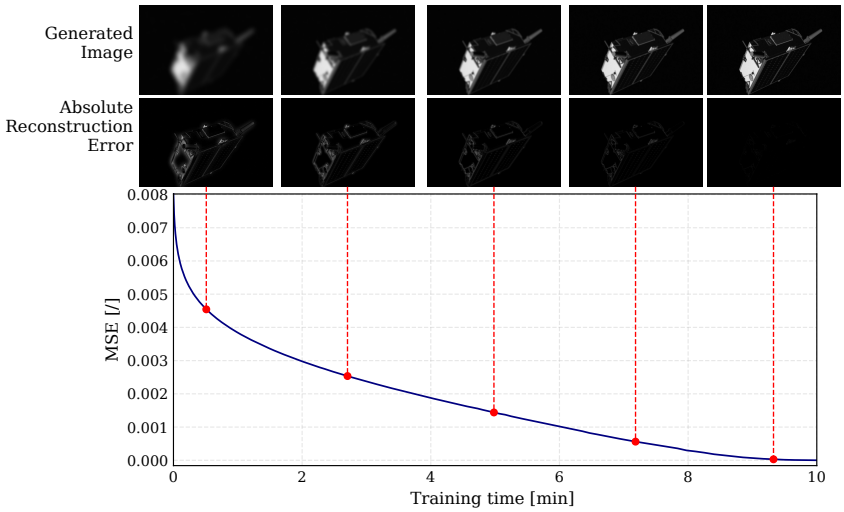


Fig. 3.2 Typical evolution of the photometric loss during training along with the corresponding images and absolute reconstruction errors. The photometric loss, *i.e.*, pixel-wise MSE, encourages the learning of the scene's low-frequency content during the first training steps. The latter steps enables the recovery of the high-frequency content, such as the spacecraft's edges.

Point-based approaches represent the scenes as an enhanced, radiance-aware point cloud. For example, Neural Point-Based Graphics [107] encodes local geometry and appearance information through point descriptors to integrate radiance into the point cloud. This idea was further refined by Point-NeRF [108] which does not only learn the features but also the point positions. While these point-based techniques offer a low memory usage due to their sparse space sampling, they often struggle with high-frequency content because of that sparse sampling.

Voxel-based approaches represent the 3D scene as a regular grid or as a hierarchical octree. For example, PlenOctrees [109] convert a trained NeRF into an octree structure containing color and density values, enabling real-time rendering through a simpler inference mechanism. Direct Voxel Grid Optimization [110] goes further by predicting the density along a shallow color network directly within the voxel grid. Voxel-based approaches are intrinsically limited by their discretization of the 3D space since the grid 3D resolution inherently determines the minimal size of the high-frequency features that can be represented. Any increase of that minimal resolution comes at a cost of an increased complexity and larger memory footprint.

Most recent approaches rely on **shape**-based representations to approximate the scene geometry. 3D Gaussian Splatting [111] describes the scene through anisotropic 3D Gaussians. Each Gaussian is defined by a shape, a color and an opacity so that an image can be rendered at a low computational cost by splatting the Gaussians on the image plane. Later works aimed at increasing the geometric sharpness and rasterization efficiency through the use of 3D convexes [112] or triangles as primitives [113]. These shape-based explicit representations provide excellent geometric fidelity and real-time rendering speed but their reliance on primitives with few degrees of freedom limits their ability to model complex illumination conditions.

Explicit representations offer strong advantages. They have a low rendering complexity which turns into a real-time image generation speed and can easily be edited by adding or removing primitives. However, these benefits, which are inherent to their discrete structure, comes at the cost of a reduced ability to handle complex, view-dependent, illumination conditions such as the ones encountered on-orbit without substantially increasing the number of primitives and hence the complexity.

3.2.2 Implicit Representations

An implicit neural representation describes a scene as continuous functions that map 3D spatial coordinates, and possibly viewing direction, to geometric and radiance properties. There exist several types of implicit representations designed for different purposes. Most of them are geared toward **appearance modeling** or put the emphasis on **geometric reconstruction**.

Radiance fields, such as *Neural Radiance Fields (NeRF)* [52], approximate the scene's volumetric density and view-dependent color through continuous functions. A NeRF maps a 3D location and a viewing direction to a density and emitted color. Images are synthesized by sampling points along camera rays and integrating densities and colors using differentiable ray-tracing techniques. This formulation makes them particularly **well suited for modeling scenes with complex appearances**, such as subtle illumination effects and view-dependent phenomena, which are typically more difficult to capture with explicit representations.

In contrast, *occupancy networks* [114] and *Signed Distance Functions* [115] are implicit representations tailored for **accurate geometric reconstruction**. Occupancy networks [114] represent a surface as the decision boundary

of a classifier, implicitly delimiting the scene's contours. SDF-based methods such as DeepSDF [115] represent the scene through a learned function whose zero-level set corresponds to the surface, with positive or negative values indicating points outside or inside the object. Although these geometry-centric methods were not initially developed for novel view synthesis, they provide high-quality shape priors and have therefore been extended into novel view synthesis methods. Hybrid methods such as VolSDF [116] or NeuS [117], integrate SDFs into the differentiable image generation process so that they often lead to more accurate shape reconstruction than NeRFs.

Implicit representations offer several advantages for 3D scene modeling as they provide continuous geometric and radiometric fields and naturally grasp complex lighting and view-dependent appearances. Their limitations, such as slower training and rendering, limited editability, are not critical for our application, as we do not need to edit the model nor to train it and/or generate images on-orbit but on-ground, where the computational cost is not an issue. Since we focus on an accurate appearance modeling under challenging on-orbit illumination conditions, we favor NeRFs over SDF-based representations because of their inherent ability to represent such challenging appearances.

3.3 Neural Radiance Fields

This section presents the fundamentals of Neural Radiance Fields. We first describe the image synthesis process from camera rays to rendered pixels. Then, we introduce the underlying neural field architecture and its main design variants, discuss extensions for handling variable illumination through appearance embeddings and highlight the partial decoupling between geometry and appearance. Finally, we conclude with the training formulation.

Image Synthesis

Figure 3.3 illustrates the synthesis of an image I of dimensions $W \times H$ from a camera pose (q, t), where q denotes orientation and t denotes position, using a NeRF M_Φ. For each pixel, a ray is cast from the camera center through the pixel center. The ray is defined by its origin $c_{ij} = t$ and unit direction $d_{ij} = R(q)\, d^{\mathrm{P}}_{ij}$, where d^{P}_{ij} represents the direction of a ray

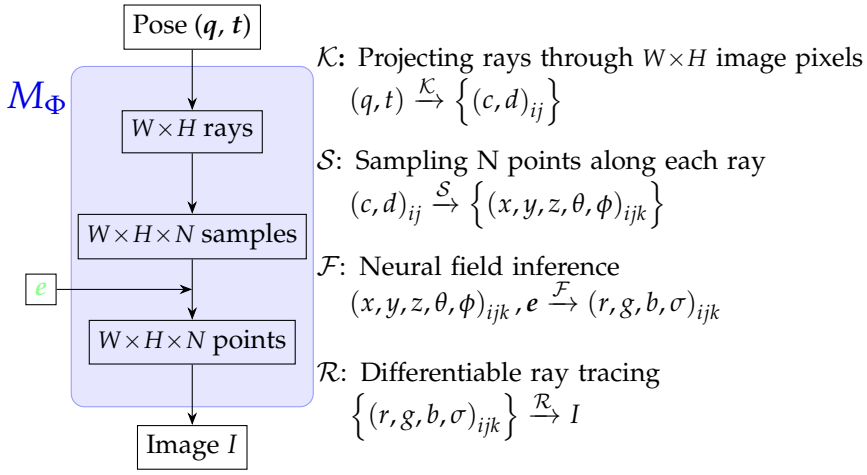


Fig. 3.3 Image generation through a Neural Radiance Field. Given a pose (q,t), the NeRF first projects rays, represented by the camera center c_{ij} and unit direction d_{ij}, through the center of every pixel ij of the corresponding virtual camera. Then, for each ray, N^S points are sampled along that ray. Each point is defined by its position (x,y,z) and 2 viewing angles (θ,ϕ). For each point, the neural field $\mathcal{F}$ predicts the point's color r,g,b and density σ. Finally, through differentiable ray-tracing techniques, the pixel value I_{ij} is recovered for each pixel of the image.

originating from the center of a world-aligned camera and passing through pixel ij. Along each ray, N points are sampled. For each sample, the 3D position (x,y,z) and viewing direction (θ,ϕ) are given to a neural field $\mathcal{F}$, which predicts the density σ and color (r,g,b) of the point. Finally, the pixel value is computed by aggregating the color and density of all sampled points along the ray using differentiable ray-tracing techniques.

Neural field architecture

While many architectures exist, the neural field, which predicts the density and color of a point from its position and viewing angle, follows a common core principle. Each position (x,y,z) and viewing direction (θ,ϕ) are first encoded into position F_{pos} and direction features F_{dir}, respectively. A MultiLayer Perceptron (MLP), fed with the position features, then outputs the density σ along with intermediate density features F_σ. These features are concatenated with the direction ones and passed to a second MLP that predicts the color (r,g,b).

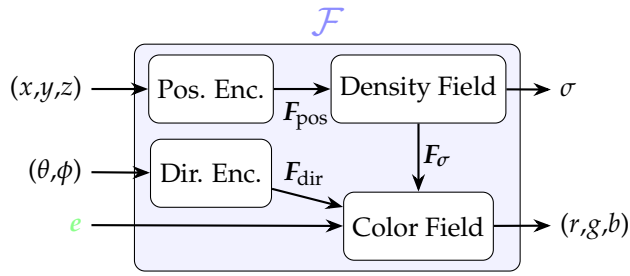


Fig. 3.4 Overview of the neural field $\mathcal{F}$ architecture. The position (x, y, z) and viewing angles (θ, ϕ), are first mapped to position and direction features F_{pos} and F_{dir}, respectively. The point density σ, along with density features F_{σ}, are regressed by a MLP that approximates the scene's density field. Finally, the point's color (r,g,b) is estimated from the density and direction features by a second MLP that approximates the color field. In in-the-wild NeRFs, this color MLP also takes as input an appearance embedding which implicitly model the image appearance such as the illumination conditions (see Section 3.3).

Key differences between NeRF implementations lie in the specific formulations of their neural field components. The original NeRF [52] employed sinusoidal positional encoding for both spatial coordinates and viewing directions, combined with large MLPs to predict density and color. Although effective, this design led to slow training and rendering of novel views. Instant-NGP [118] addresses these limitations by introducing a learnable multi-resolution hash grid for encoding positions, and spherical harmonics for encoding viewing directions. These innovations enable the use of smaller MLPs, significantly accelerating both training and inference while maintaining reconstruction quality.

Handling variable lighting conditions

While such architectures are well suited for static scenes, they are not designed to handle illumination changes that naturally occur in unconstrained environments. To address this limitation, NeRF-in-the-wild [119] introduced the concept of appearance embeddings to model appearance discrepancies across training images. An appearance embedding is a learnable vector associated with a specific image, which is fed into the color MLP alongside the direction and density features. As the appearance embeddings for all training images are jointly learned with the NeRF, they capture the illumination specificities of their corresponding images.

In this thesis, we rely on a Neural Radiance Field based on the *K*-Planes [53] architecture, because it incorporates appearance embeddings while remaining computationally efficient through a positional encoding scheme similar to Instant-NeRF [118]. Figure 3.4 illustrates the high-level architecture of the *K*-Planes-based neural field. Specifically, the positional and directional encoders consist of 3-plane interpolation and spherical harmonics, respectively.

Partial decoupling of appearance and geometry

A salient feature of NeRFs is their ability to partially decouple the scene's density from its appearance. Indeed, as depicted in Figure 3.4, some field components such as the appearance embedding, the direction encoding, and the color field do not affect the scene's density. Hence, any modification of these components would affect the scene's appearance without affecting its underlying geometry. In Chapter 6, we show that this partial geometry-appearance decoupling can be exploited to generate images exhibiting diverse appearances under diverse viewpoints while ensuring the scene's 3D consistency.

Training

Since the entire image generation process is differentiable, NeRFs can be trained by back-propagating the photometric loss between the observed images and those rendered from the corresponding viewpoints, *i.e.*,

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \mathcal{L}_{\text{Photo}} \left(I_a, M_\Phi(q_a, t_a, e_a) \right) \tag{3.2}$$

where Φ^* denotes the optimal NeRF weights. In practice, however, NeRFs are trained on batches composed of pixels randomly sampled from the training images along their corresponding rays, *i.e.*,

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \mathcal{L}_{\text{Photo}} \left(I_{aij}, \mathcal{R}(\mathcal{F}(\mathcal{S}(c_{aij}, d_{aij}, e_a))) \right) \tag{3.3}$$

This strategy ensures that each batch contains pixels from diverse viewpoints, promoting multi-view consistency and ultimately improving training convergence.

3.4 Final Remarks

In this chapter, we presented how NVS representations enable the synthesis of novel views of a scene from a limited set of images. In particular, we highlighted the interest of implicit representations such as NeRFs over explicit representations such as 3D Gaussian Splatting. While the former is inherently more complex due to its ray-tracing image synthesis paradigm, it comes with a better ability to represent harsh and varying illumination conditions, especially through learnable appearance embeddings. In our contributions, the NVS representation can run on the ground, so that the computational complexity is not an issue compared to the in-orbit illumination conditions. Furthermore, NeRFs have the intrinsic ability of partially decoupling the scene's geometry and appearance. This ability is exploited in Chapter 6, as it enables the generation of images with diverse appearances while preserving the scene's geometry. For these reasons, this thesis exploits a NeRF as the NVS representation used to address the limitations of current spacecraft pose estimation networks.

In the next chapter, we present the datasets and metrics used to validate our NVS-based contributions on the spacecraft pose estimation networks discussed in the previous chapter. These pose estimation networks, NVS models, datasets, and metrics form the foundations of this work.

4

Datasets & Metrics

DATA-DRIVEN spacecraft pose estimation depends on high-quality datasets that support both the training of neural models and the rigorous evaluation of their performance. This chapter introduces the datasets relevant to this thesis. Together with the pose estimation and novel view synthesis background presented in Chapters 2 and 3, these datasets constitute the foundation required to develop and evaluate the contributions proposed in the second part of the thesis.

First, Section 4.1 reviews existing datasets designed for data-driven proximity operations and highlights their respective scopes and limitations. Then, Section 4.2 presents the specific datasets used throughout this thesis and defines the nomenclature adopted in the experimental sections. Section 4.3 introduces the evaluation metrics employed to quantify pose estimation performance across the considered domains while Section 4.4 presents the image quality metrics used in our experiments.

4.1 Landscape Overview

Despite methodological similarities with general 6D pose estimation, spacecraft pose estimation presents characteristics that differ from terrestrial use cases. For example, while most generalistic 6D pose estimation methods

try to generalize to different object geometries, spacecraft pose estimation models are intrinsically designed for a single target. Similarly, while the main challenge in space is the harsh illumination conditions that can occur in orbit, on the ground, the illumination conditions are smoother.

Consequently, widely exploited terrestrial pose estimation datasets such as LineMOD [120] and YCB [121] cannot serve as relevant proxies, and mission-relevant performance must be evaluated using datasets specifically designed for the space environment.

Because the field remains relatively niche, only a limited number of datasets representative of proximity operation conditions have been released publicly. Some datasets address related but distinct tasks, *e.g.*, spacecraft recognition [122], component segmentation [123, 124], or alternative sensing modalities such as event cameras [125], whereas most available datasets focus directly on training spacecraft pose estimation models. The majority of spacecraft pose estimation datasets [46, 48, 92, 126–129] consist exclusively of synthetic renderings. While effective for learning image-to-pose mappings, they do not enable a rigorous evaluation of real-world performance due to the domain gap between synthetic imagery and in-orbit observations (see Section 2.3.1).

To the best of our knowledge, only four publicly available datasets combine synthetic images with data acquired under illumination conditions representative of the space environment. The first one, CubeSat-CDT [130], is not relevant for our application as it only depicts a symmetric 1U-cubesat that cannot be considered as economically sustainable target for OOS nor ADR. It is therefore not exploited in this thesis. The second dataset [131] was released recently under an European Space Agency (ESA) contract, but its image sets do not introduce features beyond those of the two datasets employed here. The two datasets used in this thesis are the remaining ones: SPEED+ [45] and SHIRT [58], which are described in the following section.

4.2 TANGO-based Datasets

This section introduces the two datasets used to validate our experiments. Both datasets are relevant to distinct parts of the thesis and are therefore employed throughout multiple chapters. Since they were generated by the same team from Stanford University using similar tools and procedures, they share several characteristics, which are outlined in Section 4.2.1. Sections 4.2.2 and 4.2.3 describe the SPEED+ and SHIRT datasets,

respectively, while Section 4.2.4 defines the nomenclature adopted in the experimental sections when referring to their image sets. Finally, Section 4.2.5 discusses the limitations that are affecting the realism and relevance of these spacecraft pose estimation datasets.

4.2.1 Overview

In a seminal work on spacecraft pose estimation, Sharma and D'Amico released SPEED [46], a dataset composed of synthetic images of the TANGO spacecraft from the PRISMA mission [132]. PRISMA was an IOD mission in which several strategies for conducting on-orbit rendezvous and proximity operations were tested using two satellites: TANGO (Target) and MANGO (Chaser). Section 4.2.1 shows the 3D model of TANGO, along with the K=11 keypoints commonly used for pose estimation.

It consists of a spacecraft body of 74 cm$\times$56 cm$\times$28 cm with a solar panel on top with dimensions of 74 cm$\times$71 cm$\times$4 cm. The spacecraft is equipped with three antennas on the corners of its +Z face and two thinner antennas on the +X and -X faces.

The same team later released SPEED+ [45] and SHIRT [58] which are summarized in Table 4.1 and described in the following sections. Both datasets consists of grayscale 8-bit images of 1200$\times$1920 pixels of TANGO along with the corresponding pose labels. The images are either synthetically generated or acquired in a HIL setup.

The synthetic images are produced using an OpenGL-based renderer configured to emulate the camera used in the HIL environment. Because it relies on rasterization, the renderer provides only an approximate reproduction of on-orbit illumination. It also employs a simplified CAD model of the spacecraft that lacks fine textural details, such as the MLI blanket depicted in Table 4.1.

The HIL images, referred to as *real* images, were captured at the Testbed for Rendezvous and Optical Navigation (TRON) [101] at Stanford University. This facility reproduces on-orbit illumination on a high-fidelity mock-up of the target spacecraft: direct solar illumination is simulated using a metal-halide arc lamp, while diffuse Earth-albedo illumination is generated through light boxes. Pose labels for these images are estimated from infrared markers placed on the robotic arms that control both the camera and the target. Hence, the pose labels suffer from some uncertainty.

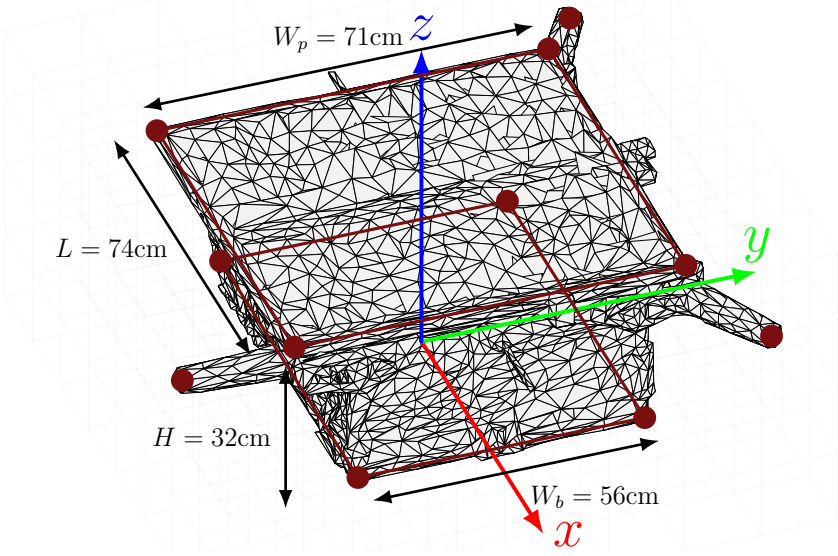
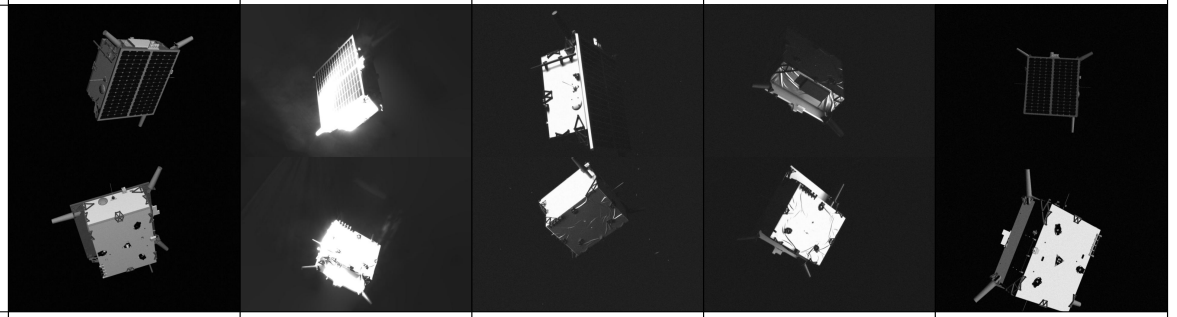


Fig. 4.1 3D model of the TANGO spacecraft along with the K=11 keypoints usually used for pose estimation. The spacecraft is made of two core elements: the spacecraft body and its solar panel. It is also equipped with five antennas. Three large antennas are fixed on three +Z corners of the spacecraft body while two thinner ones are placed in the middle of the +X and -X faces. The keypoints correspond to the 8 corners of the spacecraft body and the extremity of each of the 3 largest antennas.

For each image I_i, the corresponding pose label consists of an orientation label and a position label. The orientation label is a right-handed quaternion q_i, whichrepresents the rotation from the target reference frame to the camera reference frame. Quaternions are privileged for representing orientation, as they do not suffer from the gimbal lock that affects Euler angles. The position label is a 3D vector t_i that represents the position of the target's reference frame origin in the camera reference frame.

In addition, for each image, the dataset contains the corresponding foreground segmentation mask m_i. This binary mask indicates which parts of the image correspond to the target spacecraft and which parts depict only the background, which is not relevant for estimating the target pose. Each mask is obtained by projecting the 3D CAD model onto the image plane under the considered pose label through the OpenGL renderer. These binary masks are not available at inference but can be used during training, *e.g.*, through an auxiliary segmentation task.

Table 4.1 Overview of the 5 image sets used in this thesis along with the adopted set nomenclature. The 3 sets from SPEED+ are generated from poses randomly sampled in SE(3) while the 2 sets from SHIRT's ROE2 correspond to an approach trajectory. Images are either synthesized using the CAD model or captured in a HIL setup using a spacecraft mock-up. The HIL configuration enables two illumination domains: *Lightbox*, reproducing diffuse Earth-albedo lighting, and *Sunlamp*, emulating direct solar illumination. We keep the original SPEED+ nomenclature (*Synthetic*, *Sunlamp* and *Lightbox*) to focus on the domain-inherent challenges. To differentiate the SHIRT sets, we add to them the suffix "_ROE2" to highlight their sequential nature while preserving the domain information (*Synthetic_ROE2*, *Lightbox_ROE2*).

Dataset	SPEED+ [45]			SHIRT [58]: ROE2	
Orig. Set Name	*Synthetic*	*Sunlamp*	*Lightbox*	*Lightbox*	*Synthetic*
Pose distr.	Random SE(3) Sampling			Uniform sampling over trajectory	
Pose labels	Ground-Truth	Realistic Pseudo-labels			Ground-Truth
Target	Simplified CAD	Realistic mockup (HIL)			Simplified CAD
Illumination	OpenGL-based	Direct (Sun)	Diffuse (Albedo)		OpenGL-based
# images	59,960	2,791	6,740	2,371	
Examples					
Thesis Nomenclature	*Synthetic*	*Sunlamp*	*Lightbox*	*Lightbox_ROE*	*Synthetic_ROE*

4.2.2 SPEED+

Designed to support the development and evaluation of spacecraft pose estimation methods that are robust against the domain shift described in Section 2.3.1, SPEED+ [45] contains 3 image sets of different natures. The *Synthetic* set is made of 59,960 OpenGL-rendered images generated from poses randomly sampled in $SE(3)$. The first HIL set, *Sunlamp*, includes 2,791 images acquired in the TRON facility under direct illumination provided by a metal-halide arc lamp emulating sunlight. The second HIL set, *Lightbox*, contains images captured under diffuse illumination conditions designed to reproduce the effect of Earth's albedo.

4.2.3 SHIRT

Being primarily dedicated to the evaluation of navigation filters designed to temporally aggregate the pose estimates, SHIRT [58] contains two trajectories defined by their ROE. Each ROE (Relative Orbital Elements) describes the relative motion of a spacecraft relative to another. The first ROE consists in a v-bar hold position where the chaser spacecraft holds its position while the target is tumbling around one of its principal axes. The second ROE describes a slow approach of the chaser where the target is tumbling around two of its principal axes. As the first ROE brings little viewpoint diversity compared to the second, it is not considered in this thesis.

4.2.4 Adopted Nomenclature

Because SPEED+ contains single-frame domains while SHIRT provides sequential trajectories with overlapping illumination conditions, a unified nomenclature is required to avoid ambiguity when referring to image sets in the experimental chapters. This nomenclature is described in Table 4.1. Image sets from SPEED+ are referred to directly by their illumination domains, *i.e.*, *Synthetic*, *Lightbox*, and *Sunlamp*. These names are unambiguous because SPEED+ consists exclusively of single-frame images grouped by illumination conditions. Regarding SHIRT's ROE2, we adopt the notation *Synthetic_ROE2* and *Lightbox_ROE2* to highlight that their images are generated as the second trajectory of SHIRT under either synthetic or diffuse (*Lightbox*) illumination conditions. This unified nomenclature distinguishes trajectories from single-frame domains while preserving the informative domain names.

4.2.5 Limitations

Although SPEED+ [45] and SHIRT [58] are the best spacecraft pose estimation datasets currently available due to their operational relevance and reliance on a HIL setup, they suffer from four main limitations.

First, neither SPEED+ nor SHIRT provide the **target CAD model**. As a result, evaluating the geometric accuracy of a reconstructed three-dimensional model, such as the one developed in Chapter 5, is difficult. Moreover, CAD-based data augmentation methods cannot be considered to mitigate the domain gap.

Second, SHIRT is limited to **only two image sequences** corresponding to distinct operational scenarios. This restriction prevents an extensive evaluation of the performance and robustness of a navigation filter on a wide range of approach configurations.

Third, despite the realism of HIL imagery, these datasets do not capture all **illumination phenomena affecting on-orbit images**. In particular, the absence of atmospheric diffusion, which cannot be reproduced on ground, significantly impacts the visual artifacts observed in real flight data. Similarly, sensor non-idealities such as noise, saturation, or acquisition failures are not modeled, although they may have a non-negligible impact on estimation accuracy.

Fourth, both SPEED+ and SHIRT depict a **single satellite with a fixed geometry**. Consequently, the transposability of results obtained on these datasets to targets with different geometries, such as axisymmetric upper-stage rockets, remains questionable.

Despite these limitations, SPEED+ and SHIRT currently represent the most suitable publicly available datasets. We expect that these shortcomings will be progressively alleviated in the future, for example through forthcoming IOD missions.

4.3 Pose Estimation Metrics

In this thesis, we rely on the pose estimation metrics defined for the SPEED+ dataset [45]. Three dataset-level metrics are considered: the average translation and rotation errors, $E^{(\mathrm{T})}$ and $E^{(\mathrm{R})}$, and a pose score that captures both position and rotation errors, $S^{(\mathrm{P})}$. Each metric is computed over a whole set of N image-label pairs.

As explained in Section 2.2.4, any pose estimation network F_Θ estimates from an image I_i the relative orientation as a quaternion $\mathbf{q}_i$ and the relative position as a translation vector $\mathbf{t}_i$. It therefore predicts an estimated quaternion $\hat{\mathbf{q}}_i$ and an estimated translation $\hat{\mathbf{t}}_i$. The average translation and rotation errors, $E^{(\mathrm{T})}$ and $E^{(\mathrm{R})}$, are computed as the average of the image-wise translation and rotation errors, $e_i^{(\mathrm{T})}$ and $e_i^{(\mathrm{R})}$ over the test set, *i.e.*,

$$E^{(\mathrm{T})} = \frac{1}{N}\sum_{i=1}^{N} e_i^{(\mathrm{T})}, \quad \text{and} \quad E^{(\mathrm{R})} = \frac{1}{N}\sum_{i=1}^{N} e_i^{(\mathrm{R})}. \tag{4.1}$$

For each image I_i in the set, the translation error, $e_i^{(\mathrm{T})}$, is computed as the norm of the error between the predicted and ground-truth positions, $\hat{t}_i$ and t_i, respectively, *i.e.*,

$$e_i^{(\mathrm{T})} = \left|\left|\hat{\mathbf{t}}_i - \mathbf{t}_i\right|\right|_2, \tag{4.2}$$

The rotation error, $e_i^{(\mathrm{R})}$ is the angular error between the predicted and ground-truth quaternions, $\hat{q}_i$ and q_i, respectively, *i.e.*,

$$e_i^{(\mathrm{R})} = 2\arccos\left(\left\langle \hat{\mathbf{q}}_i, \mathbf{q}_i \right\rangle\right). \tag{4.3}$$

The pose score $S^{(\mathrm{P})}$ averages over the whole set the sum of the image-wise normalized translation and orientation errors corrected to take into account the calibration uncertainty of the HIL setup ($\tilde{e}_i^{(\mathrm{T})}$ and $\tilde{e}_i^{(\mathrm{R})}$), *i.e.*,

$$S^{(\mathrm{P})} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{e}_i^{(\mathrm{T})} + \tilde{e}_i^{(\mathrm{R})}\right). \tag{4.4}$$

Indeed, since the ground-truth positions and rotations on the HIL domains are not perfectly accurate, the estimates are considered as perfect if they are below a certain threshold, *i.e.*,

$$\tilde{e}_i^{(\mathrm{T})} = \begin{cases} 0, & \text{if } e_i^{(\mathrm{T})} < \left|\left|\mathbf{t}_i\right|\right|_2 \cdot \tau^{(\mathrm{T})} \\ \frac{e_i^{(\mathrm{T})}}{\left|\left|\mathbf{t}_i\right|\right|_2}, & \text{otherwise,} \end{cases} \tag{4.5}$$

and,

$$\tilde{e}_i^{(\mathrm{R})} = \begin{cases} 0, & \text{if } e_i^{(\mathrm{R})} < \tau^{(\mathrm{R})} \\ e_i^{(\mathrm{R})}, & \text{otherwise,} \end{cases} \tag{4.6}$$

For the HIL sets, these thresholds $\tau^{(\mathrm{T})}$ and $\tau^{(\mathrm{R})}$ are set to 2.173 mm/m

and 0.00295 radians, respectively. This corresponds to the calibration uncertainty of the HIL setup used to generate these sets. For the synthetic sets, these thresholds are set to zero as there is no uncertainty on the pose labels.

This definition of the pose score relies on an arbitrary weighting made by the SPEED+ [45] authors to represent both angular and translational errors through a single metric. It implicitly implies that a 1° angular error has the same impact as a $17.5 cm/m$ normalized translation error. This pose score metric, although it results from an arbitrary choice, is nevertheless accepted within the community and is therefore followed in this thesis. However, we point out that future works should consider how the weighting between both components affects the pose estimation quality as well as novel metric formulations.

4.4 Image Quality Metrics

As explained in the introduction, this thesis exploits Novel View Synthesis (NVS) models to address the limitations encountered by current spacecraft pose estimation methods. Since our contributions rely on image synthesis tools, we introduce in this section the image quality metrics used throughout the thesis. Sections 4.4.1 to 4.4.3 respectively describe the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Just Noticeable Blur (JNB) metrics and discuss their relevance. The PSNR and SSIM are full-reference metrics, meaning that they are computed using a pair of reference and reconstructed images, while the JNB score is a no-reference metric, *i.e.*, it is directly computed on the reconstructed image, without requiring a reference image.

4.4.1 PSNR

The Peak Signal-to-Noise Ratio (PSNR) is a full-reference image quality metric based on the Mean Squared Error (MSE) between a reference image I and a reconstructed image I' of W pixels by H pixels. The MSE is defined as the average of the squared pixel-wise error computed between the reconstructed and reference images, *i.e.*,

$$\text{MSE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \left(I_{ij} - I'_{ij} \right)^2 . \tag{4.7}$$

The PSNR captures this average pixel-wise error through a logarithmic scale that is more convenient to interpret. Hence, it is defined as

$$\text{PSNR} = 10\log_{10}\left(\frac{L^2}{\text{MSE}}\right), \tag{4.8}$$

where L denotes the maximum possible pixel intensity, *e.g.*, 255 for an image with an 8-bit depth.

The PSNR is therefore a simple metric that measures absolute pixel-wise fidelity and strongly penalizes large numerical deviations. However, it is not aligned with the human visual system, as it does not account for spatial structure or perceived quality.

4.4.2 SSIM

Unlike the PSNR, the Structural Similarity Index Measure (SSIM) [133] is a perceptually motivated full-reference metric that compares images through their luminance, contrast, and structural statistics.

Let μ_I and $\mu_{I'}$ be the mean intensities of the reference and reconstructed images,

$$\mu_I = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} I_{ij}, \qquad \mu_{I'} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} I'_{ij}, \tag{4.9}$$

and σ_I^2 and $\sigma_{I'}^2$ be their corresponding variances,

$$\sigma_I^2 = \frac{1}{WH-1}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(I_{ij}-\mu_I\right)^2, \tag{4.10}$$

$$\sigma_{I'}^2 = \frac{1}{WH-1}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(I'_{ij}-\mu_{I'}\right)^2, \tag{4.11}$$

and $\sigma_{II'}$ be the covariance between these images,

$$\sigma_{II'} = \frac{1}{WH-1}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(I_{ij}-\mu_I\right)\left(I'_{ij}-\mu_{I'}\right). \tag{4.12}$$

The SSIM index aggregates three factors corresponding to the luminance ($l(I, I')$), contrast ($c(I, I')$) and structural ($s(I, I')$) statistics, weighted by

some fixed constants α, β, γ, *i.e.*,

$$\text{SSIM}(I, I') = l(I, I')^{\alpha}\, c(I, I')^{\beta}\, s(I, I')^{\gamma}. \tag{4.13}$$

The luminance and contrast factors are defined as:

$$l(I, I') = \frac{2\mu_I \mu_{I'} + C_1}{\mu_I^2 + \mu_{I'}^2 + C_1} \qquad \text{and} \qquad c(I, I') = \frac{2\sigma_I \sigma_{I'} + C_2}{\sigma_I^2 + \sigma_{I'}^2 + C_2} \tag{4.14}$$

They respectively capture differences in average brightness and local contrast through the image means and variances. The structural factor is defined as:

$$s(I, I') = \frac{\sigma_{II'} + C_3}{\sigma_I \sigma_{I'} + C_3}, \tag{4.15}$$

and captures the spatial organization of local patterns within both images.

In the previous expressions, C_1, C_2 and C_3 are positive constants used to stabilize the division and are defined in the original paper as $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, and $C_3 = (K_2 L)^2/2$, with the constants $K_1 = 0.01$ and $K_2 = 0.03$ being empirically defined. This leads to the final SSIM formula:

$$\text{SSIM}(I, I') = \frac{(2\mu_I \mu_{I'} + C_1)\,(2\sigma_{II'} + C_2)}{(\mu_I^2 + \mu_{I'}^2 + C_1)\,(\sigma_I^2 + \sigma_{I'}^2 + C_2)}, \tag{4.16}$$

As a conclusion, the SSIM improves over the PSNR by comparing images through perceptually relevant components such as luminance, contrast, and structural similarity, rather than relying solely on pixel-wise error. However, it still suffers from some limitations. For example, it may assign high scores to uniformly blurred images whose local structural statistics are preserved.

4.4.3 JNB Score

Unlike the PSNR and SSIM, the Just Noticeable Blur (JNB) score [134] is a no-reference image quality metric that is specifically designed to assess the sharpness of a synthesized image. The metric estimates whether the amount of blur at an image edge exceeds a perceptual threshold beyond which blur becomes noticeable to a human observer. The computation of the JNB metric is out of the scope of this thesis. Hence, in this section, we only present the conceptual approach behind the metric. For further details, the reader is referred to the original paper [134].

As the metric focuses on edges, the image is first processed through a Sobel filter to detect them. The metric is computed on N_b non-overlapping blocks of 64×64 pixels. These N_b blocks correspond to the image blocks for which the number of edge pixels within the block is larger than some pre-defined threshold τ. The set of pixels that correspond to edges inside a block i is defined as $\mathcal{E}_i$.

For each edge block i, we first compute a noticeable blur threshold w_i^JNB that depends on the local luminance and contrast within the image block. Then, for each edge pixel e within that block, we compute the local blur $w(e)$, that is the extent of the transition between both sides of the edge. Then, for each edge pixel, we compute the difference between the observed blur and the threshold. Positive differences are then summed over the block. Finally, the JNB score is averaged across the N_b blocks.

Formally, the JNB score is therefore computed as:

$$\text{JNB} = \frac{1}{N_\text{b}} \sum_{i=1}^{N_\text{b}} \frac{1}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} \max\left(0, w(e) - w_i^\text{JNB}\right). \tag{4.17}$$

In summary, the JNB metric provides a sharpness-oriented evaluation by focusing on whether blur is noticeable to a human observer at image edges. A larger JNB score indicates increased blur, whereas a smaller score corresponds to sharper image edges. By targeting perceptually visible blur rather than overall similarity or signal accuracy, it complements PSNR and SSIM and is particularly relevant when assessing blur-related degradation in synthesized images.

4.5 Closing Remarks

This chapter introduced the publicly available datasets relevant to spacecraft pose estimation and justified the selection of SPEED+ and SHIRT for this thesis because of their complementary nature. Although they suffer from some limitations, such as the lack of real images taken in orbit, they combine synthetic renderings, realistic HIL acquisitions, and trajectory-based sequences so that they provide the diversity required for our experiments. Together with the pose estimation and image synthesis evaluation metrics, these datasets provide the empirical framework used throughout the thesis to assess how the proposed NVS-based methods address the limitations of current spacecraft pose estimation methods.

This concludes the first part of this thesis which was dedicated to the foundations of our work: spacecraft pose estimation networks (Chapter 2), NVS-representations such as NeRFs (Chapter 3), spacecraft pose estimation datasets and relevant metrics (Chapter 4).

PART II
Contributions

5

3D Model Reconstruction from 2D Monocular Images

AUTONOMOUS Rendezvous & Proximity Operations (RPOs) are becoming increasingly important for future On-Orbit Servicing (OOS) and Active Debris Removal (ADR) missions, as explained in Chapter 1. In this context, reconstructing a 3D model of the target from close-range monocular imagery can significantly support tasks such as grasp-point selection [135, 136], pose estimation [137] or AI-based GNC algorithm training or validation [6].

In this chapter, we focus on the practical difficulties that arise when 3D model reconstruction methods are applied to on-orbit imagery with an emphasis on the illumination variability over time and the epistemic uncertainty affecting the pose estimates.

While we exploit this method in Chapter 6 to enable the training of a pose estimation network from a small set of images depicting its target, the internal representation of future Simultaneous Localization and Mapping (SLAM) systems might also build on our proposed reconstruction method.

The content of this chapter is adapted from our paper *"NeRF-based Spacecraft Reconstruction from Close-Range Monocular Imagery Under Illumination Variability and Pose Uncertainty"* [1], currently under review, to align with the overall structure and objectives of the thesis.

5.1 Introduction

3D reconstruction from 2D images has recently attracted significant attention within the computer vision community. In particular, the Novel View Synthesis (NVS) methods introduced in Chapter 3 have rapidly matured. Their ability to infer a 3D-aware scene representation from a set of 2D monocular images has enabled a wide range of applications. While previous works explored the use of NVS-based 3D model reconstruction methods in the context of RPOs [137–139], they validated their methods on over-controlled benchmarks which do not explicitly address two key challenges that arise in realistic operational scenarios.

First, they did not consider the **challenging illumination conditions** that occur in orbit. Apart from the Earth albedo, which dimly contributes to the spacecraft illumination, the target is directly illuminated by the Sun. Due to the lack of atmospheric diffusion, only the spacecraft parts that are lit by the Sun are visible. Furthermore, those parts may be over-exposed because of the specular reflections that are typical of the materials used to build the spacecraft. An additional challenge related to the illumination conditions is their variability. Indeed, due to the long times required to safely approach the target spacecraft, both the chaser and the target move significantly in their orbits. As a result, the illumination conditions may dramatically vary during the operation.

Second, NVS-based reconstruction methods rely on a set of images along their relative poses. Previous works assumed a perfect knowledge of these pose labels. However, in a real use case, they are unknown and can only be estimated, *e.g.*, through human annotations. This **uncertainty on the 6D pose labels** results in an uncertainty on the 3D representation, and hence, in blurry 2D image reconstruction [140].

We propose an NVS-based method for recovering a 3D representation of an unknown target spacecraft from a set of 2D images depicting that target. Unlike previous works, our method is designed to handle *(i) the illumination variability* and the *(ii) pose uncertainty* that are inherent to real use cases. To handle the varying illumination conditions, our method adopts a Neural Radiance Field (NeRF) [52] with in-the-wild abilities [119], *i.e.*, an *appearance embedding* is learned for each image in order to model implicitly the photometric variations between the training images. Similarly, for each image, we introduce a *learnable pose correction term* to correct the

estimated pose label. By learning these terms along the NeRF, the labels are progressively denoised, thereby reducing the uncertainty of the 3D representation, without affecting the model complexity. The proposed method is validated on a synthetic and two Hardware-In-the-Loop (HIL) sets that correspond to different orbital lighting scenarios.

This chapter is structured as follows. Section 5.2 positions our work with regard to prior work and highlights their limitations. Section 5.3 presents our method for reconstructing a 3D model of an unknown target spacecraft from monocular images taken under different viewpoints with an emphasis on the handling of the illumination variability as well as the pose labels uncertainty. Section 5.4 validates our method and its key components. Finally, Section 5.5 concludes.

5.2 Related Works

Offline 3D model reconstruction from monocular images aims to recover a 3D representation of an object given a set of images and their associated pose labels. Numerous studies have addressed this task in the context of orbital rendezvous and proximity operations. This section summarizes these works and highlights how the proposed approach overcomes their limitations.

NVS-based methods have been widely adopted for this task. As explained in Chapter 3, these methods learn a representation optimized to reconstruct the available images, which can then be used to render novel views of the target from unseen viewpoints. NVS approaches rely on either implicit or explicit representations. Implicit models describe the scene as a continuous function, whereas explicit models represent it using geometric primitives. Consequently, implicit methods synthesize images via volumetric integration, while explicit methods use rasterization.

Implicit representations, such as NeRFs [52], were the first considered for NVS-based 3D reconstruction. Mergy *et al*. [138] pioneered this direction by demonstrating NeRF's ability to learn spacecraft geometry from monocular images. They also showed that training from unposed images is possible using Generative Radiance Fields (GRAFs) [141], although at a prohibitive computational cost due to an adversarial formulation. Subsequent works [142–145] leveraged advanced NeRF variants such as Instant-NGP [118] and D-NeRF [146], but their evaluations were limited to synthetic datasets or images captured under smooth illumination, neglecting the chal-

lenges of orbital lighting. Other studies addressed issues such as RGB-D input [147] or low-light and blurry on-orbit imagery [148], yet they still relied on synthetic data that poorly reflects real illumination conditions. In contrast, our method is specifically tailored for on-orbit lighting and is evaluated on images exhibiting these conditions.

Explicit representations have recently gained attention for this task. Several works [139, 149–151] demonstrated that 3D Gaussian Splatting [111] can reconstruct spacecraft geometry from monocular synthetic images. Similarly, De Smijter *et al*. [151] validated this capability for 3D Convex Splatting [112], which uses convex primitives instead of Gaussians. While explicit representations offer advantages such as a lower computational cost and an explicit geometric description, we adopt NeRF-based models because their field-based formulation is better suited to handle adverse illumination conditions.

Finally, several works aim to recover an abstract 3D structure of an unknown spacecraft from monocular images without relying on NVS representations [128, 152–155]. Although conceptually appealing and promising for future developments, these approaches remain insufficiently robust to the diverse illumination conditions encountered in orbit. They are also less suited than NVS-based solutions in dealing with rendering-based tasks. That is, NVS-based models enable rendering novel viewpoints, facilitating inspection and supporting the training of relative pose estimation networks [6, 7]. They can also visualize decision cues [3] or even perform pose estimation directly via inversion [137].

Although pose inaccuracies have not been extensively studied in space applications, this issue has been addressed within the computer vision community. Unlike SLAM-based approaches [156–159], which aim to reconstruct a scene from a stream of unposed images, or adversarial methods [141, 160, 161] that reconstruct scenes from unposed image sets, our objective is different: we seek to learn the scene from images whose relative pose labels are slightly inaccurate. Inspired by prior works such as I-NeRF [162], Parallel Inversion [163], and BARF [164], our method leverages the differentiability of NeRFs to optimize pose correction terms applied to the inaccurate labels. This mechanism introduces no additional complexity while significantly enhancing reconstruction sharpness.

5.3 Method

This section describes our method for recovering an implicit representation of an uncooperative, unknown, target spacecraft from a set of monocular images depicting that target in orbit. Sections 5.3.1 and 5.3.2 describe the problem and present an overview of the method while Section 5.3.3 details a key element of our method, namely, the fine-tuning of the pose labels through learned correction terms.

5.3.1 Problem Statement

Figure 5.1 illustrates the problem of recovering an implicit representation M_Φ of an unknown target spacecraft from a set of N monocular images $\{I_a\}_{a=1}^{N}$ taken under different viewpoints $\{(q,t)_a\}_{a=1}^{N}$, and different orbital lighting conditions. Such a representation can then be used to generate novel views of the target taken under unseen viewpoints. While this problem has partially been addressed [138], *e.g.*, through Neural Radiance Fields (see Section 3.3), two issues that arise on real use cases have not been addressed yet.

First, an underlying conventional assumption of implicit representations is the constant nature of the lighting conditions. All the training images are expected to be taken under similar illumination conditions. However, illumination conditions may abruptly and significantly change in orbit. For example, in Figure 5.1, images I_6 and I_7 depict the same two perpendicular faces of the target. However, only one face is illuminated in each image while the other is completely under-exposed. These discrepancies between the training images seriously affect the training of the NeRF, and consequently, the quality of the model reconstructed by such an illumination-agnostic NeRF.

Second, NeRFs are usually trained on a set of images paired with their corresponding relative poses $(q,t)_a$, *i.e.*, orientation q and position t. However, in a real use case, the ground-truth positions and orientations are unknown, only the captured images are available. Relative poses can be estimated photogrammetrically, for example through COLMAP [165], but the noise affecting these estimates remains an impediment. This epistemic uncertainty on the pose labels results in an inconsistent 3D representation. When such a representation is rendered, this leads to blur in the generated images.

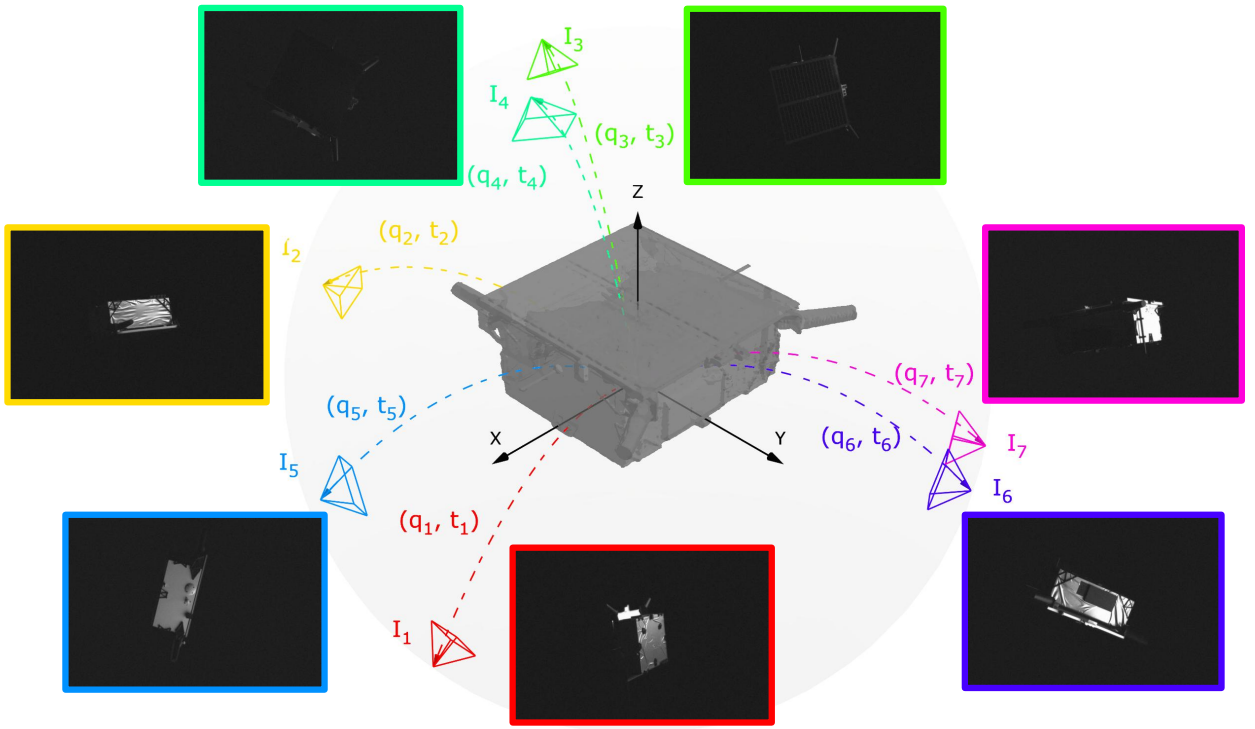


Fig. 5.1 3D reconstruction from 2D on-orbit imagery: problem overview. Our method learns a 3D, implicit, representation of a target spacecraft from a set of images taken under different viewpoints. The challenges are two-fold. First, due to the orbital lighting conditions, two images taken under close viewpoints may exhibit drastically different illuminations. Second, the relative pose between the camera and the target are unknown and can only be roughly estimated.

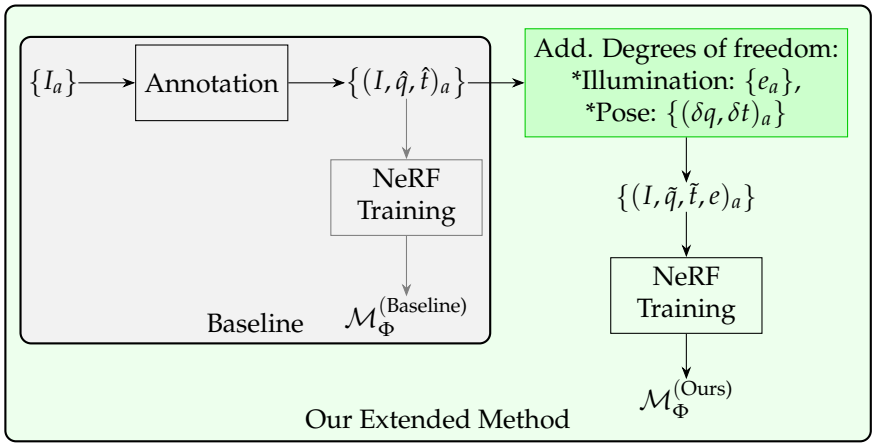


Fig. 5.2 Learning 3D implicit representation: method overview. Our method extends the usual NeRF training paradigm by adding degrees of freedom to the model. While per-image appearance embeddings $\{e_a\}_{a=1}^N$ are learned to capture photometric variations between images taken under diverse illumination conditions, per-image pose correction terms $\{(\delta q, \delta t)_a\}_{a=1}^N$ are learned to refine the roughly annotated pose labels.

5.3.2 Method Overview

As depicted in Figure 5.2, our method recovers an implicit representation M_Φ from a set of images $\{I_a\}_{a=1}^N$ by extending the NeRF-based paradigm. This extension addresses the two previously highlighted challenges, namely, the inability of standard NeRFs to model varying and harsh lighting conditions, as well as the 3D uncertainty inherent with the use of estimates of the relative poses of the training images. Both challenges are addressed through a conceptually simple yet effective mechanism, which consists in increasing the degrees of freedom of the neural radiance field, with regard to both illumination variations as well as pose labels uncertainty.

To model the photometric variations between images taken under similar viewpoints but different illumination conditions, we rely on appearance embeddings [119]. Each embedding is associated with a given image and implicitly captures the illumination conditions that are specific to that image. These additional, per-image, degrees of freedom allow the model to handle challenging and varying illumination conditions.

Similarly, as detailed in the next section, to mitigate the impact of the pose labels uncertainty, we learn for each image a pose correction term. This

term, which encompasses the correction of both the camera orientation and position, is learned along the representation and is able to refine the coarse pose estimate. This reduces pose uncertainty and improves the consistency of the 3D representation, resulting in sharper renderings.

5.3.3 Mitigating Pose Uncertainty Trough Learnable Correction Terms

The Neural Radiance Field M_Φ should normally be trained on ground-truth, *i.e.*, perfectly accurate, pose labels, $\{(q_a, \boldsymbol{t}_a)\}_{a=1}^N$. However, due to errors inherent to human annotations, the NeRF is trained on pseudo-pose labels $\{(\hat{q}_a, \hat{\boldsymbol{t}}_a)\}_{a=1}^N$. The epistemic uncertainty on the pose labels results in an uncertainty on the 3D reconstruction [140], which is ultimately responsible for a blur observed in the images generated through this model.

To alleviate this uncertainty on the pose labels and therefore recover a sharper representation, we do not train the NeRF on the pseudo-labels directly but on refined pose labels, $\{(\tilde{q}_a, \tilde{\boldsymbol{t}}_a)\}_{a=1}^N$. As illustrated in Figure 5.3, for each image I_a, the refined, learned, pose label is computed from the pseudo-label through a pose correction term, $(\delta q_a, \boldsymbol{\delta t}_a)$, learned along the NeRF training, *i.e.*,

$$\tilde{\boldsymbol{t}}_a = \hat{\boldsymbol{t}}_a + (\boldsymbol{\delta t})_a \quad \text{and} \quad \tilde{q}_a = (\delta q)_a \, \hat{q}_a \tag{5.1}$$

where $\{(\boldsymbol{\delta t})_a\}_{a=1}^N$ are learnable parameters and $\{(\delta q)_a\}_{a=1}^N$ are correction terms formulated as quaternions, with scalar parts $\{(\delta q)_a^0\}_{a=1}^N$ and vector parts $\{(\boldsymbol{\delta q})_a^{\text{vec}}\}_{a=1}^N$, *i.e.*,

$$(\delta q)_a = \begin{pmatrix} (\delta q)_a^0 \\ (\boldsymbol{\delta q})_a^{\text{vec}} \end{pmatrix} \tag{5.2}$$

where $\{(\boldsymbol{\delta q})_a^{\text{vec}}\}_{a=1}^N$ are learnable parameters.

To ensure that the quaternion represents a valid rotation, the scalar component of the quaternion correction term, $(\delta q)_a^0$, is not directly learned. Instead, it is computed based on the vector part to enforce a unit-norm, *i.e.*,

$$(\delta q)_a^0 = \sqrt{1 - \left|\left|(\boldsymbol{\delta q})_a^{\text{vec}}\right|\right|^2}. \tag{5.3}$$

These pose correction terms can be seen as additional degrees of freedom granted to the model as they give the ability of fine-tuning the pose labels to improve the representation consistency, and in turn the rendering sharpness.

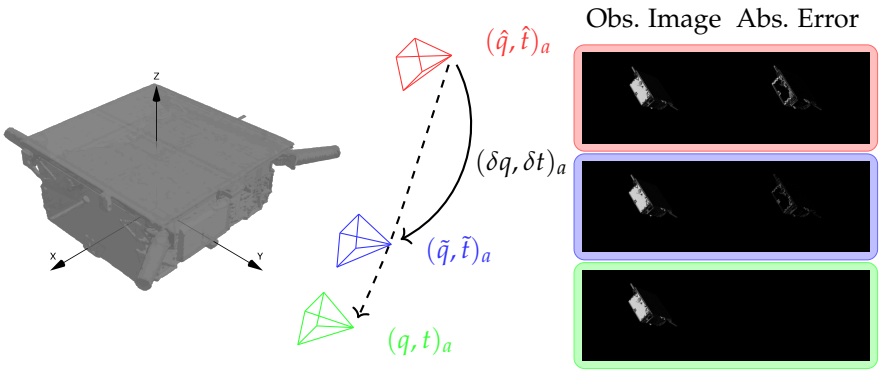


Fig. 5.3 Solving pose misalignment through learnable correction terms. (Exaggerated for visualization.) The optimization goal of reducing the photometric error between the generated image and the observed one encourages the model to progressively refine the pose label.

Indeed, as illustrated in Figure 5.3, any misalignment of the pose labels results in an increased photometric error. Hence, the model, trained on the photometric loss, is encouraged to refine the pose labels as it contributes to reducing the photometric loss.

In practice, as explained in Chapter 3, neural radiance fields are not directly trained on images paired with their pose labels but rather on batches made of pixels, and their corresponding rays, sampled across all the training images to promote multi-view consistency and hence improve convergence. Therefore, we do not apply the pose correction terms directly on the pose labels but on the rays as it is equivalent but easier to integrate in a NeRF.

The rays computed from the pseudo-pose labels ($\hat{\boldsymbol{c}}_{aij}$,$\hat{\boldsymbol{d}}_{aij}$), *i.e.*, ray origin $\hat{\boldsymbol{c}}_{aij}$ and direction $\hat{\boldsymbol{d}}_{aij}$ as computed using the pseudo-pose labels ($\hat{q}_i$, $\hat{\boldsymbol{t}}_i$), are refined through the image-wise pose correction terms into fine-tuned rays ($\tilde{\boldsymbol{c}}_{aij}$, $\tilde{\boldsymbol{d}}_{aij}$), *i.e.*,

$$\tilde{\boldsymbol{c}}_{aij} = \hat{\boldsymbol{c}}_{aij} + (\delta \boldsymbol{t})_a \quad \text{and} \quad \begin{pmatrix} 0 \\ \tilde{\boldsymbol{d}}_{aij} \end{pmatrix} = (\delta q)_a \otimes \begin{pmatrix} 0 \\ \hat{\boldsymbol{d}}_{aij} \end{pmatrix} \otimes (\delta q)_a^* \tag{5.4}$$

Refining the pose labels does not significantly increase the training complexity as it is conducted during the training of the NeRF. In addition, it does not affect the inference complexity since the pose correction terms are not used to render novel views. The only hazard that arises from the

pose fine-tuning is the risk of divergence that could happen if the pose labels move away from the true camera poses during the first training steps. However, by imposing lower learning rates on the pose correction terms, this risk is reduced to a negligible level and has never been observed, *i.e.*,

$$l_r^{(\delta t)} = \beta^{(\delta t)}\, l_r \quad \text{and} \quad l_r^{(\delta q)} = \beta^{(\delta q)}\, l_r, \tag{5.5}$$

where $l_r^{(\delta t)}$ and $l_r^{(\delta q)}$ are the learning rates for the position and orientation correction terms, $\beta^{(\delta t)}$ and $\beta^{(\delta q)}$ are factors lower or equal to 0.01 to ensure convergence, and l_r is the NeRF learning rate.

5.4 Experiments

This section describes and analyzes the experiments conducted to evaluate the proposed 3D reconstruction method. Section 5.4.1 describes our validation setup while Section 5.4.2 evaluates the benefits brought by our method compared to a standard, off-the-shelf, NeRF. Section 5.4.3 assesses our method's resilience to a limited number of images. An ablation study, conducted on the proposed pose correction mechanism (see Section 5.3.3), is carried out in Section 5.4.4.

5.4.1 Validation Setup

This section describes the training and evaluation setup used to assess our method. In the following, "baseline" refers to a model trained without appearance embeddings nor pose correction terms, as opposed to our model.

Datasets

The three datasets used in this chapter, *Lightbox_ROE2*, *Sunlamp* and *Synthetic_ROE2*, have been described in Chapter 4 and are summarized in Table 4.1. *Lightbox_ROE2* and *Synthetic_ROE2* share the same viewpoint distribution, apart from calibration errors, as they consist in images taken by a chaser spacecraft during a slow approach towards a tumbling target. Nevertheless, the visual appearance of this target differs significantly. While images of *Lightbox_ROE2* were acquired in a testbed that mimics the illumination conditions encountered in orbit, those contained in *Synthetic_ROE2* were synthetically generated. The key differences between both sets are

therefore the realistic illumination conditions as well as the pose uncertainty, which are the specific challenges addressed by our method. Finally, *Sunlamp* consists in images taken under even more challenging illumination conditions. Its acquisition setup emulates the direct illumination from the Sun on a specular target. The *Sunlamp* images are therefore of poorer quality and are characterized by under as well as over-exposure, including sun flares.

These three sets serve different purposes. *Lightbox_ROE2* is representative of a real use case and allows us to validate the ability of our method to handle varying illumination conditions as well as pose uncertainty. Similarly, *Sunlamp* is representative of the most challenging illumination conditions that can be encountered on-orbit. By demonstrating the ability of our method in handling such challenging illumination conditions, we show that our method is robust to a wide range of illumination scenarios. Finally, even if the *Synthetic_ROE2* set is over-simplified compared to the real ones, it offers a controllable environment to perform ablation studies on the pose fine-tuning abilities of our method. This allows us to demonstrate the ability of our method to cope with a wide range of noise on the pseudo-pose labels.

Training

Unless otherwise mentioned, each model is trained on 400 images randomly sampled in the considered set. All images are preprocessed to remove the background using foreground segmentation masks. Each model, along with the 400 appearance embeddings and pose correction terms, is trained for 30,000 optimization steps of 4096 rays per batch. The training takes 30 minutes on an NVIDIA L40s.

Evaluation

Since the 3D model of TANGO is not available, direct evaluation of the accuracy of the learned 3D representation is infeasible. Instead, we evaluate the 2D image reconstruction quality as a proxy of the 3D accuracy. For this purpose, we randomly sample 200 pose labels, along with the corresponding images, from a realistic set and evaluate the ability of our model in reconstructing the images from the pose labels. The *Lightbox* set (see Chapter 4) was chosen for this purpose as it contains realistic images of the same target under viewpoints and illumination conditions that are different from the ones encountered in the different training sets.

This photometric evaluation suffers from an inherent ambiguity. Indeed, the same object viewed under the same pose corresponds to multiple images, depending on the illumination conditions. We solve this ambiguity through

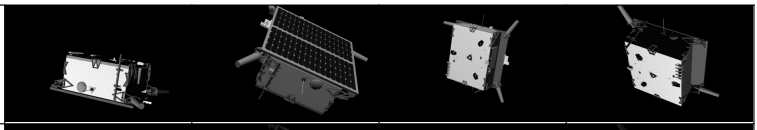
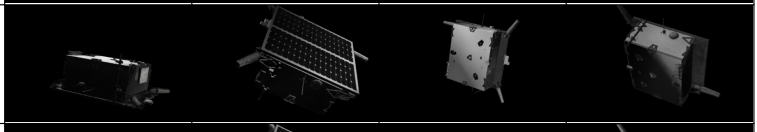
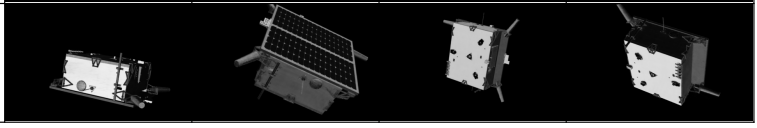
a widely accepted practice in the novel view synthesis community [119]. For each generated image, we first train an appearance embedding using half of the reference image pixels to implicitly learn the illumination conditions. This embedding is then used to generate the full image using illumination conditions aligned with the reference image, thereby enabling a valid evaluation of the reconstruction quality.

Each generated image is then compared to the reference one through qualitative as well as quantitative assessment. For each pair of images, we compute the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [133] and Just Noticeable Blur (JNB) score [134]. As described in Section 4.4, PSNR and SSIM are full-reference image quality metrics while the JNB score is a no-reference metric which quantifies the amount of blur in an image, and hence, the sharpness of its content. Unless otherwise mentioned, the values correspond to the metric averaged over the 200 test images.

5.4.2 Validation

This section compares the image reconstruction quality of our method against the baseline. Table 5.1 shows images generated from identical viewpoints for both models trained on synthetic images with accurate pose labels. Both approaches successfully capture the 3D geometry of the target spacecraft. However, while our method accurately handles illumination conditions, the baseline exhibits artifacts such as shadows in the rendered

Table 5.1 Qualitative assessment of our reconstruction method applied on *Synthetic_ROE2* images, compared to the baseline reconstruction. Both methods perform well on synthetic data, although the baseline struggles to handle illumination variations.

Reference	
Baseline	
Ours	

images. This improvement stems from our model's ability to decouple scene geometry from image-specific illumination.

Across all image pairs, reconstructions produced by our method consistently achieve higher PSNR and SSIM values than those generated by the baseline. Over 200 images, our method reaches an average PSNR of 30.3 dB, compared to 22.0 dB for the baseline, confirming a substantial quality improvement. The same trend holds for SSIM: 0.948 for our method versus 0.927 for the baseline. A key factor in this illumination consistency is the use of appearance embeddings, which align lighting conditions between reference and generated images, thereby improving perceptual quality.

Table 5.2 presents results on the *Lightbox_ROE2* dataset, which better reflects real mission conditions, including orbital lighting and pose uncertainty. Our method successfully reconstructs both geometry and appearance, whereas the baseline struggles with illumination, producing averaged lighting effects. Incorporating appearance embeddings improves PSNR from 24.35 dB to 30.90 dB and SSIM from 0.944 to 0.952.

Fine-tuning the pose labels further enhances representation sharpness and image detail. For example, solar panels reconstructed by the baseline appear texture-less, while our method recovers structural details and even

Table 5.2 Qualitative assessment of our reconstruction method, with and without pose fine-tuning, applied on *Lightbox_ROE2* images, compared to the baseline reconstruction. While the baseline struggles to handle varying illumination, our method achieves it. In addition, pose fine-tuning enables the learning of a sharper representation as it reduces the 3D uncertainty.

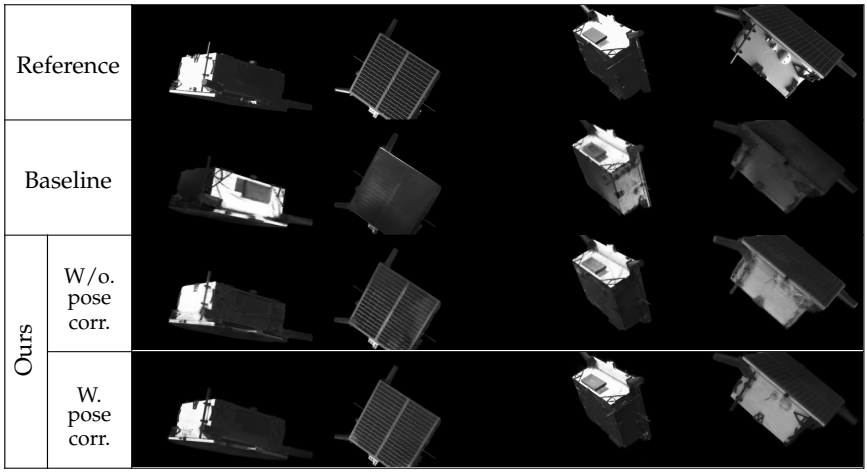


Table 5.3 Enlargement of the images depicted in Table 5.2 to highlight their high-frequency content. Fine-tuning the pose labels results in a sharper representation, which finally lowers the level of blur on the generated images.

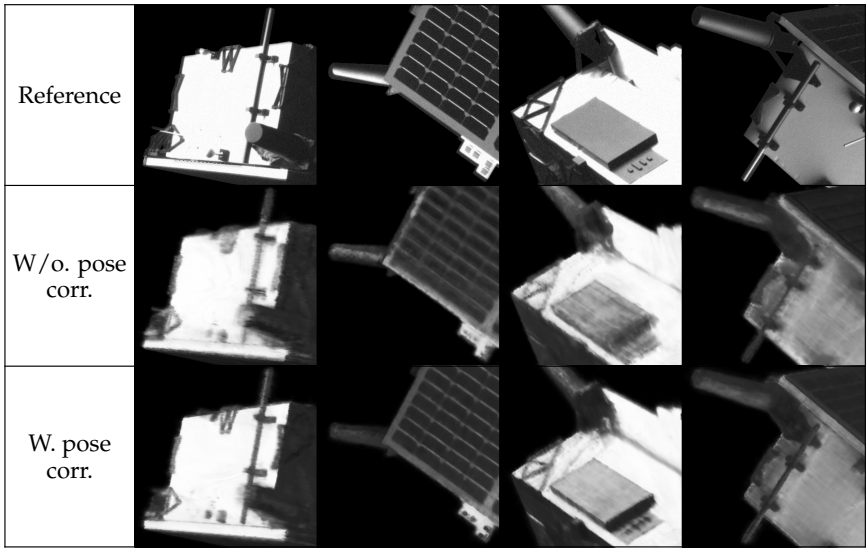
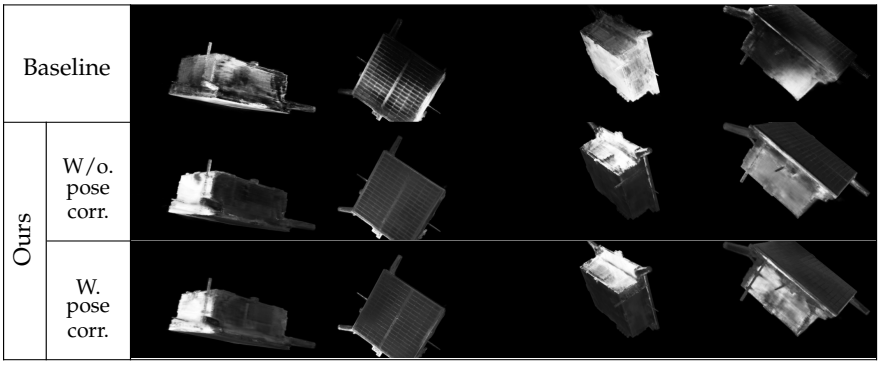


Table 5.4 Qualitative assessment of our reconstruction method, with and without pose correction, applied on *Sunlamp* images, compared to the baseline reconstruction. The baseline fails to capture the illumination conditions while our method is able to cope with the challenging illumination conditions encountered in *Sunlamp*. Furthermore, correcting the pose labels improves the representation's sharpness.



individual solar cells. When combined with pose correction, these details become sharper, as illustrated in Table 5.3, which magnifies high-frequency content, *i.e.*, image details. Features such as solar cells, torque rods, antennas, and Multi-Layer Insulation (MLI) wrinkles are significantly sharper when pose correction is applied. These capabilities are analyzed in Section 5.4.4. Using pose correction, generated images achieve an average PSNR of 30.54 dB, SSIM of 0.950, and JNB score of 0.0333. Although PSNR and SSIM slightly decrease, the JNB score improves by 5.8%, indicating a notable sharpness gain.

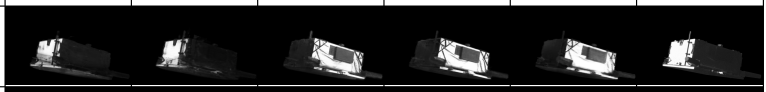
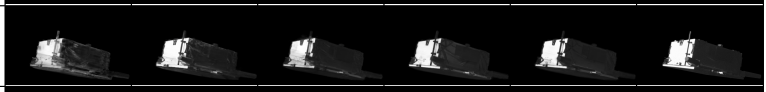
To further evaluate robustness under challenging illumination, we repeated experiments on *Sunlamp*, which simulates direct solar lighting. As shown in Table 5.4, the baseline fails to handle these conditions, producing only coarse approximations of the target shape. In contrast, our method reconstructs a consistent representation, recovering overall geometry and subtle details, though not as well as on *Lightbox_ROE2*. Quantitatively, the baseline achieves an average PSNR of 18.84 dB and SSIM of 0.932, while our method reaches an average PSNR of 27.98 dB and SSIM of 0.951.

Our pose correction mechanism on *Sunlamp* is not as effective as on *Lightbox_ROE2*, likely because its benefits depend on a sufficiently accurate underlying representation. With pose correction, the average PSNR decreases slightly, from 27.97 dB to 27.14 dB, while the SSIM remains unchanged. However, the JNB score improves from 0.0402 to 0.0389, a 3.2% reduction, consistent with previous observations on *Lightbox_ROE2*. Despite minor reductions of the image quality metrics, pose fine-tuning enhances image sharpness.

5.4.3 Sensitivity to Training Image Count

Table 5.5 illustrates images generated by both the baseline and our method when trained with varying numbers of *Lightbox_ROE2* images. Both approaches reconstruct geometrically accurate models even with fewer training images. However, while our method preserves the target spacecraft's appearance under limited data, the baseline struggles as the image count decreases. Overall, image quality for our method degrades moderately with fewer training images: the average PSNR, and SSIM drop from 30.54 dB and 0.950 at 400 training images to 27.46 dB and 0.933 at 25 images. The JNB score evolves from 0.0337 to 0.0496, which indicates that a reconstructed model is not as sharp when the number of images decreases. This con-

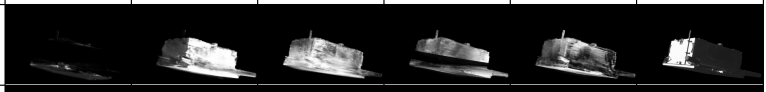
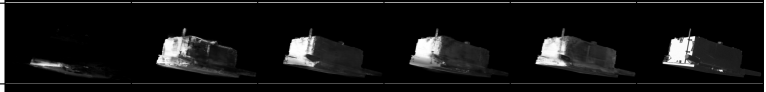
Table 5.5 Qualitative assessment of our reconstruction method applied on *Lightbox_ROE2* images, compared to the baseline reconstruction, for different number of training images. Our method is more robust to limited training data.

# images	25	50	100	200	400	Reference
Baseline						
Ours						
Baseline						
Ours						

firms that performance declines with reduced data, though the degradation remains reasonable.

A similar trend is observed on *Sunlamp*, as illustrated in Table 5.6, where the effect of reduced image count is even more pronounced. The baseline trained on 400 images captures geometry and partially reproduces appearance. However, with fewer images, it only approximates the target shape. In contrast, our method not only performs better with 400 images but also

Table 5.6 Qualitative assessment of our reconstruction method applied on *Sunlamp* images, compared to the baseline reconstruction, for different number of training images. Our method is more resilient to scarce training sets, even on images exhibiting adverse illumination conditions.

# images	25	50	100	200	400	Reference
Baseline						
Ours						
Baseline						
Ours						

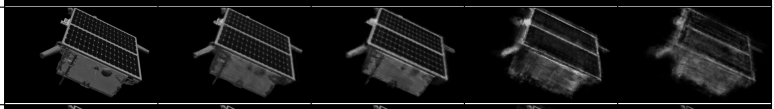
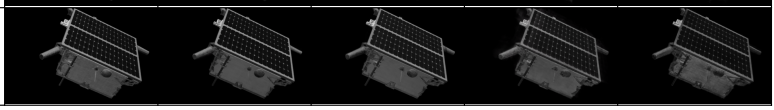
exhibits greater robustness to data scarcity. With 50 images or more, it reconstructs both geometry and appearance. Quantitatively, PSNR and SSIM decrease from 27.14 dB and 0.950 at 400 images to 26.68 dB and 0.946 at 50 images. The JNB score increases from 0.0389 at 400 images to 0.0440 at 50 images. This is consistent with the observations made on *Lightbox_ROE2*. While reconstruction quality diminishes as image count decreases, this degradation remains limited.

5.4.4 Ablation Study: Pose Correction

To further assess the effectiveness of the proposed pose correction mechanism and its robustness across varying noise levels, we conduct an ablation study. For this purpose, we use the *Synthetic_ROE2* dataset, which contains noise-free pose labels, enabling a controlled evaluation of noise impact on reconstruction quality. In the following experiments, we introduce synthetic noise to pose labels and train our method with and without pose correction to quantify its influence on reconstruction accuracy.

Table 5.7 shows images generated from models trained on labels corrupted with orientation noise, with and without pose correction. Reconstructions without pose correction appear blurry, whereas those trained

Table 5.7 Images generated by our model trained either using pseudo-pose labels or fine-tuned labels. Each line depicts an image taken under the same pose, for the same model trained on image sets with different levels of noise on their relative orientation labels.

Pose Corr.	$E^{(R)} = 0.0°$	$E^{(R)} = 0.2°$	$E^{(R)} = 0.4°$	$E^{(R)} = 0.8°$	$E^{(R)} = 1.6°$
✗					
✓					
✗					
✓					

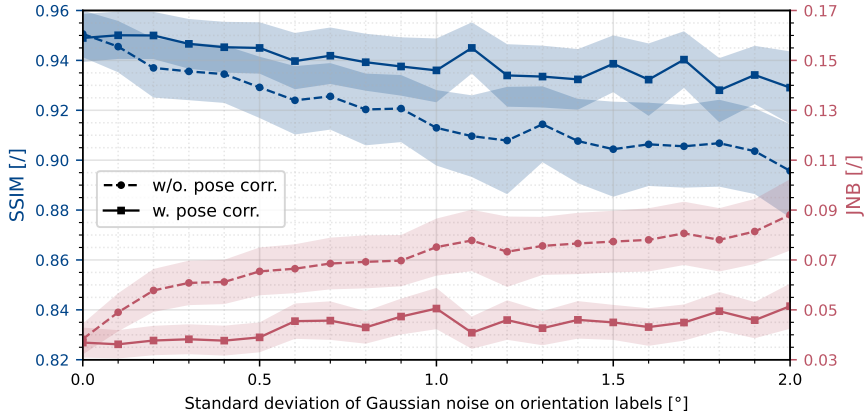


Fig. 5.4 Relationship between the average angular error on the pose labels and the average image quality metrics (SSIM: higher is better, JNB: lower is better, see Section 4.4), with or without performing pose correction. Shaded areas correspond to $[\mu - 0.2\,\sigma, \mu + 0.2\,\sigma]$. Fine-tuning the pose labels improves both the average metrics and reduces their deviation.

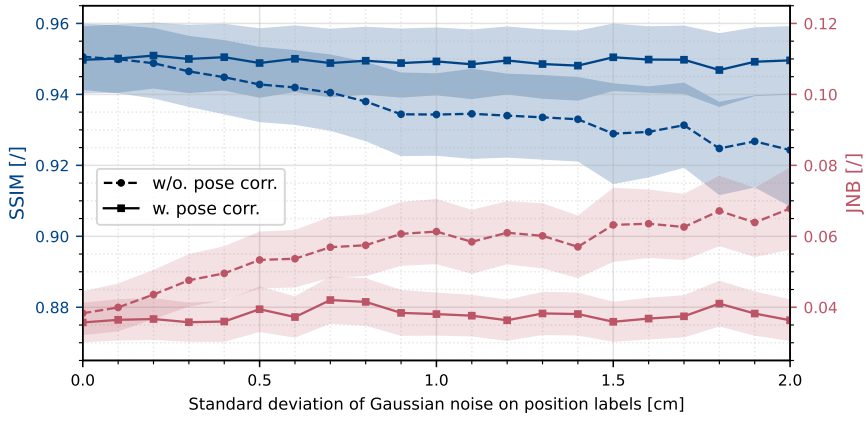
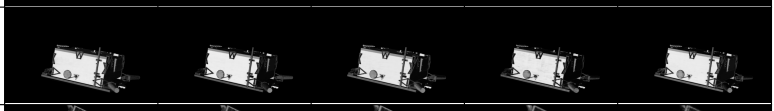
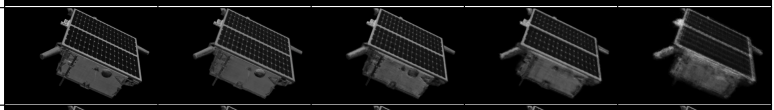
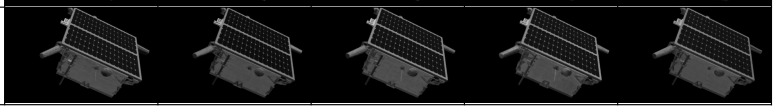


Fig. 5.5 Relationship between the average position error on the pose labels and the average image quality metrics, with or without performing pose correction. Shaded areas correspond to $[\mu - 0.2\,\sigma, \mu + 0.2\,\sigma]$. Fine-tuning the pose labels improves both the average metrics and reduces their deviation.

Table 5.8 Images generated by our model trained either using pseudo-pose labels or fine-tuned labels. Each line depicts an image taken under the same pose, for the same model trained on image sets with different levels of noise on their relative position labels.

Pose Corr.	$E^{(T)}$ = 0 mm	$E^{(T)}$ = 2 mm	$E^{(T)}$ = 4 mm	$E^{(T)}$ = 8 mm	$E^{(T)}$ = 16 mm
✗					
✓					
✗					
✓					

with pose correction are significantly sharper—often matching the reference image. For orientation noise up to 0.8°, reconstructions are visually indistinguishable from the reference; at 1.6°, minor artifacts appear. These results indicate that pose correction effectively compensates for substantial orientation noise, well beyond typical annotation uncertainty. Figure 5.4 illustrates the relationship between noise standard deviation and image quality metrics for models trained with and without pose correction. Across all noise levels, pose correction consistently improves reconstruction quality in terms of both image quality, *i.e.*, SSIM, and sharpness, *i.e.*, JNB score.

Similarly, Table 5.8 presents results for models trained on position-corrupted labels. As with orientation noise, pose correction significantly mitigates the impact of position label uncertainty, producing sharper representations. At low noise levels, reconstructions from pose-corrected models are virtually artifact-free; at higher levels, only minor artifacts remain. Figure 5.5 confirms these observations quantitatively: pose correction improves reconstruction quality across all position noise levels, reinforcing its effectiveness in handling pose uncertainty.

5.5 Conclusions

In this chapter, we proposed a method for reconstructing a 3D model of an unknown spacecraft from monocular 2D imagery. The approach specifically addresses two critical challenges overlooked by state-of-the-art techniques when applied to datasets representative of operational scenarios. These challenges consist of significant illumination variability that undermines the quality and consistency of on-orbit imagery and the pose uncertainty that reduces the reconstruction accuracy.

To overcome these challenges, our method extends Neural Radiance Fields with additional per-image degrees of freedom. For each image, it learns alongside the NeRF an appearance embedding that implicitly captures image-specific illumination conditions and a pose correction term that refines the corresponding noisy pose label. We demonstrated that while introducing additional complexity, these parameters substantially enhance robustness to both illumination variability and pose uncertainty on a dataset representative of orbital conditions.

Beyond nominal conditions, we also evaluated the method's robustness in three challenging scenarios: limited image availability (Section 5.4.3), severe illumination variations (Table 5.4), and increased pose uncertainty (Section 5.4.4). These experiments demonstrate that our approach outperforms prior methods under all these conditions, thereby confirming the effectiveness of its two key components, namely, the appearance embeddings and the learnable pose correction terms.

Contribution to Research Question #1

Research Question #1

How can spacecraft pose estimation be enabled when an explicit prior model of the target's geometry and appearance cannot be assumed?

This chapter demonstrated that an implicit representation M_Φ can be learned from a small set of monocular images acquired on-orbit. For this purpose, the representation should exploit learnable appearance embeddings and pose correction terms to compensate for the varying illumination and pose uncertainty, respectively. Through its novel view synthesis ability,

the proposed representation enables the generation of the images required to train a spacecraft pose estimator F_Θ without requiring an explicit prior on the target's geometry and appearance.

In the next chapter, we demonstrate that such a representation M_Φ enables the learning of a pose estimator F_Θ without requiring the access to a CAD model. Furthermore, we show that M_Φ can be used to improve the Out-Of-Domain (OOD) generalization capabilities of that pose estimator F_Θ.

6

NeRF-based Viewpoint and Appearance Augmentation

LEARNING-BASED spacecraft pose estimation requires large training sets that capture wide variations in viewpoint and appearance, yet only a small number of real images might be typically available before proximity operations. This scarcity enforces the use of synthetic training sets generated using the target Computer-Aided Design (CAD) model. This dependency on explicit target models intrinsically limits the operational scope of pose estimation networks, since they can only be deployed on targets with reliable and accessible geometric models. Due to their limited appearance diversity, these synthetic sets further constrain the reliability of current methods when they encounter the illumination conditions of real on-orbit imagery.

In this chapter, we introduce an augmentation strategy that relies on an implicit scene representation learned from a limited image set to generate both novel viewpoints and diversified appearances. This approach removes the dependence on Computer-Aided Design (CAD)-based rendering pipelines for training and enhances the robustness of spacecraft pose estimators when processing real on-orbit images.

The content of this chapter is adapted from our paper *"CAD-Free Learning of Spacecraft Pose Estimators via NeRF-Based Augmentation"* [2], currently under review, to align with the overall structure and objectives of the thesis.

6.1 Introduction

Recent spacecraft pose estimation works have shown that learning-based methods can achieve accurate monocular pose estimation [5, 43, 47, 49]. These approaches rely on training sets containing tens of thousands of labeled images that depict the target under a wide range of viewpoints and illumination conditions. Since collecting such large real datasets before a mission is not possible, these training sets always consist of synthetic views rendered from a detailed CAD model of the spacecraft.

This training paradigm suffers from two fundamental limitations. First, the entire training process depends on the availability of an accurate CAD model. Many On-Orbit Servicing (OOS) and Active Debris Removal (ADR) targets do not have publicly available geometry, and some are legacy or damaged spacecraft for which no reliable CAD exists. Second, synthetic images do not reproduce the lighting variability, specular reflections, and shadow dynamics encountered in orbit [101, 166]. As a result, networks trained exclusively on synthetic data exhibit poor Out-Of-Domain (OOD) generalization capabilities, *i.e.*, they fail to generalize to real conditions.

We propose a different approach to building training sets for monocular pose estimation. We introduce a NeRF-based image augmentation method that requires only a few tens to a few hundreds of real images of the target. As illustrated in Figure 6.1, a Neural Radiance Field (NeRF) [52] M_Φ is learned from this small set S_{Orig} and used to synthesize a large set S_{NeRF} of geometrically consistent novel views of the target under diverse appearance configurations. The pose estimation network F_Θ can then be learned on the augmented set S_{Augm} that combines the NeRF-based set with the original images. Because the network is trained on images that are diverse in terms of viewpoints and appearances, its OOD accuracy is significantly improved.

This approach eliminates the dependency on a detailed CAD model. It allows learning-based perception to be applied to spacecraft for which the geometry is unknown, incomplete, or proprietary. This is highly relevant for ADR, inspection missions, and servicing scenarios that involve objects with limited priors. Furthermore, when a CAD model is available, the NeRF-generated views can also complement the synthetic dataset, to improve robustness to unobserved lighting variations and to improve generalization to real on-orbit imagery.

This work builds on our preliminary works presented in a conference

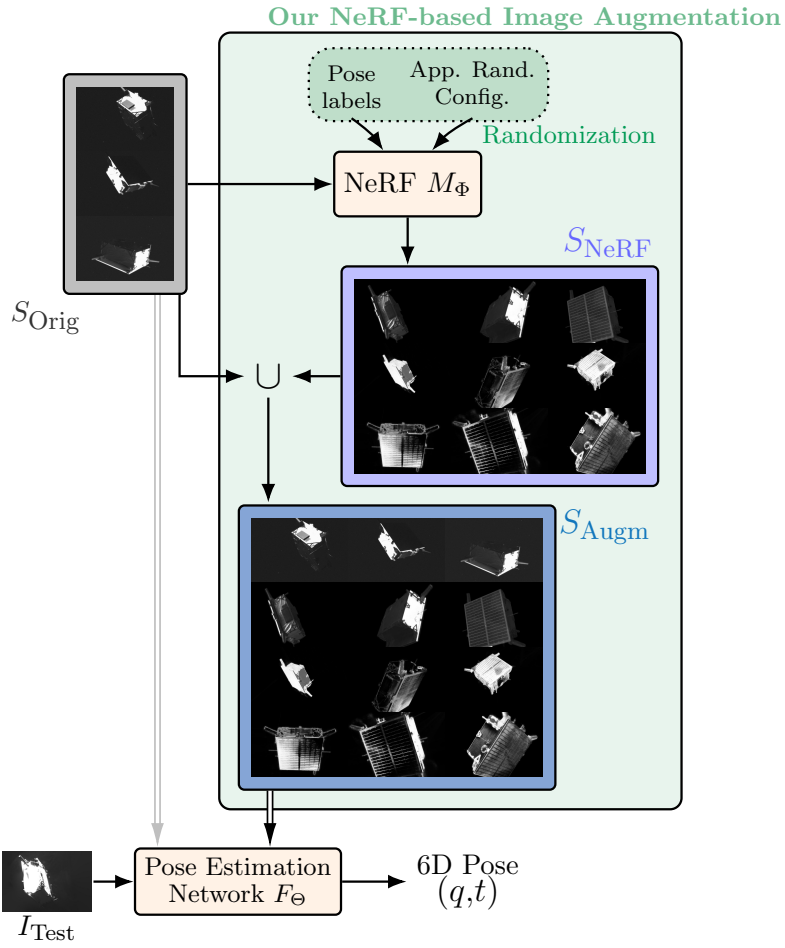


Fig. 6.1 OOD generalization improvement through offline dataset augmentation method. Instead of training a pose estimation network on dataset S_{Orig} limited in terms of both volume and diversity, we train the network on an augmented set S_{Augm} that combines S_{Orig} with a set S_{NeRF}. This set contains images generated through a NeRF trained on S_{Orig} under novel pose labels and randomized appearance conditions. This dataset augmentation enhances the pose estimator's generalization capabilities.

and published as pre-prints [6, 7]. In [6], we demonstrated that a spacecraft pose estimation network can be trained on a small set of real on-orbit images using novel viewpoints generated using a NeRF. In [7], we described a method for synthesizing images of a target spacecraft under randomized appearances by exploiting the density-appearance partial decoupling enabled by NeRFs. These augmented images enable the improvement of the generalization capabilities of a pose estimation network trained on a large set of synthetic images. In this chapter, we validate our NeRF-based augmentation method in two scenarios of interest for proximity operations, respectively assuming a small number of realistic images, and a large number of CAD-based synthetic views. Across both scenarios, NeRF augmentation improves pose-estimation accuracy and broadens the range of missions in which learning-based RPO perception can be deployed.

The rest of this chapter is organized as follows. Section 6.2 positions our works with respect to prior work while Section 6.3 describes our NeRF-based augmentation method. Section 6.4 evaluates the benefits brought by our method on domain generalization capabilities of a spacecraft pose estimation network under different data availability assumptions. Finally, Section 6.5 concludes.

6.2 Related Works

Real on-orbit images of a target spacecraft are difficult to acquire before the proximity phase of an OOS or ADR mission. As a result, only a small set of real images might be available, which results in insufficient coverage of possible viewpoints and appearances. For this reason, most spacecraft pose estimation methods rely on a detailed CAD model to generate a large synthetic training dataset [5, 43, 47, 49]. This reliance on a CAD model limits applicability to missions where accurate geometry is available, which excludes many uncooperative, degraded, or poorly documented targets. Furthermore, synthetic imagery has limited appearance diversity and does not reproduce the illumination and reflectance conditions encountered in orbit [83]. These two limitations, *i.e.*, **data scarcity** and **limited appearance diversity**, are addressed by our NeRF-based solution. They have also been studied in prior work, as detailed below.

6.2.1 Mitigating Data Scarcity

Data scarcity primarily results in poor coverage of the viewpoint space, causing networks to overfit and extrapolate. Specifically, the network adapts too closely to the few sampled viewpoints, leading to degraded performance when predicting on novel views. To address this, several geometric augmentation strategies have been proposed. Two-dimensional augmentations [167] such as cropping or scaling operate strictly in the image plane and provide limited viewpoint diversity. Three-dimensional approaches rely on multi-camera setups [168] or generative models that synthesize new views, such as diffusion-based [169] or adversarial methods [170]. NeRF-based techniques [171–173] exploit a NeRF to learn the scene's configuration, *i.e.*, density and appearance, from the few available images. The NeRF can then be used to generate novel training images under unseen viewpoints, thereby complementing the space coverage. These *methods improve viewpoint coverage* but they focus exclusively on generating new views and *do not address the limited appearance diversity* of the available images.

6.2.2 Increasing Appearance Diversity

The lack of appearance diversity in synthetic datasets causes pose estimation networks to rely on spurious visual cues and generalize poorly to real on-orbit images. Domain randomization [91] addresses this limitation by introducing variability during rendering. CAD-based pipelines [174–176] randomize texture, illumination, background, noise, or blur. Other approaches introduce diversity directly in the image space, for example using style transfer [88, 177], random convolutions [178], randomized layers [179], or Fourier perturbations [89]. Image-space methods introduce appearance diversity without requiring a CAD model, but they remain restricted to two-dimensional operations and cannot randomize the underlying 3D scene appearance.

Domain adaptation, which exploits information on the target domain such as unlabeled images to complement the source domain, has also been explored for spacecraft pose estimation [83, 180, 181]. However, these techniques require access to target-domain data and incur computational costs that exceed the capabilities of current space-grade hardware [182, 183]. For operational missions, domain generalization through image augmentation therefore remains the only practically viable option.

6.2.3 Related Works Limitations

Existing approaches address viewpoint coverage or appearance diversity in isolation but do not overcome the fundamental limitations of current spacecraft pose estimation methods. Many rely on CAD-based rendering to compensate for data scarcity, yet these rendered datasets provide limited appearance variability. Methods that introduce appearance diversity without a CAD model operate strictly within the two-dimensional image grid, which constrains the range of variations they can generate. As a result, none of these techniques can train a target-specific pose estimator when only a small set of real images is available and no reliable CAD model exists.

6.3 Method

This section presents our method for generating an augmented training set that provides broad viewpoint coverage and significant appearance variability from an initially limited collection of images. By enriching the available data through this image generation pipeline, we enhance the OOD generalization capabilities of a pose estimation network trained on the resulting dataset.

Our method relies on three successive stages. First, the input images are pre-processed to make them suitable for learning a neural scene representation. Second, two neural radiance fields are trained on the pre-processed data, as depicted in Figure 6.2. Finally, the first NeRF, M_Φ, which captures the scene appearance, is responsible for generating foreground images from arbitrary viewpoints under diverse appearance conditions while the second NeRF, M_Ψ, which models the scene geometry, predicts the corresponding foreground masks. When training the pose estimation network, the masks are used to insert a random background behind the foreground image in order to make the pose estimator robust to background distractors.

The remainder of this section follows a three-stage structure. Section 6.3.1 describes the required pre-processing steps, and Section 6.3.2 details the learning strategy for training M_Φ and M_Ψ. Finally, Section 6.3.3 presents the image-synthesis procedure that preserves scene geometry while randomizing its appearance.

6.3.1 Data Pre-processing

We first transform the available input images $\{I_a^*\}_{a=1}^{N_{\text{orig}}}$ of width W and height H into a dataset that can be used for learning a NeRF. We start by estimating the relative pose $(\hat{q}, \hat{t})_a$, and foreground mask $\hat{m}_a$[1] of each image I_a^*. This can for example be done by using COLMAP [165] or human annotations. The mask is then applied to remove the background from the original image in order to prevent the reconstruction of background content that is not relevant to the spacecraft.

Using the camera calibration matrix, we cast rays through each pixel ij of the image. Each ray is defined by the camera center $\hat{c}_a$ and the direction $\hat{d}_{aij}$ that corresponds to pixel ij. This produces a NeRF training set, in which each element consists of the ray origin $\hat{c}_a$, the ray direction $\hat{d}_{aij}$, the pixel value I_{aij}, and the binary mask value m_{aij}.

6.3.2 Radiance Field Learning

We aim to learn a representation that can generate a novel view of the scene and the corresponding foreground mask for any given relative pose. In addition, the representation must allow the randomization of the scene appearance while preserving the underlying geometry. As discussed in Chapter 3, NeRFs satisfy these requirements because they decouple scene geometry from appearance.

Learning a NeRF from on-orbit imagery is challenging because illumination conditions vary strongly across views and because the relative pose labels contain uncertainty. To address these difficulties, we follow the methodology developed in the previous chapter and extend a NeRF with additional degrees of freedom. As introduced in Chapter 5, per-image appearance embeddings allow the model to learn image-specific illumination conditions instead of averaging illumination across all views. Learnable pose correction terms enable the model to progressively refine the input pose labels and reduce the overall 3D uncertainty.

As depicted in Figure 6.2, for each ray $(\hat{c}, \hat{d})$ in the NeRF training set, we refine the ray origin $\tilde{c}$ and the ray direction $\tilde{d}$ through a learnable pose correction term and we select the appearance embedding $\mathbf{e}$ associated with the image from which the ray is cast. We then sample N_{P} points along the

[1]The foreground mask is a binary image that differentiates the foreground, *i.e.*, the spacecraft, from the background.

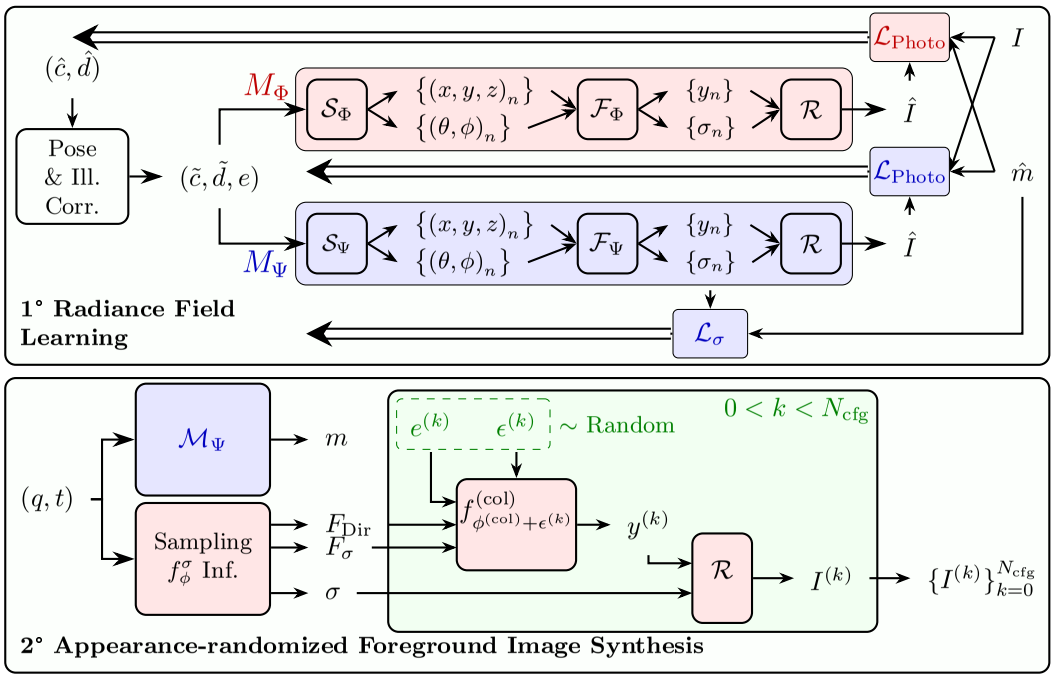


Fig. 6.2 NeRF-based viewpoint and appearance augmentation: method overview. As explained in Section 6.3.2, We first learn two NeRFs M_Φ and M_Ψ. M_Φ is dedicated to the representation of the target appearance through its supervision on a photometric loss. M_Ψ complements this supervision with a loss on the density of the background points, so as to refine the density field. Then, as explained in Section 6.3.3, both M_Φ and M_Ψ are used to synthesize training samples. For each pose label, we generate a segmentation mask m using M_Ψ and N_{cfg} foreground images $I^{(k)}$ through M_Φ. To render diverse appearances, we randomize either the illumination through the appearance embedding $\mathbf{e}^{(k)}$, or the color through the color MLP's weights $\epsilon^{(k)}$.

ray. Each sampled point is characterized by its position (x, y, z) and its two viewing angles (θ, ϕ). For each point, the neural field predicts its density σ and its color y. Differentiable ray-tracing techniques aggregate these quantities to predict the pixel value $I^{(\text{photo})}$. The photometric loss between the predicted and observed values is computed and back-propagated through the NeRF M_Φ to update its weights. After multiple iterations, M_Φ learns a representation of the scene geometry and appearance that enables the synthesis of novel views.

Although M_Φ produces visually convincing images, preliminary experiments showed that the masks extracted from the predicted densities are noisy. Since these masks are later used to randomize the background or as supervision for pose estimator training, this level of noise is not acceptable. For this reason, we introduce a second NeRF, denoted M_Ψ, which focuses exclusively on the geometry. Its architecture is identical to that of M_Φ, but the training objective differs. In addition to the photometric loss, we supervise the density using an auxiliary *per-ray* loss $\mathcal{L}_\sigma$, defined based on the mask value $m(\hat{\mathbf{c}}, \hat{\mathbf{d}})$ available for the ray of interest $(\hat{\mathbf{c}}, \hat{\mathbf{d}})$,

$$\mathcal{L}_\sigma = \sum_{n=1}^{N_\text{P}} \sigma_n^2 \, (1 - m(\hat{\mathbf{c}}, \hat{\mathbf{d}})). \tag{6.1}$$

It enforces zero density for points that contribute only to background pixels. During training, $\mathcal{L}_\sigma$ is averaged over the mini-batch and combined with the photometric loss using unit weights, i.e., $\mathcal{L}_\Psi = \mathcal{L}_\text{photo} + \mathcal{L}_\sigma$. This constraint prevents the network from hallucinating spurious density outside the target envelope in order to improve the image reconstruction loss. As a consequence, the images synthesized by M_Ψ have lower visual quality, but the masks extracted from them are significantly sharper. The pose-correction terms are updated through M_Φ; M_Ψ does not contribute to these pose correction terms.

6.3.3 Appearance-randomized Image Generation

To augment the training set with both viewpoint diversity and appearance variability, we begin by randomly sampling N_lab pose labels in $\text{SE}(3)$. For each sampled pose (q, t), we render the foreground mask m with the density-supervised NeRF M_Ψ by thresholding the accumulated opacity at τ, and we synthesize N_cfg foreground images for that viewpoint with the appearance-supervised NeRF M_Φ, reusing its density and density features while randomizing appearance.

For this purpose, we use the sampler and the density field of M_Φ to compute the density σ together with the direction and density features F_{dir} and F_σ. Since the density field defines the projection of the scene geometry into the image, all remaining steps of the image generation process can be randomized without altering the target shape. We rely on two mechanisms to randomize appearance: illumination randomization through the appearance embedding and color randomization through the color network.

Illumination Randomization During training, the network learns a set of appearance embeddings $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_N\} \subset \mathbb{R}^d$. At inference time, we randomize the appearance by sampling a new embedding $\mathbf{e}^{(k)}$ using one of four strategies. The first strategy selects an embedding uniformly at random, meaning that we draw $i \sim \text{Unif}\{1, \ldots, N\}$ and set $\mathbf{e}^{(k)} = \mathbf{e}_i$. The second strategy interpolates between two embeddings by sampling $i \neq j$ and a coefficient $\alpha \in [0, 1]$ and forming

$$\mathbf{e}^{(k)} = \mathbf{e}_i + \alpha(\mathbf{e}_j - \mathbf{e}_i). \tag{6.2}$$

The third strategy extrapolates between two embeddings using the same expression but with a coefficient α outside the interval $[0, 1]$. The final strategy samples $\mathbf{e}^{(k)}$ from a Gaussian model fitted to $\mathcal{E}$, treating the learned embeddings as realizations of a multivariate normal distribution with empirical mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}_{\text{emb}}$, and drawing $\mathbf{e}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\text{emb}})$.

Color Randomization A second approach for appearance randomization consists in injecting Gaussian noise into the parameters of the color MLP. Instead of using the deterministic color network $f^{(\text{col})}_{\phi^{(\text{col})}}$, we perturb its parameters with a noise vector $\varepsilon^{(k)} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\Phi)$ and evaluate the colors using the perturbed network $f^{(\text{col})}_{\phi^{(\text{col})}+\varepsilon^{(k)}}$. The color of each sampled point is then computed using this noise-injected parameter vector.

The point-wise colors $\{y^{(k)}_n\}$ predicted by the randomized color network are combined with the preserved densities $\{\sigma_n\}$ to compute the final pixel value $I^{(k)}$. By applying different randomization configurations $(\mathbf{e}^{(k)}, \varepsilon^{(k)})$, we generate N_{cfg} images of the target spacecraft that all correspond to the same pose but exhibit different appearances.

6.4 Experimental Validation

This section describes the experiments conducted to validate our proposed augmentation method. First, Section 6.4.1 presents our experimental setup. Then, Section 6.4.2 demonstrates how our method enables the training of a pose estimator on a dataset with severe limitations regarding the number of available viewpoints as well as the image diversity. Section 6.4.3 and Section 6.4.4 then evaluate our method's capabilities when applied on datasets that are even more limited in terms of number of viewpoints and quality of the images, respectively. To isolate the performance improvement originating from the increase in number of viewpoints, we then show in Section 6.4.5 how our method improves the OOD generalization capabilities of a pose estimation network when an arbitrary large number of viewpoints are available, *e.g.*, because a CAD model is available. Finally, Section 6.4.6 performs an ablation study on the use of the second NeRF dedicated to the synthesis of segmentation masks.

6.4.1 Experimental Setup

Our method enables the training of pose estimation network F_Θ from a set of images that is too small and lacks appearance diversity. To validate our method, we conduct our experiments on 4 image sets from SPEED+ [45] (*Synthetic*, *Lightbox*, *Sunlamp*) and SHIRT [58] (*Lightbox_ROE2*) which were described in Chapter 4. The evaluation metrics consist of the angular and translation errors and the pose score, $E^{(R)}$, $E^{(T)}$ and $S^{(P)}$, respectively, which were defined in Section 4.3.

Unless otherwise mentioned, since *Lightbox_ROE2* is representative of an on-orbit acquisition scenario through its trajectory-based pose distribution and illumination conditions, we exploit N_{Orig}=400 images randomly sampled in *Lightbox_ROE2* as the original train set S_{Orig}. As the Hardware-In-the-Loop (HIL) sets of SPEED+ (*Lightbox* and *Sunlamp*) depict realistic illumination while being independent from *Lightbox_ROE2*, they are used to evaluate the domain generalization abilities of a network trained using our augmentation method.

We also conduct similar experiments on models trained using the *Sunlamp* and *Synthetic* sets. The former enables the evaluation of the method's ability to deal with more challenging illumination conditions. The latter

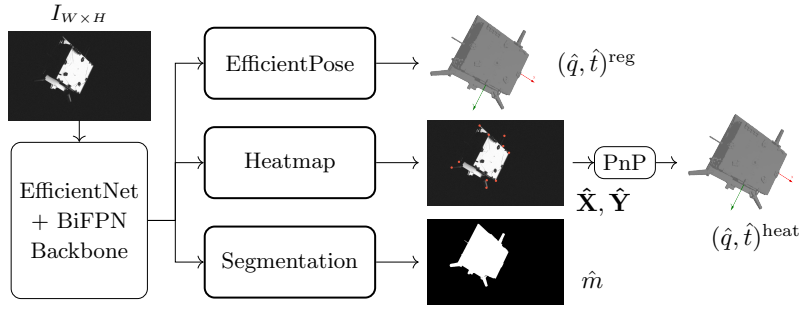


Fig. 6.3 Architectural overview of SPNv2 [49]. The network extracts multi-scale feature maps using an EfficientNet backbone and a BiFPN. They are processed through three heads: (i) a pose-regression head, (ii) a keypoint head whose predictions are used by a PnP solver to compute the pose, and (iii) a segmentation head providing auxiliary supervision. The model outputs a hybrid pose combining the regressed translation with the keypoint-based orientation. Further details are available in Figure 2.15.

evaluates the method's ability in improving the test-time accuracy of a pose estimator trained on a large set of synthetic images generated using the target CAD model.

We train two NeRFs, M_Φ and M_Ψ, focusing respectively on the learning of the object appearance and on the scene density. Both NeRFs follow the *K*-Planes implementation [53]. Once trained, they are used to generate the augmented set S_{NeRF} by sampling 48,000 pose labels in $\text{SE}(3)$. In practice, these labels are taken from the train set of the *Synthetic* set to provide a reference for comparison. For each pose label, a segmentation mask and $N_{\text{cfg}} = 64$ images corresponding to that pose under different appearance-randomization configurations (24 illumination augmentations ($\mathbf{e}^{(k)}$) and 40 color augmentations ($\epsilon^{(k)}$)) are generated. Training both NeRFs takes 50 minutes on an NVIDIA L40s while the rendering of each mask takes 1.5 second and the rendering of the $N_{\text{cfg}} = 64$ images takes 3.8 seconds. Both images and masks are generated at a 768×512 resolution which corresponds to the input resolution of the pose estimation network F_Θ.

The pose estimation network F_Θ used in our experiments is SPNv2 [49], a state-of-the-art spacecraft pose estimation network designed to achieve good OOD generalization capabilities, and is depicted in Figure 6.3. SPNv2 is built around an EfficientNet [103] backbone and a BiFPN [104] for multi-scale feature fusion, followed by three heads: an EfficientPose [105] regres-

Table 6.1 OOD performance comparison of SPNv2 trained on the *Lightbox_ROE2* set under four augmentation strategies: no appearance, illumination-only, color-only, and combined illumination-color, augmentations. Applying both illumination and color augmentations yields the strongest OOD generalization performance.

Ill. Augm.	Color Augm.	*Lightbox* $S^{(\mathrm{P})}[/]$	*Lightbox* $E^{(\mathrm{R})}[^\circ]$	*Lightbox* $E^{(\mathrm{T})}[cm]$	*Sunlamp* $S^{(\mathrm{P})}[/]$	*Sunlamp* $E^{(\mathrm{R})}[^\circ]$	*Sunlamp* $E^{(\mathrm{T})}[cm]$
✗	✗	0.168	7.55	23.43	0.529	25.62	57.12
✓	✗	0.141	6.53	17.73	0.396	19.53	36.58
✗	✓	0.144	6.31	21.37	0.276	12.90	35.61
✓	✓	**0.109**	**4.80**	**15.37**	**0.191**	**9.01**	**23.33**

sion head, a keypoint-prediction head refined through a PnP solver [78], and a segmentation head used for auxiliary supervision. At inference, the model outputs a hybrid pose in which the translation comes from the regression head while the orientation is taken from the PnP-based branch. Designed for robustness under domain shift through multi-task learning and augmentations, SPNv2 is used as a baseline to assess the impact of our augmentation method on OOD generalization.

In our experiments, SPNv2 is trained for 20 epochs where each epoch consists of the N_{Orig} images from the S_{Orig} set combined with the 48,000 new pose labels used to generate S_{NeRF}. For each pose label, a sample is drawn at random from of the N_{cfg} images generated for that pose label so that the network is exposed to different appearance-randomization configurations. We add the Earth in the background of half of the 48,000 images using the foreground segmentation masks predicted by M_{Ψ}. These masks are also used a supervision signal for training SPNv2, as in the original paper [49].

Once trained, the pose estimation networks are evaluated on the *Lightbox* and *Sunlamp* datasets using as metrics the average translation error $E^{(\mathrm{T})}$, angular error $E^{(\mathrm{R})}$ and SPEED score $S^{(P)}$ described in Section 4.3.

6.4.2 CAD-free learning of a pose estimator

This section evaluates our method's ability to generate a sufficiently large and diverse dataset so as to enable the training of a robust spacecraft pose estimation network from a reduced set of images taken in orbit, thereby alleviating the CAD model requirement encountered by such models. Such a set suffers from limitations regarding both its viewpoint coverage and

appearance diversity. For this purpose, we conduct our experiments on a subset of *Lightbox_ROE2* containing 400 images taken under realistic illumination conditions and viewpoints that correspond to a slow approach towards a tumbling target spacecraft.

We first trained the pose estimator on these 400 images using the standard training procedure of SPNv2. Unsurprisingly, the so-trained network performs extremely poorly, with pose scores in the order of 2.0, which roughly corresponds to 80°, on all test sets. This indicates that the network severely overfitted the 400 images. Table 6.1 indicates the test-time performance of SPNv2 trained using different configurations of our NeRF-based augmentation method designed to prevent that overfitting.

First, by generating 48,000 images using only the NVS ability of a NeRF trained on these 400 images, we can learn a pose estimation network properly. It achieves a reasonable pose score of 0.168 on *Lightbox*, but is significantly worse on *Sunlamp* as the pose score rises to 0.529. This is expected because the illumination conditions within *Lightbox_ROE2* are close to the ones encountered in *Lightbox* but different from the challenging illumination conditions encountered in *Sunlamp*.

Then, Table 6.1 also presents the performance metrics achieved by a pose estimation network trained using our appearance-diversification strategies. Illumination augmentations through randomized appearance embeddings lead to an improvement of the generalization capabilities on both sets. On *Lightbox* and *Sunlamp*, the pose scores are of 0.141 and 0.396, respectively. These improvements are even better using appearance augmentations enabled by the randomization of the color MLP. Using only these color augmentations, we achieve pose scores of 0.144 and 0.276, respectively.

Finally, Table 6.1 highlights that combining both forms of appearance augmentations leads to the best generalization capabilities. On *Lightbox*, we achieve a pose score of 0.109 while on *Sunlamp*, the pose score reaches 0.191. This indicates that the proposed geometry-preserving appearance-randomization strategy reduces by 35% and 57% the pose scores on *Lightbox* and *Sunlamp*, respectively, compared to performing only NeRF-based viewpoint augmentation.

6.4.3 Impact of Smaller Train Sets

In the previous section, we evaluated the ability of our augmentation method to enable the learning of a pose estimation network from images acquired during a rendezvous approach. It assumed access to a reasonably

large set of 400 images taken on-orbit and depicting the target spacecraft. In practice, acquiring these 400 images of the target might be complicated. Hence, in this section, we assess the evolution of this capability when the number of available images is reduced.

Table 6.2 compares the performance achieved by a pose estimation network trained using the same viewpoint and appearance strategy on 5 datasets of decreasing sizes. As the number of available images decreases, the pose score is progressively degraded. This is expected as the level of information in the train set is reduced and the quality of the learned representation is therefore affected. However, this decrease remains gradual. Even at 100 images, the network generalization abilities remain attractive. When decreasing the number of images from 400 to 100 images, the pose scores evolve from 0.109 on *Lightbox* and 0.191 on *Sunlamp* sets to 0.205 and 0.241, respectively. Reducing by a factor 4 the number of images that need to be acquired therefore only costs an increase of the pose score of 88% and 26%, respectively on those sets.

If we further decrease the number of images, the performance drops but even when only 25 images are available, the image-to-pose mapping can still be learned, even if its accuracy on both domains is limited. Although the number of available images is divided by a factor 16, the pose scores evolve from 0.109 and 0.191 to 0.403 and 0.282, respectively on both sets. While this represents an increase of 270% on *Lightbox*, this increase is limited to 48% on *Sunlamp*. These experiments therefore demonstrate that our method is able to cope with the really low data regimes that could be required in the context of proximity operations with a poorly defined target spacecraft.

Table 6.2 OOD performance comparison of SPNv2 trained on five datasets of decreasing sizes augmented through the same viewpoint and appearance augmentations. The performance is smoothly reduced when the number of available images decreases.

N_{Orig}	*Lightbox*			*Sunlamp*		
	$S^{(\text{P})}[/]$	$E^{(\text{R})}[^\circ]$	$E^{(\text{T})}[cm]$	$S^{(\text{P})}[/]$	$E^{(\text{R})}[^\circ]$	$E^{(\text{T})}[cm]$
400	**0.109**	**4.80**	**15.37**	**0.191**	**9.01**	23.33
200	0.173	7.75	23.70	0.297	13.97	38.03
100	0.205	9.66	22.39	0.241	11.68	24.98
50	0.241	11.07	30.13	0.394	19.00	44.76
25	0.403	19.25	42.52	0.282	14.24	**21.22**

6.4.4 Impact of Adverse Illumination Conditions

In the previous section, we examined our method's behavior when the training set is even smaller. Beyond quantity, image quality is also of tremendous importance. Indeed, we previously conducted our experiments on image sets acquired using a testbed emulating on-orbit illumination under diffuse conditions. In practice, on-orbit images may be significantly affected by direct solar exposure and exhibit strong contrasts. In this section, we therefore evaluate our method's resilience to harsh illumination conditions within the available training set.

For this purpose, we carried out an experiment using 400 images randomly sampled in *Sunlamp* to train the pose estimation network. The so-trained pose estimation network achieves good performances given the poor visual quality of the available images. On *Lightbox*, the pose score is equal to 0.291 while it is equal to 0.096 when computed on the *Sunlamp* images that are not used in the training set. These images therefore share similar illumination conditions as the training set but are not seen by the pose estimation network at training.

This experiment demonstrates that the proposed augmentation method is able to generate an augmented set that is sufficiently rich in terms of viewpoints and appearances so as to enable the training of a robust pose estimation network even on datasets made of images of poor visual quality.

6.4.5 Augmentation of Synthetic Train Sets

Our augmentation method addresses the two limitations that prevent the training of a pose estimation network from a small set of real images depicting the target spacecraft and impose the reliance on CAD model to generate a relevant training set. It enables the generation of novel viewpoints as well the randomization of the target appearance through an implicit 3D model learned from these on-orbit images. In the previous sections, we demonstrated the coupled impact of both augmentation types. In this section, we aim at decoupling the impact of both augmentations. We therefore conduct our experiments on a training set made of 48,000 images from the *Synthetic* set. With this large set, we do not need to generate any novel viewpoint, and therefore only render the current ones under different appearances so as to only evaluate the impact of appearance augmentation.

Table 6.3 OOD performance comparison of SPNv2 trained on 48,000 *Synthetic* images with and without using our NeRF-based appearance-randomization strategy. Applying our augmentation strategy improves the OOD generalization capabilities.

Augmentation Strategy	*Lightbox*			*Sunlamp*		
	$S^{(P)}[/]$	$E^{(R)}[^\circ]$	$E^{(T)}[cm]$	$S^{(P)}[/]$	$E^{(R)}[^\circ]$	$E^{(T)}[cm]$
No	0.193	8.90	**22.98**	0.295	14.56	25.73
Color & Ill.	**0.155**	**6.75**	24.36	**0.183**	**8.96**	**17.28**

Table 6.3 compares the OOD generalization capabilities of a pose estimation network trained either on the *Synthetic* set or using that set augmented with our appearance-randomization method. On *Lightbox*, our method significantly reduces the pose score by 19.7%. On *Sunlamp*, this improvement is even more pronounced with a pose score improved by 38%. This highlights that the potential of our method extends beyond enabling the training of pose estimators without relying on a CAD model. With this experiment, we showed that it can also improve the OOD generalization capabilities of a pose estimator trained on a CAD-based dataset, even if that pose estimator already exhibit strong OOD generalization capabilities through the use of a wide collection of image augmentation techniques.

6.4.6 Ablation on Density Supervision

This section presents an ablation study on the use of the second NeRF M_Ψ dedicated to the representation of the target density and the synthesis of the foreground masks m used to (i) supervise the auxiliary segmentation task and (ii) enable the randomization of the image background so as to improve the network resilience to the background content, *e.g.*, the Earth.

Table 6.4 OOD performance comparison of SPNv2 trained using masks generated by either M_Ψ, the NeRF dedicate to geometry modeling, M_Φ, the NeRF dedicated to appearance modeling, or using no mask at all. Using masks generated by M_Ψ leads to the best performance.

Masks Origin	*Lightbox*			*Sunlamp*		
	$S^{(P)}[/]$	$E^{(R)}[^\circ]$	$E^{(T)}[cm]$	$S^{(P)}[/]$	$E^{(R)}[^\circ]$	$E^{(T)}[cm]$
M_Ψ	**0.109**	**4.80**	**15.37**	**0.191**	**9.01**	**23.33**
M_Φ	0.189	8.20	29.12	0.253	11.88	32.27
No masks	0.424	19.25	59.04	0.387	18.15	52.23

Table 6.4 compares performance metrics achieved by the pose estimation network when trained with our augmentation strategy using different mask synthesis strategies. It highlights the metrics achieved by the estimation network when trained using (i) the masks generated by M_Ψ, (ii) the masks generated by M_Φ, or, (iii) no mask at all. Compared to using clean segmentation masks produced by M_Ψ, using those predicted by M_Φ degrades the pose scores as they evolve from 0.109 and 0.191 to 0.189 and 0.253, respectively on both sets. Using no masks at all, *i.e.*, neither for auxiliary segmentation nor for background augmentation, the pose scores on both sets are significantly degraded as they reach 0.424 and 0.387, respectively. This emphasizes the importance of segmentation masks when learning a pose estimation network and that their quality is so critical that we need a second NeRF M_Ψ dedicated to their generation.

6.5 Conclusions

This chapter introduced an original approach to spacecraft pose estimation that removes the need for a CAD model and enables target-specific networks to be learned from only a small set of real or real-like images. The core idea is to exploit a Neural Radiance Field (NeRF) learned from these images as a three-dimensional scene representation in which geometry and appearance are naturally decoupled. This representation allows our method to generate novel viewpoints and to randomize appearance directly in the scene without altering the underlying geometry.

By combining viewpoint augmentation with appearance randomization, the proposed approach produces an augmented training set that can be used to train spacecraft pose estimation networks for targets whose geometry is unknown, undocumented, or inaccessible. The same mechanism also improves OOD generalization when used to augment large CAD-based datasets, since it introduces appearance variability that cannot be produced by conventional rendering or two-dimensional image transformations.

We demonstrated that robust pose estimation networks can be trained from a few tens to a few hundreds of images, including images captured under challenging illumination conditions. Finally, a further analysis highlighted the importance of accurate segmentation masks produced by a second NeRF model dedicated to modeling the scene's geometry.

Contribution to Research Question #1

Research Question #1

How can spacecraft pose estimation be enabled when an explicit prior model of the target's geometry and appearance cannot be assumed?

In Chapter 5, we showed that an NVS representation M_Φ can be learned from on-orbit monocular imagery, provided that the representation exploits appearance embeddings and pose correction terms to compensate for the varying illumination conditions and pose uncertainty. In this chapter, we build on this representation M_Φ to generate a large training set encompassing diverse viewpoints and appearance conditions. We then demonstrated how this large and diverse dataset enables the training of a robust pose estimation network F_Θ without relying on an explicit model of the target.

Contribution to Research Question #2

Research Question #2

How can dataset-level information be aggregated to enable augmentations whose diversity surpasses the limits of image-wise transformations and improves Out-Of-Domain generalization?

This chapter demonstrated that a dataset-level augmentation method can be built using a NeRF M_Φ learned from a small image set and exploited to synthesize both novel viewpoints and diversified appearances. The learned representation aggregates all available information into a unified three-dimensional model that decouples geometry and appearance, supports the sampling of pose labels in $\mathrm{SE}(3)$ for Novel View Synthesis (NVS), and enables appearance diversification through randomized appearance embeddings and noise injection in the color MLP. As a result, the augmented set S_{NeRF} captures variations that two-dimensional image-wise transforms cannot generate, since it modifies the underlying three-dimensional target appearance and illumination while preserving its geometry.

With dataset-level augmentation in place, robustness is improved but human confidence is still impaired by a lack of understanding on what the network sees. The next chapter introduces a NeRF-based approach to visualize its decision cues and answers to our third research question.

7

Visualizing 3D Decision Cues in Pose Estimation Networks

AUTONOMOUS Rendezvous & Proximity Operations (RPOs), including those involved in On-Orbit Servicing (OOS), rely on the ability of a chaser spacecraft to accurately estimate the relative pose of a target. When the target is uncooperative and only passive sensors are available, this task becomes particularly challenging, especially under complex illumination and viewing conditions.

These difficulties have motivated the use of data-driven pose estimation methods that exploit neural networks trained on synthetic imagery. While state-of-the-art methods have proven robust against domain shifts, their inherent lack of transparency complicates assessing their behavior in mission conditions, making thorough verification and validation critical yet especially challenging.

The content of this chapter is based on our paper *"NeRF-based Visualization of 3D Cues Supporting Data-Driven Spacecraft Pose Estimation"*, published at the IEEE International Conference on Space Robotics (ISpaRo) 2025, and adapted to align with the overall structure and objectives of the thesis.

7.1 Introduction

Data-driven methods have recently become the dominant approach for estimating the relative pose of a target spacecraft from monocular images. These techniques typically rely on neural networks trained on large synthetic datasets generated from a CAD model of the target, enabling them to predict position and orientation directly from a single input image. Their robustness to challenging illumination conditions has made them attractive alternatives to traditional hand-crafted pipelines [46, 65].

Data-driven methods bring several challenges. As discussed in the previous chapter, training on synthetic imagery does not guarantee reliable performance on real images, and the resulting networks must remain lightweight enough to run on space-grade hardware [96, 182]. While these issues have received significant attention, a third challenge has been comparatively overlooked: the inherent lack of explainability of neural networks. This raises a critical question for deployment in safety-critical scenarios: *which features does a pose estimation network actually rely on, and can we trust these features to remain informative once it is deployed in orbit?*

This chapter addresses this question by introducing a method for gaining insights into the 3D cues exploited by spacecraft pose estimation networks. As illustrated in Figure 7.1, our approach trains an image generator M_Φ conditioned on a pose label, *i.e.*, orientation q and position t, so that the resulting image $\hat{I}$ leads a pretrained neural network F_Θ to recover the same pose $(\hat{q}, \hat{t})$. The underlying assumption is that if such a generator consistently produces images that are correctly interpreted by the pose estimator, then it must have learned to reproduce the target features most critical to that estimator's predictions.

The contributions of this chapter are two-fold. First, we introduce a method for visualizing the key target features exploited by a spacecraft pose estimation network. Second, we evaluate this method on several state-of-the-art networks to gain insights into the target 3D cues that guide their decisions.

7.2 Related Works

Despite a considerable interest in the computer vision community, explaining the predictions of a neural network remains an open question.

Explainability methods aim at determining which input pixels have the strongest influence on the prediction of a trained neural network. Although several methods exist, they can be classified in two main categories. On one hand, class activation maps [184–186] reflect by design the derivatives of the network predictions with respect to the input pixels. On the other hand, perturbation-based methods [187, 188] consist in modifying the input image and observing how this perturbation has affected the network prediction. These methods have been developed for both convolutional neural networks [189], as well as transformer-based architectures [190] with a strong emphasis on classification tasks. Unlike those works, our method does not aim at identifying which input pixels contributed to the network decision for a particular input image. Instead, it aims at synthesizing a 3D scene whose rendering in 2D is consistent with the pose estimation model, meaning that the pose predicted from a projected image corresponds to the actual viewpoint parameters used to project the scene.

In this work, the image generator consists in a Neural Radiance Field (NeRF) [52]. A NeRF aims at learning the geometry and appearance of a scene from a set of a few training images along their pose label in order to render novel views of that scene from unseen viewpoints. As explained in Chapter 3 (see Figure 3.3), to synthesize an image taken from a given viewpoint, a NeRF projects rays across every pixel of that image. It then samples points along the rays and feeds them in a MultiLayer Perceptron (MLP) which outputs their colors and densities. Finally, through differentiable

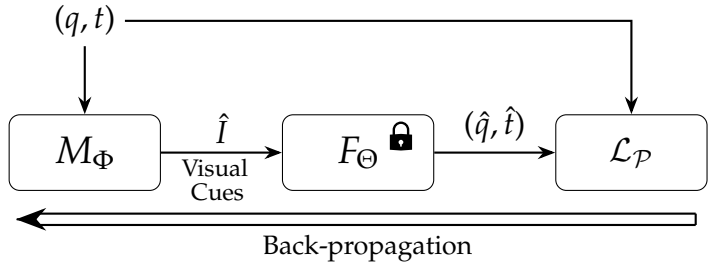


Fig. 7.1 To visualize the 3D cues exploited by a spacecraft pose estimation network F_Θ, our method relies on an image generator M_Φ which takes as input a 6D pose and outputs an image. By back-propagating the difference between the pose predicted on that image (by the frozen pose estimator) and the input pose, the generator is trained. Once it has been trained, the generator can synthesize images containing the 3D cues on which the pose estimation network primarily relies.

ray-tracing techniques, the color of each pixel is computed from the color and density of all the points along the corresponding rays. The NeRF is iteratively trained on the training images by back-propagating the photometric loss between the training image and the image generated by the NeRF under the corresponding pose label. Although several works [6, 7] explored the use of NeRFs in the context of spacecraft pose estimation, they only considered the Neural Radiance Field as a tool for data augmentation. Here, their ability of learning a 3D scene through back-propagation is exploited to visualize the features that support the predictions of a spacecraft pose estimation network.

7.3 Method

This section describes our method for visualizing the features that are primarily exploited by a spacecraft pose estimation network F_Θ. As depicted in Figure 7.1, we rely on a generation network M_Φ which takes as input a 6D pose, *i.e.*, orientation q and position t, and outputs an image $\hat{I}$ of resolution WxH. Our goal is to train this generator so that the rendered images contain the features primarily exploited by the pose estimation network F_Θ. The generation network M_Φ is defined by a set of weights Φ. To determine the optimal generator weights Φ^* that achieve that goal, we train it by iteratively updating its weights through the back-propagation of the difference between the input pose and the pose predicted by the frozen pose estimation network $(\hat{q}, \hat{t}) = F_\Theta(\hat{I})$, *i.e.*,

$$\Phi^* = \arg\min_{\Phi} \mathcal{L}_\mathcal{P}\left(\Phi; \left(\hat{q}, \hat{t}\right), (q, t)\right) \tag{7.1}$$

where $\mathcal{L}_\mathcal{P}$ computes the distance between the input and predicted poses in the $\mathrm{SE}(3)$ space.

Although the proposed feature visualization approach is described in general terms, training an image generation network under this formulation may lead to convergence issues. To increase the probability that the training of the generator converges, we rely on an NeRF-based image generator, which is described in Section 7.3.1. Section 7.3.2 introduces the training procedure required to encourage the image generation network to learn 3D-consistent features and therefore improve its probability of converging.

7.3.1 Image Generator Architecture

Our approach can rely on any generation network that takes as input a 6D pose and outputs an image, *e.g.*, pose-conditioned GANs (Generative Adversarial Networks) [191, 192]. However, taking inspiration from previous works on NeRF-based pose-conditioned image generation [141, 193, 194], our method relies on a Neural Radiance Field [52]. NeRFs were initially designed to render novel views of a scene, given a set of training images depicting that scene. Here, the NeRF has to learn the 3D features exploited by a pose estimation network through the gradients back-propagated from that network.

The architecture of our NeRF-based image generator follows the *K*-planes [53] implementation presented in Chapter 3 (Figures 3.3 and 3.4). Given the input pose, rays are projected from the camera focal point through each pixel. For each ray, N points are sampled along the ray through a coarse density model of the scene. For each point, position and direction features are computed from its 3D position and viewing angles. The position features are fed in a first MLP which predicts density features associated with the point density. These density features, concatenated to the direction ones, are given to a second MLP to infer the color of the point. Finally, through ray-tracing techniques, the value of each pixel of the image is computed from the densities and colors of the points sampled on the corresponding ray.

Using a Neural Radiance field as image generator offers several key advantages. First, the NeRF representation is inherently **3D consistent** due to its architecture which relies on rays projected through the scene and the decoupling of the color field from the density one. Second, due to recent architectural improvements, such as the learned positional encoding [53, 118], NeRFs are **computationally efficient** and can therefore be trained on a consumer-grade GPU. Third, NeRFs are **modular**. Hence, to increase the convergence probability, we first train a NeRF on the synthetic dataset on which the pose estimator was trained. Then, we initialize a novel instance of that NeRF from scratch, except for the sampler and the learned positional encoding, which are copied from the pre-trained model and are not updated during the generator optimization. These three architectural characteristics significantly alleviate the convergence issues related to the problem formulation.

7.3.2 Ensuring 3D Perception Through Gradient Accumulation

Even if the convergence issues are partially solved through the image generator architecture, convergence is not guaranteed because of a key difference in the training strategies of NeRFs and pose estimators. Indeed, to ensure 3D consistency and promote convergence, each update of the NeRF weights should be derived from gradients computed across multiple viewpoints. When learned from a set of views, NeRFs can be trained with batches composed of pixels randomly sampled across all input images, effectively mixing viewpoints and therefore promoting 3D consistency. In contrast, pose estimators operate on entire images. Each update to the network's weights typically requires data from at least one full image, and ideally from multiple training images, although this is not strictly necessary. This constraint is at the heart of the difficulty of training a NeRF with a pose estimation loss.

Ideally, training a NeRF with a pose estimation loss should involve batches of full images corresponding to multiple viewpoints to ensure 3D perception. However, this would imply tremendous GPU memory requirements. We reduce these requirements by relying on two complementary mechanisms. First, to reduce the per-image generation complexity, we render images at half the resolution expected by the pose estimator and upsample them afterwards. Furthermore, we generate only the pixels that actually correspond to the target spacecraft. The remaining ones, corresponding to the outer space, are set to black. Second, to mimic larger batch sizes containing images taken from multiple viewpoints, we perform gradient accumulation.

Algorithm 1 describes the gradient accumulation strategy required to train the image generation network on a large effective batch size, thereby promoting the 3D consistency of the learned 3D cues. Each step of the algorithm corresponds to the processing of a single full image. First, an image $\hat{I}_s$ is rendered by the image generator M_Φ from a pose label $(q, t)_s$ sampled in the training dataset $\mathcal{D}$. Then, a loss $\mathcal{L}_s$ is computed between the prediction of the pose estimator F_Θ on that image and the pose label. Finally, the gradients of the image generator's weights are computed through backpropagation and accumulated. Every N steps, the generator weights Φ are updated by the optimizer O according to the average of the stored gradients.

Algorithm 1 Training M_Φ Through Gradient Accumulation

Require: Image generator M_Φ, pose estimation network F_Θ, optimizer O, loss $\mathcal{L}$, dataset D, accumulation steps N, total steps T

Initialize model parameters Φ
Initialize gradient accumulator $G \leftarrow 0$
for step $s = 1$ to T **do**
    Sample pose $(q, t)_s$ from dataset D
    Generate image $\hat{I}_s \leftarrow M_\Phi((q, t)_s)$
    Compute loss: $L_s \leftarrow \mathcal{L}(F_\Theta(\hat{I}_s), (q, t)_s)$
    Compute gradients: $g_s \leftarrow \nabla_\Phi L_s$
    Accumulate gradients: $G \leftarrow G + g_s$
    if $s \mod N = N - 1$ **then**
        Update parameters: $\Phi \leftarrow \Phi - O(G/N)$
        Reset accumulator: $G \leftarrow 0$
    end if
end for

7.4 Experiments

This section presents a set of experiments conducted with our feature visualization method on a spacecraft pose estimation network. Section 7.4.1 describes our validation setup, *i.e.*, architecture of the image generation and pose estimation networks and the generator optimization strategy. Section 7.4.2 demonstrates that our method recovers the 3D features that are primarily exploited by a pose estimation network and are sufficient to predict the correct relative pose. Section 7.4.3 exploits our visualization method to gain some insights on the learning mechanisms of spacecraft pose estimation networks.

7.4.1 Validation Setup

Our experiments are conducted on a pose estimator trained on the *Synthetic* set of SPEED+ [45] (see Chapter 4) and evaluated on the two corresponding HIL sets, *Lightbox* and *Sunlamp*. The pose estimation network F_Θ considered in these experiments is **SPNv2** [49], **a state-of-the-art spacecraft pose estimation network** extensively described in Section 2.5.2. Unless otherwise mentioned, SPNv2 is trained using the same parameters as in the original paper [49].

The image generation network M_Φ consists in a NeRF [52] that follows the K-planes implementation [53] (see Chapter 3, Figures 3.3 and 3.4) and is trained using our method. First, we train a NeRF on 200 synthetic images from SPEED+, resized to 256x384 pixels. Then, we initialize a new NeRF and update the sampler weights and the learned positional encoding with the corresponding pretrained weights. The weights of the sampler and positional encoding are frozen during training, only the density and color fields are trained. For each forward pass, only the foreground pixels are given to the network to reduce the GPU memory footprint. A full image of 256x384 pixels is assembled from the foreground ones, the background ones being set to black. The image is then upsampled by a factor 2 to achieve the resolution used during the pose estimation network training, *i.e.*, 512x768 pixels. The image is then normalized as expected by SPNv2 and fed into the pose estimation network which outputs K heatmaps and a pose.

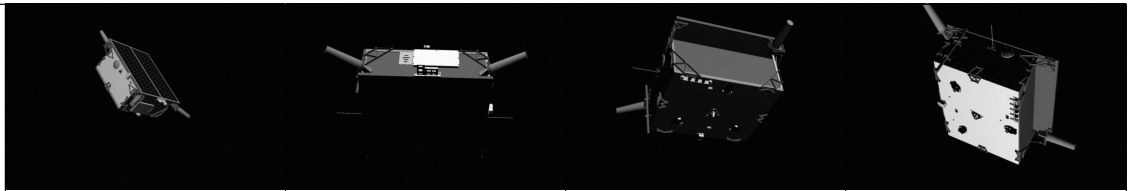
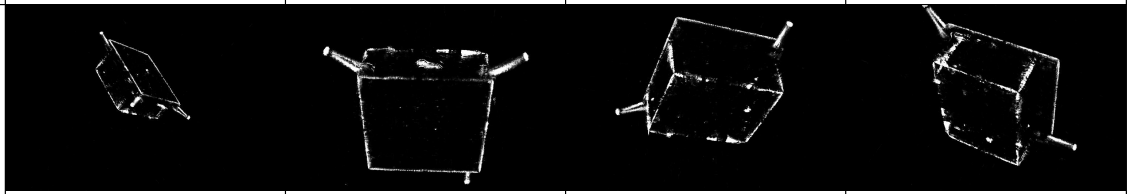
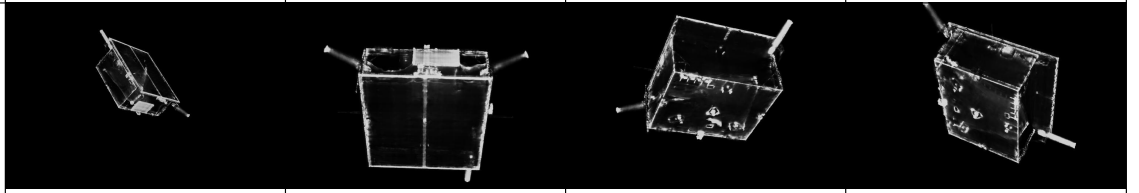
The loss back-propagated through F_Θ and M_Φ consists in the sum of two losses associated to the heatmaps and pose, both of them weighted by a factor β. This factor, determined through empirical evaluation, is set to 0.01. Indeed, experiments demonstrated that a factor of 0.01 or lower than this value leads to a probability of convergence of approximately 90 %. The first term is the L2-error between the predicted and target heatmaps while the second term is the SPEED loss [49] which combines the angular error with the normalized translation error. The weights of M_Φ are updated for T/N = 1,000 optimization steps, *i.e.*, each optimization step corresponding to N=10 images, while the weights of F_Θ are frozen. The optimization takes place on a single NVIDIA L40s and lasts about 30 minutes.

In the following sections, the angular error ($E^{(\mathrm{R})}$, in degrees) and the translation errors ($E^{(\mathrm{T})}$, in centimeters), introduced in Chapter 4, are computed on the prediction inferred by the pose estimation network on the corresponding image.

7.4.2 Validation

In this section, we demonstrate that our visualization method is able to recover the key 3D features that are primarily exploited by a given spacecraft pose estimation network. For this purpose, we trained two image generation networks M_Φ, based on the same NeRF architecture. Their only difference is the fact that the first network learns the whole neural field $\mathcal{F}$ (see Chapter 3, Figure 3.4) while the second network uses a pretrained positional encoding

Table 7.1 **(Top):** *Synthetic* images and the corresponding error metrics achieved by a pose estimator F_Θ. **(Middle-Bottom):** 3D cues rendered by a NeRF M_Φ trained with our method, under the same viewpoints considering two training configurations for M_Φ, either (i) the neural field $\mathcal{F}$ of M_Φ is trained from scratch **(Middle)** or (ii) the positional encoding of $\mathcal{F}$ is copied from the pretrained model and only the fields parameters are trained **(Bottom)**. In both cases, the learned visual cues are sufficient to enable the pose estimator to accurately predict the pose.

SPEED+ Synthetic images					
		$E^{(T)}$ = 1.6 cm $E^{(R)}$ = 0.38°	$E^{(T)}$ = 2.1 cm $E^{(R)}$ = 1.87°	$E^{(T)}$ = 2.1 cm $E^{(R)}$ = 0.94°	$E^{(T)}$ = 2.5 cm $E^{(R)}$ = 0.47°
3D cues used by F_Θ as learned by M_Φ	$\mathcal{F}$ trained w. fixed pos. enc.				
		$E^{(T)}$ = 1.9 cm $E^{(R)}$ = 0.83°	$E^{(T)}$ = 12.5 cm $E^{(R)}$ = 1.12°	$E^{(T)}$ = 10.9 cm $E^{(R)}$ = 0.83°	$E^{(T)}$ = 11.4 cm $E^{(R)}$ = 1.68°
	$\mathcal{F}$ fully trainable				
		$E^{(T)}$ = 2.8 cm $E^{(R)}$ = 0.94°	$E^{(T)}$ = 1.9 cm $E^{(R)}$ = 0.49°	$E^{(T)}$ = 4.8 cm $E^{(R)}$ = 1.19°	$E^{(T)}$ = 4.2 cm $E^{(R)}$ = 1.00°

which is frozen during training, *i.e.*, only the density and color fields are learned.

Table 7.1 depicts the visual cues learned by both networks for different viewpoints, along with the corresponding images from the SPEED+ dataset [45]. Not only do both networks converge, but they also generate similar visual cues. Those mainly consist in structural edges and asymmetrical features such as antennas, sun sensors, or a torque rod. However, the second model, *i.e.*, which uses a pretrained positional encoding, is able to recover smaller details. Furthermore, it only requires 1,000 optimization steps instead of the 3,000 steps required by the first one due to the regularization effect of the positional encoding. For these reasons, the next experiments are always conducted with an image generator for which the positional encoding was pretrained and is not updated during the optimization.

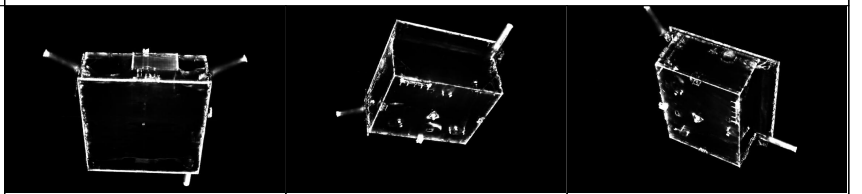
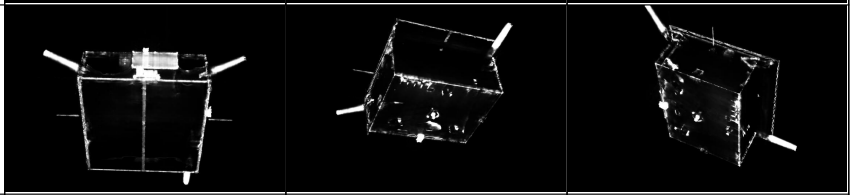
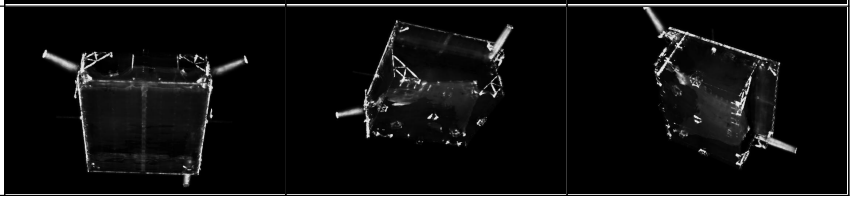
Table 7.1 also mentions the error metrics (see Section 4.3) computed on the predictions made by the pose estimator on the rendered visual cues. It shows that the pose estimation network achieves a comparable accuracy on the generated visual cues as on the synthetic images. These visual cues can therefore be interpreted as the main target features on which the pose estimator relies.

7.4.3 Insights on the Supervision of Pose Estimators

In the previous section, we considered an image generation network trained through the back-propagation of a loss that grasps the pose estimator ability of both detecting pre-defined keypoints and predicting the target relative pose. In this section, we consider two image generation networks learned through the back-propagation of either the heatmap-based keypoint regression loss or the pose regression one.

Table 7.2 depicts the images rendered by these image generation models. Visual cues from both models share similarities with the ones obtained through the combined loss. Indeed, they show that the model mostly relies on edges for both tasks. However, a key difference between both visualization models lies in the details that the model trained on the pose loss recovers. Indeed, this model learns subtle asymmetrical features such as the two thin antennas on the spacecraft sides. Together with the three strongly highlighted main antennas, they form a set of pose-relevant singularities. These singularities are especially useful for the pose estimation task as they are asymmetrical features that are distant from the spacecraft center,

Table 7.2 3D cues rendered by a NeRF M_Φ trained with our method on a pose estimator F_Θ, considering two training configurations for both M_Φ and for F_Θ. Training M_Φ through the pose loss encourages the learning of pose-relevant singularities (peripheral or asymmetrical), while the heatmap loss forces the reconstruction of edges. A pose estimation network trained only on a pose regression task relies on less robust and detailed features, which results in poorer generalization capabilities.

F_Θ Training	*Lightbox* Metrics	*Sunlamp* Metrics	M_Φ Supervision	Visual Cues
Heatmap & Pose	$E^{(R)} = 9.4°$ $E^{(T)} = 24$cm	$E^{(R)} = 15.5°$ $E^{(T)} = 28$cm	Heatmap	
			Pose	
Pose	$E^{(R)} = 24.5°$ $E^{(T)} = 46$ cm	$E^{(R)} = 34.9°$ $E^{(T)} = 38$ cm	Pose	

thereby enabling more accurate pose predictions. In contrast, the heatmap-based keypoint regression task does not really benefit from those features so that the image generation network trained on this loss does not learn to reconstruct them. Instead, it mainly relies on edges that are resilient to photometric variations and are relevant to the position of keypoints.

As we have seen, both tasks rely on similar features, except for the pose-relevant singularities exploited by the pose estimation head. An obvious question arises from this observation: *"Is the double supervision really required to train the pose estimation network ?"*. To answer that question, we first trained a SPNv2 model with only a pose estimation head. Then, we trained an image generator using the pose estimation loss. Table 7.2 depicts the visual cues learned by the image generation network.

The pose-supervised 3D cues no longer retain the distinct pose-relevant features previously observed. The thin side antennas of the spacecraft are missing from the reconstruction, and the three primary antennas lack their earlier contrast. Additionally, artifacts resembling some elements of the synthetic SPEED+ images have emerged, such as the solar panel texture. This indicates that the pose estimator trained only on the pose regression task has overfitted to the training set by learning a representation that is not as robust as the one learned through the pose estimator supervision of both tasks. This visual observation is consistent with the pose estimation accuracy metrics measured on the HIL sets of SPEED+ and reported in table 7.2 which show that the network poorly generalizes. Through the visualization of the 3D target features on which the model relies, our method therefore has the potential of indicating to which extent a pose estimation network has learned robust features, or, conversely, has overfitted to the training set distribution.

7.5 Conclusions

In this chapter, we introduced a method for visualizing the 3D cues which drive the predictions of a given spacecraft pose estimation network. This visualization is of a particular interest for understanding the decision process of spacecraft pose estimation networks and therefore alleviates the "black-box" effect that hampers their operational adoption.

By training an image generation network M_Φ using the gradients back-propagated through the pose estimator, we encourage the image generator

to recover the target 3D features that contribute the most to the pose estimator predictions. To solve the convergence issues induced by a challenging problem formulation, the image generation network consists in a Neural Radiance Field, trained through gradient accumulation to ensure the 3D consistency of the learned features.

Experiments carried out with a state-of-the-art spacecraft pose estimation network F_Θ demonstrate that the proposed method is able to recover the 3D cues on which the estimator relies. In addition, our method helps in gaining insights on the relationship between the pose estimation network supervision and the visual cues exploited by that estimator, and demonstrated the interest of Multi-Task Learning (MTL) in improving the generalization capabilities of a pose estimation network.

Contribution to Research Question #3

Research Question #3

How can we visualize the spacecraft features that a pose estimation network primarily relies on, in order to gain insights into its decision process and increase confidence in its predictions?

In this chapter, we developed a method to reveal the features that a pose estimation network primarily exploits. We train a pose-conditioned image generator M_Φ while keeping the estimator F_Θ frozen and back-propagate a pose loss so that the generated image leads the estimator to predict the input pose. Using a NeRF enforces three-dimensional consistency and gradient accumulation across viewpoints promotes coherent geometry. Because it must reproduce what the estimator considers sufficient for correct prediction, the rendered content exposes the decisive three-dimensional cues.

The visualizations show that the estimator relies mainly on robust structural edges and asymmetrical features that provide strong pose constraints. They also highlight that multi-task supervision encourages more robust cues and better generalization, while pose-only supervision yields fragile features and poorer OOD accuracy. This provides a powerful tool to visualize the basis of the network's decisions, diagnose overfitting, and increase confidence before deployment.

PART III
Perspectives & Conclusion

8

Conclusion

SATELLITES are critical assets in modern society, and the increase in orbital traffic has reinforced the need for autonomous and reliable close-proximity operations in OOS and ADR missions. This thesis addresses the vision-based estimation of the relative 6D pose of an uncooperative spacecraft through neural networks. It focuses on three challenges that hinder the adoption of such methods in autonomous RPOs: (i) removing the need for explicit CAD models, (ii) improving out-of-domain generalization, and (iii) increasing the transparency of pose estimators. These questions are investigated under a unified perspective based on NeRFs.

This chapter concludes the thesis. Section 8.1 synthesizes the findings and Section 8.2 summarizes the contributions. Section 8.3 discusses their practical limitations and associated future work. Section 8.4 outlines long-term perspectives, and Section 8.5 offers the closing remarks.

8.1 Synthesis of Findings

In Chapter 1, we highlighted three research questions that fundamentally hinder the adoption of data-driven pose estimation methods in close-proximity operations. Throughout this thesis, we address these questions under the prism of Novel View Synthesis (NVS) methods and more specifically NeRFs. This section synthesizes our findings.

Research Question #1

How can spacecraft pose estimation be enabled when an explicit prior model of the target's geometry and appearance cannot be assumed?

In Chapter 5, we show that a Neural Radiance Field M_Φ can be reconstructed from on-orbit imagery despite harsh illumination conditions and pose label uncertainty. In Chapter 6, we demonstrate that this representation can produce the training set for a pose estimator, thereby enabling training without an explicit CAD model of the target.

Underlying Principle. This approach fundamentally changes the source of information. Instead of requiring an explicit CAD model, it uses on-orbit imagery to learn a surrogate that captures the target's geometry and appearance at the level of fidelity required for view synthesis. The surrogate can then generate the same type of images that a CAD-based renderer would produce, which allows the pose estimator to be trained without changes to its design or training procedure.

Operational impact. This paradigm shift removes the need to exchange proprietary models between organizations, broadens OOS to targets outside the operator's fleet, and makes ADR feasible from a perception perspective. The result is a larger set of eligible targets and fewer legal or security constraints around model access.

Research Question #2

How can dataset-level information be aggregated to enable augmentations whose diversity surpasses the limits of image-wise transformations and improves Out-Of-Domain generalization?

We propose and evaluate a NeRF-based dataset-level augmentation method that diversifies appearance while preserving geometry. Training on the augmented set improves OOD generalization of the pose estimator across domain shifts, as detailed in Chapter 6. Compared to a training scenario where only viewpoints are randomized, the estimator achieves a reduction of 35% and 57% on *Lightbox* and *Sunlamp*, respectively.

Underlying Principle. NeRFs parameterize density and radiance as distinct fields. This enables augmentation of the target's appearance without affecting its geometry by randomizing the appearance embeddings and

perturbing the weights of the color MLP. Unlike image-grid augmentations, this generates images with diverse appearances while preserving the underlying geometry that should be learned by the pose estimator.

Operational impact. NeRF-based augmentation improves robustness to illumination conditions encountered in orbit. It reduces mission risk at deployment without changing the estimator architecture or runtime at the cost of an increase of the training complexity induced by the NeRF learning and the dataset generation that take place on the ground.

Research Question #3

How can we visualize the spacecraft features that a pose estimation network primarily relies on, in order to gain insights into its decision process and increase confidence in its predictions?

In chapter 7, we show that NVS models such as NeRFs can be repurposed to reveal the target features that contribute to the decision process of a spacecraft pose estimation network.

Underlying Principle. We back-propagate the pose error from an estimator F_Θ evaluated on images rendered by a NeRF M_Φ conditioned on ground-truth poses. The estimator being frozen, the gradients force M_Φ to encode the target cues that reduce the pose loss. The representation can then be queried to visualize these 3D cues and to analyze the estimator's decision process.

Operational impact. The approach shows that a state-of-the-art pose estimation network primarily relies on geometric features that are robust to illumination and texture variations, along with subtle asymmetric pose-relevant singularities. We further show that this combination arises from a hybrid pose prediction strategy that uses keypoint and pose supervisions. This work improves the estimator's transparency, which increases confidence in data-driven spacecraft pose estimation and facilitates its adoption.

8.2 Contributions

Figure 8.1 summarizes our three key contributions.

1. **Implicit 3D modeling from on-orbit monocular imagery.**
   We introduce a NeRF-based method to learn a 3D implicit target representation M_Φ^{Data} from 2D monocular images. Learnable appearance

embeddings and pose correction terms enable the model to handle the illumination variability and pose label uncertainty that affect on-orbit imagery. As demonstrated in Chapter 5, our method successfully recovers the target geometry and appearance on scarce image sets suffering from poor illumination conditions and pose uncertainty.

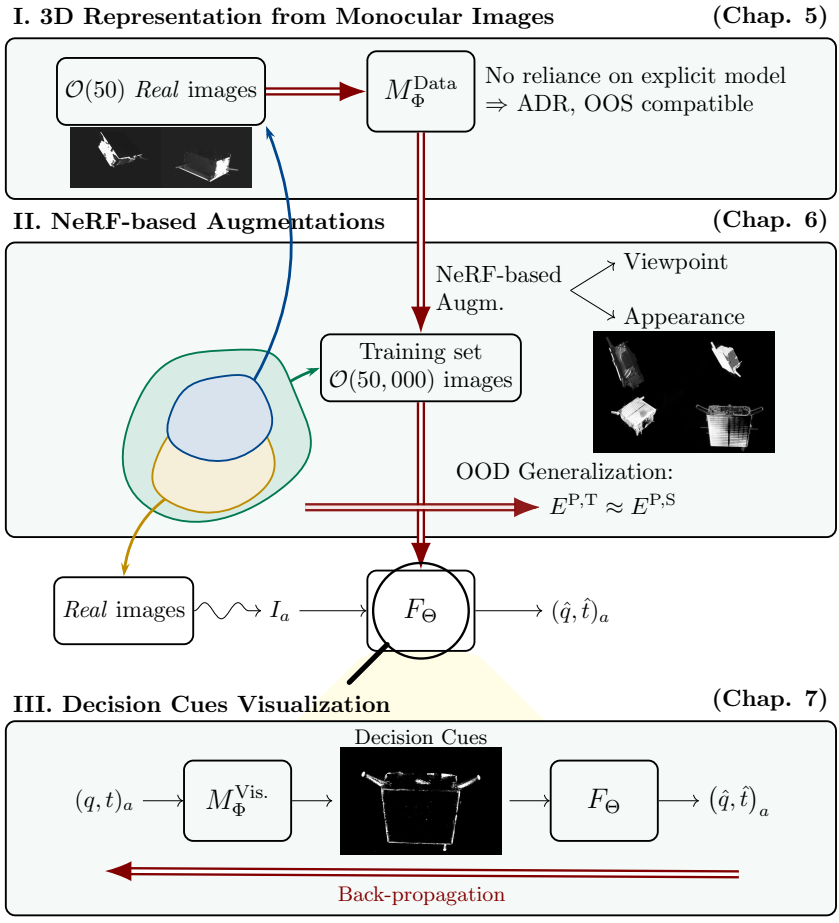


Fig. 8.1 Overview of the three contributions developed in this thesis. Two data-driven models, *i.e.*, implicit representation M_Φ and the pose estimator F_Θ, interact to (i) enable data generation without an explicit target model, (ii) improve out-of-domain generalization through dataset-level augmentation, and (iii) reveal the target visual features exploited by F_Θ to improve its transparency.

2. **Dataset-level augmentation beyond image-wise transforms.**
   We develop a dataset-level augmentation method based on a NeRF M_{Φ}^{Data} which enables viewpoint and appearance augmentation. The target depicted in the generated images is geometrically consistent, but its appearance is diversified through the randomization of the appearance embedding and color field. The method enables (i) the training of an estimator F_{Θ} without relying on a CAD model via viewpoint augmentation, and (ii) the improvement of the OOD generalization of F_{Θ} through appearance randomization, as shown in Chapter 6.
3. **Visualization of 3D cues exploited by a pose estimation network.**
   We repurpose a NeRF M_{Φ}^{Vis} to visualize the three-dimensional cues that drive the estimator's decisions by back-propagating the pose loss through a frozen estimator F_{Θ}. The analysis of these cues highlights a reliance on robust geometrical cues and subtle pose-relevant singularities, as discussed in Chapter 7.

8.3 Limitations & Future Works

This section highlights the limitations of this thesis' contributions and recommends research directions to address them.

Limitation 1: Lack of validation data

Our contributions were validated on the SPEED+ [45] and SHIRT [58] datasets. Despite being among the most comprehensive datasets for spacecraft pose estimation, owing to the availability of a large number of independent frames as well as sequences of both synthetic and HIL images of a relevant target spacecraft, they present two main limitations. First, the realism of the HIL images is limited because they are generated on the ground. Second, both SPEED+ and SHIRT contain images of a single spacecraft.

Limitation 1a: Lack of real on-orbit validation data

Our NeRF-based reconstruction and augmentation methods are validated on a testbed that emulates on-orbit conditions, but the imagery is not entirely realistic due to atmospheric diffusion and illumination mismatches. Furthermore, the captured images fail to faithfully reproduce the acquisition impairments observed in orbit, particularly those arising from radiation

exposure and long-term operational conditions, including changes in noise statistics, dead pixels, and readout failures.

The methods perform well under these emulated conditions, yet they must be validated on truly representative on-orbit data to confirm that their performance transfers to operational scenarios. A reliable solution would be an in-orbit acquisition campaign or an IOD to validate the full pipeline. Such validation would strengthen confidence in the methods and support future developments in spacecraft pose estimation.

Limitation 1b: Lack of diverse spacecraft geometries in validation data

SPEED+ and SHIRT only contain images of a single spacecraft, TANGO, whose geometry was illustrated in Section 4.2.1. It consists of a rectangular-prism body with a planar solar panel mounted on top. Three large antennas are located at the upper corners of the body, while two thinner antennas are centered on two opposing faces. Although this geometry may be challenging due to symmetries, it also contains asymmetrical features that mitigate such ambiguities.

While our experiments show that the proposed contributions improve a pose estimation network trained on this target geometry, we did not consider more challenging geometries, such as axisymmetrical upper-stage rockets, due to a lack of high-quality validation data for such geometries. Future works should consider the acquisition of images of additional spacecraft geometries in a HIL setup to evaluate how our methods, when applied to SPNv2, generalize to different target geometries. If SPNv2 were to perform poorly on images depicting an axisymmetrical spacecraft, future works should also consider the use of spacecraft pose estimation methods tailored to axisymmetrical targets [195, 196].

Limitation 2: Intrinsic depth ambiguity of monocular vision

Single-frame monocular estimates suffer from depth ambiguity, which leads to large translational errors on the order of 20 cm. Two directions can mitigate this ambiguity: adding depth measurements or exploiting temporal information. Depth sensors would provide richer information but increase payload mass and complexity. Learned multi-frame tracking models would require large datasets with long trajectories and illumination diversity. In contrast, a Kalman-based navigation filter offers a lightweight and explainable solution that requires no additional hardware or training data.

Preliminary experiments conducted during this thesis reveal two challenges for Kalman filtering. Single-frame estimators generate occasional outliers that can be filtered, but they also produce pose-dependent errors that are harder to model. The measurement noise follows a distribution $\mathcal{N}(\mu(q,t),\Sigma(q,t))$ with $(q,t) \in \mathrm{SE}(3)$, where both mean and covariance depend on the observed pose. Designing a navigation filter that accounts for this behavior is an essential direction for future work.

Limitation 3: Computational cost of NeRFs

Rendering the full augmented dataset of 48,000 views requires about 71 GPU-hours due to NeRFs' ray-tracing process. In comparison, 3D Gaussian Splatting (3D-GS) would likely reduce this cost to under 7 GPU-hours. We selected NeRFs because they are able to represent complex illumination more faithfully than 3D-GS, which is critical for spaceborne imagery.

Two research directions may combine NeRF-quality illumination handling with 3D-GS rendering speed. The first is to reduce the complexity imposed by the illumination conditions so that the 3D-GS representation could be suitable. For example, Park and D'Amico [150] improved the 3D-GS representation by using the ephemerides to exploit the Sun illumination angle to improve the rendering of the shadows on the target. The second is to adopt novel hybrid NVS representations [197–199] that combine the strengths of NeRF and 3D-GS. Such hybrid approaches would significantly reduce rendering time while preserving robustness to harsh illumination conditions.

8.4 Underlying Assumptions & Perspectives

"Your assumptions are your windows on the world.
Scrub them every once in a while.",
Alan Alda, 1980.

This section examines two assumptions that shape current spacecraft pose estimation and outlines long-term perspectives.

On the learning paradigm.

The current learning paradigm for spacecraft pose estimation follows a two-step approach. On the ground, a neural network F_Θ is trained on a dataset

of images generated using a model of a specific target. During this stage, F_Θ learns the target geometry. Once deployed on board, the network uses this learned representation to estimate the relative pose. This supervised learning strategy keeps the on-board computational load reasonable because the network is trained on the ground.

This paradigm, however, limits the scope of applicability. It requires either access to the target CAD model or an on-orbit tele-operated acquisition phase. Moreover, these estimators are target-specific and must be retrained for every new mission, which induces recurring engineering costs. A target-agnostic approach would therefore be highly valuable, as such a method would jointly learn the target geometry and estimate its pose from a stream of images.

This Simultaneous Localization and Mapping (SLAM) formulation has received extensive attention in computer vision and robotics, but it remains largely unexplored in space robotics. The on-orbit environment poses significant challenges to SLAM due to harsh illumination, strong distractors such as Earth in the background, and limited on-board computing capabilities. Recent advances in low-power GPUs and NVS-based SLAM methods [156–158] indicate that space-tailored SLAM systems may become feasible. A first step in this direction was recently made [159]. Although it was only validated on synthetic data, the environment-specific challenges are nevertheless solvable. Segmentation networks could be used to isolate the spacecraft and mitigate background distractors. Similarly, appearance embeddings could improve the representation ability of handling varying illumination.

On the target geometry.

Existing spacecraft pose estimation methods assume that the target is rigid. Under this assumption, the chaser-target configuration is characterized by six degrees of freedom. Networks are therefore trained to predict this 6D pose using datasets with diverse viewpoints and illumination conditions. However, many spacecraft include articulated components such as solar arrays or antennas, and others may exhibit structural flexibility. Current estimators, designed for rigid targets, cannot be applied directly in these cases.

From a mechanical point-of-view, each articulation introduces an additional degree of freedom. The operational scope of pose estimators could therefore be broadened by predicting a $6 + N$ degree-of-freedom pose. This

would require modifying the pose regression head and generating a training set with diversity in viewpoints, illumination, and articulation states. While feasible for targets with only a few joints, this approach becomes impractical for flexible structures. For such targets, an alternative is to estimate only the 6D pose of a rigid core while treating deformations as external variations. The training set would then need to include images that span a range of viewpoints, illumination conditions, and deformation configurations.

To the best of our knowledge, this question has not been addressed in the literature. One likely reason is the absence of suitable datasets. In the rigid case, considerable progress has been enabled by the SPEED [46] and SPEED+ [45] datasets released with the SPEC challenges [44, 83]. To support meaningful research on articulated or deformable spacecraft, the community should first generate and release datasets on non-rigid targets.

8.5 Closing Remarks

This thesis demonstrates how NeRFs provide a unified perspective on data-driven spacecraft pose estimation. They enable learning target representations from on-orbit imagery, generating training data without explicit CAD models, improving OOD generalization, and revealing the visual cues that drive neural pose estimators. These contributions are validated on realistic imagery under emulated on-orbit conditions. Together, they strengthen the reliability and transparency of learning-based estimators and extend their application scope across a broader set of missions. As autonomous RPOs become increasingly important for OOS and ADR, this thesis provides a foundation for more adaptable, robust, and interpretable pose estimation methods.

"Every day you may make progress. Every step may be fruitful. Yet there will stretch out before you an ever-lengthening, ever-ascending, ever-improving path. You know you will never get to the end of the journey. But this, so far from discouraging, only adds to the joy and glory of the climb."

— Winston S. Churchill

Acknowledgments

This work was funded by the Walloon Region and Aerospacelab S.A. as a *Win4Doc* project. We thank the Walloon Region and Aerospacelab S.A. for enabling and sponsoring this work, and for their commitment to scientific research.

Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

Bibliography

[1] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. Nerf-based spacecraft reconstruction from monocular imagery under illumination variability and pose uncertainty. *arXiv preprint arXiv:2605.18447*, 2026.

[2] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. Cad-free learning of spacecraft pose estimators via nerf-based augmentation. *arXiv preprint arXiv:2605.19649*, 2026.

[3] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. Nerf-based visualization of 3d cues supporting data-driven spacecraft pose estimation. In *2025 International Conference on Space Robotics (iSpaRo)*. IEEE, 2025.

[4] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. End-to-end neural estimation of spacecraft pose with intermediate detection of keypoints. In *European Conference on Computer Vision*, pages 154–169. Springer, 2022.

[5] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. Domain generalization for in-orbit 6d pose estimation. *Journal of Aerospace Information Systems*, 22(11):938–947, 2025.

[6] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. Leveraging neural radiance fields for pose estimation of an unknown space object during proximity operations. *2024 IEEE ISpaRo, available at arXiv:2405.12728*, 2024.

[7] Antoine Legrand, Renaud Detry, and Christophe De Vleeschouwer. Domain generalization for 6d pose estimation through nerf-based image synthesis. *arXiv preprint arXiv:2407.10762*, 2024.

[8] Antoine Legrand, Benoît Macq, and Christophe De Vleeschouwer. Forward error correction applied to jpeg-xs codestreams. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3723–3727. IEEE, 2022.

[9] Steven J Goodman, Richard J Blakeslee, William J Koshak, Douglas Mach, Jeffrey Bailey, Dennis Buechler, Larry Carey, Chris Schultz, Monte Bateman, Eugene McCaul Jr, et al. The goes-r geostationary lightning mapper (glm). *Atmospheric research*, 125:34–49, 2013.

[10] Rainer Hollmann, Chris J Merchant, Roger Saunders, Catherine Downy, Michael Buchwitz, Anny Cazenave, Emilio Chuvieco, Pierre Defourny, Gerrit de Leeuw, René Forsberg, et al. The esa climate change initiative: Satellite data records for essential climate variables. *Bulletin of the American Meteorological Society*, 94(10):1541–1552, 2013.

[11] David J Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4):358–371, 2013.

[12] Dmitry Rashkovetsky, Florian Mauracher, Martin Langer, and Michael Schmitt. Wildfire detection from multisensor satellite imagery using deep semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:7001–7016, 2021.

[13] Giancarlo Santilli, Cristian Vendittozzi, Chantal Cappelletti, Simone Battistini, and Paolo Gessini. Cubesat constellations for disaster management in remote areas. *Acta Astronautica*, 145:11–17, 2018.

[14] Gunter Schreier and Stefan Dech. High resolution earth observation satellites and services in the next decade–a european perspective. *Acta Astronautica*, 57(2-8):520–533, 2005.

[15] Oltjon Kodheli, Eva Lagunas, Nicola Maturo, Shree Krishna Sharma, Bhavani Shankar, Jesus Fabian Mendoza Montoya, Juan Carlos Merlano Duncan, Danilo Spano, Symeon Chatzinotas, Steven Kisseleff, et al. Satellite communications in the new space era: A survey and future challenges. *IEEE Communications Surveys & Tutorials*, 23(1): 70–109, 2020.

[16] Elliott D Kaplan and Christopher Hegarty. *Understanding GPS/GNSS: principles and applications*. Artech house, 2017.

[17] F Alby, E Lansard, and T Michal. Collision of cerise with space debris. In *Second European Conference on Space Debris*, volume 393, page 589, 1997.

[18] Eduardo Baptista. Damaged shenzhou-20 spacecraft to return to earth uncrewed for inspection. *Reuters*, https://www.reuters.com/science/damaged-shenzhou-20-spacecraft-return-earth-uncrewed-inspection-2025-12-01/, December 2025. Accessed: 2025-12-31.

[19] TS Kelso et al. Analysis of the iridium 33-cosmos 2251 collision. *Advances in the Astronautical Sciences*, 135(2):1099–1112, 2009.

[20] Sarah Thiele and Aaron C Boley. Investigating the risks of debris-generating asat tests in the presence of megaconstellations. *The Journal of the Astronautical Sciences*, 69(6):1797–1820, 2022.

[21] TS Kelso. Analysis of the 2007 chinese asat test and the impact of its debris on the space environment. In *8th Advanced Maui Optical and Space Surveillance Technologies Conference, Maui, HI*, volume 7, 2007.

[22] Donald J Kessler, Nicholas L Johnson, JC Liou, and Mark Matney. The kessler syndrome: implications to future space operations. *Advances in the Astronautical Sciences*, 137(8):2010, 2010.

[23] Federal Communications Commission. Fcc adopts new '5-year rule' for deorbiting satellites to address growing risk of orbital debris. https://www.fcc.gov/document/fcc-adopts-new-5-year-rule-deorbiting-satellites, September 2022.

[24] ISO 24113:2023 – space systems — space debris mitigation requirements. International Standard, 2023.

[25] European Commission. Proposal for an eu space act: Strengthening safety, resilience and sustainability in space. https://defence-industry-space.ec.europa.eu/eu-space-act_en, 2025.

[26] Andrew M Long, Matthew G Richards, and Daniel E Hastings. On-orbit servicing: a new value proposition for satellite design and operation. *Journal of Spacecraft and Rockets*, 44(4):964–976, 2007.

[27] Enrico Stoll, Juergen Letschnik, Ulrich Walter, Jordi Artigas, Philipp Kremer, Carsten Preusche, and Gerd Hirzinger. On-orbit servicing. *IEEE robotics & automation magazine*, 16(4):29–33, 2009.

[28] J-C Liou, Nicholas L Johnson, and NM Hill. Controlling the growth of future leo debris populations with active debris removal. *Acta Astronautica*, 66(5-6):648–653, 2010.

[29] Christophe Bonnal, Jean-Marc Ruault, and Marie-Christine Desjean. Active debris removal: Recent progress and current trends. *Acta Astronautica*, 85:51–60, 2013.

[30] Carl Glen Henshaw. The darpa phoenix spacecraft servicing program: Overview and plans for risk reduction. In *i-SAIRAS*. European Space Agency, 2014.

[31] Matt Pyrak and Joseph Anderson. Performance of northrop grumman's mission extension vehicle (mev) rpo imagers at geo. In *Autonomous Systems: Sensors, Processing and Security for Ground, Air, Sea and Space Vehicles and Infrastructure 2022*, volume 12115, pages 64–82. SPIE, 2022. doi: 10.1117/12.2631524.

[32] Jason L. Forshaw, Guglielmo S. Aglietti, Nimal Navarathinam, Haval Kadhem, Thierry Salmon, Aurélien Pisseloup, Eric Joffre, Thomas Chabot, Ingo Retat, Robert Axthelm, Simon Barraclough, Andrew Ratcliffe, Cesar Bernal, François Chaumette, Alexandre Pollini, and Willem H. Steyn. Removedebris: An in-orbit active debris removal demonstration mission. *Acta Astronautica*, 127:448–463, 2016. ISSN 0094-5765. doi: 10.1016/j.actaastro.2016.06.018.

[33] Robin Biesbroek, Sarmad Aziz, Andrew Wolahan, Ste-fano Cipolla, Muriel Richard-Noca, and Luc Piguet. The clearspace-1 mission: Esa and clearspace team up to remove debris. In *Proc. 8th Eur. Conf. Sp. Debris*, pages 1–3, 2021.

[34] Jacopo Ventura. *Autonomous proximity operations for noncooperative space targets*. PhD thesis, Technische Universität München, 2016.

[35] 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. doi: 10.1016/j.paerosci.2017.07.001.

[36] Roberto Opromolla, Giancarmine Fasano, Giancarlo Rufino, and Michele Grassi. Pose estimation for spacecraft relative navigation

using model-based algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 53(1):431–447, 2017. doi: 10.1109/TAES.2017.2650785.

[37] John A Christian and Scott Cryan. A survey of lidar technology and its use in spacecraft relative navigation. In *AIAA Guidance, Navigation, and Control (GNC) Conference*, page 4641, 2013. doi: 10.2514/6.2013-4641.

[38] Jian-Feng Shi, Steve Ulrich, Stephane Ruel, and Martin Anctil. Uncooperative spacecraft pose estimation using an infrared camera during proximity operations. In *AIAA SPACE 2015 conference and exposition*, page 4429, 2015. doi: 10.2514/6.2015-4429.

[39] Duarte Rondao, Nabil Aouf, and Mark A Richardson. Chinet: Deep recurrent convolutional learning for multimodal spacecraft pose estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 59(2): 937–949, 2022. doi: 10.1109/TAES.2022.3193085.

[40] Harvey Gómez Martínez, Gabriele Giorgi, and Bernd Eissfeller. Pose estimation and tracking of non-cooperative rocket bodies using time-of-flight cameras. *Acta Astronautica*, 139:165–175, 2017. doi: 10.1016/j.actaastro.2017.07.002.

[41] Mohsi Jawaid, Ethan Elms, Yasir Latif, and Tat-Jun Chin. Towards bridging the space domain gap for satellite pose estimation using event sensing. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11866–11873, 2023. doi: 10.1109/ICRA48891.2023.10160531.

[42] Vincenzo Pesce, Michèle Lavagna, and Riccardo Bevilacqua. Stereovision-based pose and inertia estimation of unknown and uncooperative space objects. *Advances in Space Research*, 59(1):236–251, 2017. doi: 10.1016/j.asr.2016.10.002.

[43] Sumant Sharma, Connor Beierle, and Simone D'Amico. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. In *2018 IEEE Aerospace Conference*, pages 1–12. IEEE, 2018.

[44] Mate Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Märtens, and Simone D'Amico. Satellite pose estimation challenge: Dataset, competition design, and results. *IEEE Transactions on Aerospace and Electronic Systems*, 56(5):4083–4098, 2020.

[45] Tae Ha Park, Marcus Märtens, Gurvan Lecuyer, Dario Izzo, and Simone D'Amico. Speed+: Next-generation dataset for spacecraft pose estimation across domain gap. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–15. IEEE, 2022.

[46] Sumant Sharma and Simone D'Amico. Neural network-based pose estimation for noncooperative spacecraft rendezvous. *IEEE Transactions on Aerospace and Electronic Systems*, 56(6):4638–4658, 2020.

[47] 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.

[48] 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.

[49] Tae Ha Park and Simone D'Amico. Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap. *Advances in Space Research*, 73(11):5726–5740, 2024.

[50] Lorenzo Pasqualetto Cassinis, Alessandra Menicucci, Eberhard Gill, Ingo Ahrns, and Manuel Sanchez-Gestido. On-ground validation of a cnn-based monocular pose estimation system for uncooperative spacecraft: Bridging domain shift in rendezvous scenarios. *Acta Astronautica*, 196:123–138, 2022.

[51] Maximilian Ulmer, Maximilian Durner, Martin Sundermeyer, Manuel Stoiber, and Rudolph Triebel. 6d object pose estimation from approximate 3d models for orbital robotics. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10749–10756. IEEE, 2023. doi: 10.1109/IROS55552.2023.10341511.

[52] 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.

[53] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance

fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[54] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[55] Leonard A McGee and Stanley F Schmidt. Discovery of the kalman filter as a practical tool for aerospace and industry. Technical report, National Aeronautics and Space Administration, 1985.

[56] Simon J Julier and Jeffrey K Uhlmann. New extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI*, volume 3068, pages 182–193. Spie, 1997.

[57] Lorenzo Pasqualetto Cassinis, Robert Fonod, Eberhard Gill, Ingo Ahrns, and Jesús Gil-Fernández. Evaluation of tightly-and loosely-coupled approaches in cnn-based pose estimation systems for uncooperative spacecraft. *Acta Astronautica*, 182:189–202, 2021.

[58] Tae Ha Park and Simone D'Amico. Adaptive neural-network-based unscented kalman filter for robust pose tracking of noncooperative spacecraft. *Journal of Guidance, Control, and Dynamics*, 46(9):1671–1688, 2023.

[59] Lorenzo Pasqualetto Cassinis, Tae Ha Park, Nathan Stacey, Simone D'Amico, Alessandra Menicucci, Eberhard Gill, Ingo Ahrns, and Manuel Sanchez-Gestido. Leveraging neural network uncertainty in adaptive unscented kalman filter for spacecraft pose estimation. *Advances in Space Research*, 71(12):5061–5082, 2023.

[60] Tae Ha Park and Simone D'Amico. Online supervised training of spaceborne vision during proximity operations using adaptive kalman filtering. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11744–11752. IEEE, 2024.

[61] Nuno Filipe, Michail Kontitsis, and Panagiotis Tsiotras. Extended kalman filter for spacecraft pose estimation using dual quaternions. *Journal of Guidance, Control, and Dynamics*, 38(9):1625–1641, 2015.

[62] Brent E Tweddle and Alvar Saenz-Otero. Relative computer vision-based navigation for small inspection spacecraft. *Journal of guidance, control, and dynamics*, 38(5):969–978, 2015.

[63] Siddarth Kaki and Maruthi R Akella. Inertia-free pose and angular velocity estimation using monocular vision. In *33rd AAS/AIAA Space Flight Mechanics Meeting, AAS Paper*, pages 23–210, 2023.

[64] 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.

[65] Sumant Sharma et al. Comparative assessment of techniques for initial pose estimation using monocular vision. *Acta Astronautica*, 123: 435–445, 2016.

[66] Sumant Sharma, Jacopo Ventura, and Simone D'Amico. Robust model-based monocular pose initialization for noncooperative spacecraft rendezvous. *Journal of Spacecraft and Rockets*, 55(6):1414–1429, 2018.

[67] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[68] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[70] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 770–778, 2016.

[72] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[73] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[74] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pages 2286–2296. PMLR, 2021.

[75] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

[78] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009.

[79] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8109, 2020.

[80] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 2781–2790, 2022.

[81] Fulin Liu, Yinlin Hu, and Mathieu Salzmann. Linear-covariance loss for end-to-end learning of 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14107–14117, 2023.

[82] Haoran Huang, Bin Song, Gaopeng Zhao, and Yuming Bo. End-to-end monocular pose estimation for uncooperative spacecraft based on direct regression network. *IEEE Transactions on Aerospace and Electronic Systems*, 59(5):5378–5389, 2023.

[83] Tae Ha Park, Marcus Märtens, Mohsi Jawaid, Zi Wang, Bo Chen, Tat-Jun Chin, Dario Izzo, and Simone D'Amico. Satellite pose estimation competition 2021: Results and analyses. *Acta Astronautica*, 204:640–665, 2023.

[84] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[85] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

[86] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*, 2018.

[87] Dimitrios Sakkos, Hubert PH Shum, and Edmond SL Ho. Illumination-based data augmentation for robust background subtraction. In *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pages 1–8. IEEE, 2019.

[88] Philip TG Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, volume 6, pages 10–11, 2019.

[89] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. doi: 10.1109/CVPR46437.2021.01415.

[90] Qinwei Xu, Ruipeng Zhang, Ziqing Fan, Yanfeng Wang, Yi-Yan Wu, and Ya Zhang. Fourier-based augmentation with applications to domain generalization. *Pattern Recognition*, 139:109474, 2023. doi: 10.1016/j.patcog.2023.109474.

[91] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. doi: 10.1109/IROS.2017.8202133.

[92] 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.

[93] Julien Posso, Guy Bois, and Yvon Savaria. Mobile-ursonet: an embeddable neural network for onboard spacecraft pose estimation. In *2022 IEEE international symposium on circuits and systems (ISCAS)*, pages 794–798. IEEE, 2022.

[94] George Lentaris, Ioannis Stratakos, Ioannis Stamoulias, Dimitrios Soudris, Manolis Lourakis, and Xenophon Zabulis. High-performance vision-based navigation on soc fpga for spacecraft proximity operations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1188–1202, 2019.

[95] Kiruki Cosmas and Asami Kenichi. Utilization of fpga for onboard inference of landmark localization in cnn-based spacecraft pose estimation. *Aerospace*, 7(11):159, 2020.

[96] Julien Posso, Guy Bois, and Yvon Savaria. Real-time spacecraft pose estimation using mixed-precision quantized neural network on cots reconfigurable mpsoc. In *2024 22nd IEEE Interregional NEWCAS Conference (NEWCAS)*, pages 358–362. IEEE, 2024.

[97] Windy S. Slater, Nayana P. Tiwari, Tyler M. Lovelly, and Jesse K. Mee. Total ionizing dose radiation testing of nvidia jetson nano gpus. In *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–3, 2020. doi: 10.1109/HPEC43674.2020.9286222.

[98] Michael A Felix, Windy S Slater, Derrek C Landauer, Ryan E Pinson, and Benjamin BW Rutherford. Total ionizing dose radiation testing of nvidia jetson orin nx system on module. In *2024 IEEE Space Computing Conference (SCC)*, pages 116–121. IEEE, 2024.

[99] Roberto Opromolla, Dmitriy Grishko, John Auburn, Riccardo Bevilacqua, Luisa Buinhas, Joseph Cassady, Markus Jäger, Marko Jankovic, Javier Rodriguez, Maria Antonietta Perino, et al. Future in-orbit servicing operations in the space traffic management context. *Acta Astronautica*, 220:469–477, 2024.

[100] Stéphane Estable, Clément Pruvost, Eugénio Ferreira, Jürgen Telaar, Michael Fruhnert, Christian Imhof, Tomasz Rybus, Gregory Peckover, Robert Lucas, Rohaan Ahmed, et al. Capturing and deorbiting envisat with an airbus spacetug. results from the esa e. deorbit consolidation phase study. *Journal of Space Safety Engineering*, 7(1):52–66, 2020.

[101] Tae Ha Park, Juergen Bosse, and Simone D'Amico. Robotic testbed for rendezvous and optical navigation: Multi-source calibration and machine learning use cases. *arXiv preprint arXiv:2108.05529*, 2021.

[102] Tae Ha Park and Simone D'Amico. Bridging the domain gap for flight-ready spaceborne vision. *Journal of Spacecraft and Rockets*, pages 1–19, 2025.

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

[104] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[105] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020.

[106] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[107] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European conference on computer vision*, pages 696–712. Springer, 2020.

[108] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5438–5448, 2022.

[109] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5752–5761, 2021.

[110] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5459–5469, 2022.

[111] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[112] Jan Held, Renaud Vandeghen, Abdullah Hamdi, Adrien Deliege, Anthony Cioppa, Silvio Giancola, Andrea Vedaldi, Bernard Ghanem, and Marc Van Droogenbroeck. 3d convex splatting: Radiance field rendering with 3d smooth convexes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21360–21369, 2025.

[113] Jan Held, Renaud Vandeghen, Adrien Deliege, Abdullah Hamdi, Silvio Giancola, Anthony Cioppa, Andrea Vedaldi, Bernard Ghanem, Andrea Tagliasacchi, and Marc Van Droogenbroeck. Triangle splatting for real-time radiance field rendering. *arXiv preprint arXiv:2505.19175*, 2025.

[114] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[115] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[116] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34:4805–4815, 2021.

[117] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[118] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[119] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021.

[120] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012.

[121] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[122] Mohamed Adel Musallam, Vincent Gaudilliere, Enjie Ghorbel, Kassem Al Ismaeil, Marcos Damian Perez, Michel Poucet, and Djamila Aouada. Spacecraft recognition leveraging knowledge of space environment: Simulator, dataset, competition design and analysis. In *2021*

IEEE International Conference on Image Processing Challenges (ICIPC), pages 11–15. IEEE, 2021.

[123] Hoang Anh Dung, Bo Chen, and Tat-Jun Chin. A spacecraft dataset for detection, segmentation and parts recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2019, 2021.

[124] Yadong Shao, Aodi Wu, Shengyang Li, Leizheng Shu, Xue Wan, Yuanbin Shao, and Junyan Huo. Satellite component semantic segmentation: Video dataset and real-time pyramid attention and decoupled attention network. *IEEE Transactions on Aerospace and Electronic Systems*, 59(6):7315–7333, 2023.

[125] Arunkumar Rathinam, Haytam Qadadri, and Djamila Aouada. Spades: A realistic spacecraft pose estimation dataset using event sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11760–11766. IEEE, 2024.

[126] Yinlin Hu, Sebastien Speierer, Wenzel Jakob, Pascal Fua, and Mathieu Salzmann. Wide-depth-range 6d object pose estimation in space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15870–15879, 2021.

[127] Zeeshan Ahmed, T. H. Park, Arpan Bhattacharjee, Rahman Razel-Rezai, Robert Graves, Olli Saarela, Ryo Teramoto, Kiran Vemulapalli, and Simone D'Amico. Speed-ue-cube: A machine learning dataset for autonomous, vision-based spacecraft navigation. In *46th Rocky Mountain AAS Guidance, Navigation and Control Conference*, Breckenridge, Colorado, February 2024.

[128] Tae Ha Park and Simone D'Amico. Rapid abstraction of spacecraft 3d structure from single 2d image. In *AIAA SCITECH 2024 Forum*, page 2768, 2024.

[129] Szabolcs Velkei, Csaba Goldschmidt, and Károly Vass. A large-scale, physically-based synthetic dataset for satellite pose estimation. *arXiv preprint arXiv:2506.12782*, 2025.

[130] Mohamed Adel Musallam, Arunkumar Rathinam, Vincent Gaudillière, Miguel Ortiz del Castillo, and Djamila Aouada. Cubesat-cdt: A cross-domain dataset for 6-dof trajectory estimation of a symmetric

spacecraft. In *European Conference on Computer Vision*, pages 112–126. Springer, 2022.

[131] Jérémy Lebreton, Ingo Ahrns, Roland Brochard, Christoph Haskamp, Hans Krüger, Matthieu Le Goff, Nicolas Menga, Nicolas Ollagnier, Ralf Regele, Francesco Capolupo, et al. Training datasets generation for machine learning: Application to vision based navigation. *arXiv preprint arXiv:2409.11383*, 2024.

[132] Eberhard Gill, Simone D'Amico, and Oliver Montenbruck. Autonomous formation flying for the prisma mission. *Journal of Spacecraft and Rockets*, 44(3):671–681, 2007.

[133] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[134] Rony Ferzli and Lina J Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE transactions on image processing*, 18(4):717–728, 2009.

[135] Nikos Mavrakis, Zhou Hao, and Yang Gao. On-orbit robotic grasping of a spent rocket stage: Grasp stability analysis and experimental results. *Frontiers in Robotics and AI*, 8:652681, 2021.

[136] Roshan Sah, Raunak Srivastava, Vighnesh Vatsal, Rolif Lima, and Kaushik Das. Target grasp-conditioned whole-body control of satellite manipulators using neural fields with mpc. In *2024 IEEE Aerospace Conference*, pages 1–12. IEEE, 2024.

[137] Aneesh M Heintz and Mason Peck. Spacecraft state estimation using neural radiance fields. *Journal of Guidance, Control, and Dynamics*, 46 (8):1596–1609, 2023.

[138] Anne Mergy, Gurvan Lecuyer, Dawa Derksen, and Dario Izzo. Vision-based neural scene representations for spacecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2021.

[139] Van Minh Nguyen, Emma Sandidge, Trupti Mahendrakar, and Ryan T White. Characterizing satellite geometry via accelerated 3d gaussian splatting. *Aerospace*, 11(3):183, 2024.

[140] Marcus Klasson, Riccardo Mereu, Juho Kannala, and Arno Solin. Sources of uncertainty in 3d scene reconstruction. In *European Conference on Computer Vision*, pages 271–289. Springer, 2024.

[141] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in neural information processing systems*, 33:20154–20166, 2020.

[142] Basilio Caruso, Trupti Mahendrakar, Van Minh Nguyen, Ryan T White, and Todd Steffen. 3d reconstruction of non-cooperative resident space objects using instant ngp-accelerated nerf and d-nerf. *arXiv preprint arXiv:2301.09060*, 2023.

[143] Tao Fu, Yu Zhou, Ying Wang, Jian Liu, Yamin Zhang, Qinglei Kong, and Bo Chen. Neural field-based space target 3d reconstruction with predicted depth priors. *Aerospace*, 11(12):997, 2024.

[144] Clément Forray, Pauline Delporte, Nicolas Delaygue, Florence Genin, and Dawa Derksen. Joint attitude estimation and 3d neural reconstruction of non-cooperative space objects. *arXiv preprint arXiv:2506.20638*, 2025.

[145] Yang Liu, Zhihao Sun, Lele Zhang, Lele Xi, Wei Dong, and Fang Deng. Spacecraft-nerf: High-fidelity reconstruction of spacecraft by neural radiance field based implicit representation. *IEEE Transactions on Aerospace and Electronic Systems*, pages 1–13, 2025. doi: 10.1109/TAES.2025.3539273.

[146] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021.

[147] Bing Han, Chenxi Wang, Xinyu Zhang, Zhibin Zhao, Zhi Zhai, Jinxin Liu, Naijin Liu, and Xuefeng Chen. Pose estimation and neural implicit reconstruction towards non-cooperative spacecraft without offline prior information. *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[148] Qicheng Xu, Min Hu, Yuqiang Fang, and Xitao Zhang. A neural radiance fields method for 3d reconstruction of space target. *Advances in Space Research*, 75(9):6924–6943, 2025.

[149] Arianna Issitt, Trupti Mahendrakar, Amy Alvarez, Ryan T White, and Alex Sizemore. On optimal observation orbits for learning gaussian splatting-based 3d models of unknown resident space objects. In *AIAA SCITECH 2025 Forum*, page 1780, 2025.

[150] Tae Ha Park and Simone D'Amico. Improved 3d gaussian splatting of unknown spacecraft structure using space environment illumination knowledge. In *2025 International Conference on Space Robotics (iSpaRo)*. IEEE, 2025.

[151] Elias De Smijter, Renaud Detry, and Christophe De Vleeschouwer. On the geometric accuracy of implicit and primitive-based representations derived from view rendering constraints. *arXiv preprint arXiv:2509.10241*, 2025.

[152] Nidhi Mathihalli, Audrey Wei, Giovanni Lavezzi, Peng Mun Siew, Victor Rodriguez-Fernandez, Hodei Urrutxua, and Richard Linares. Dreamsat: Towards a general 3d model for novel view synthesis of space objects. *arXiv preprint arXiv:2410.05097*, 2024.

[153] Tae Ha Park, Emily Bates, and Simone D'Amico. Improving zero-shot abstraction of unknown spacecraft 3d shape as primitive assembly. In *International Workshop on Satellite Constellations and Formation Flying*, 2024.

[154] Emily Bates and Simone D'Amico. Removing ambiguities in concurrent monocular single-shot spacecraft shape and pose estimation using a deep neural network. In *2025 AAS GNC conference (Feb 2025)*, 2025.

[155] Pol Francesch Huc, Emily Bates, and Simone D'Amico. Fast learning of non-cooperative spacecraft 3d models through primitive initialization. *arXiv preprint arXiv:2507.19459*, 2025.

[156] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6229–6238, 2021.

[157] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 12786–12796, 2022.

[158] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023.

[159] Kuldeep R Barad, Antoine Richard, Jan Dentler, Miguel Olivares-Mendez, and Carol Martinez. Object-centric reconstruction and tracking of dynamic unknown objects using 3d gaussian splatting. In *2024 International Conference on Space Robotics (iSpaRo)*, pages 202–209. IEEE, 2024.

[160] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023.

[161] Hongyu Fu, Xin Yu, Lincheng Li, and Li Zhang. Cbarf: cascaded bundle-adjusting neural radiance fields from imperfect camera poses. *IEEE Transactions on Multimedia*, 26:9304–9315, 2024.

[162] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021.

[163] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. *arXiv preprint arXiv:2210.10108*, 2022.

[164] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021.

[165] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[166] Sophie Duzellier, Paulo Gordo, Rui Melicio, Duarte Valério, Mark Millinger, and António Amorim. Space debris generation in geo: Space materials testing and evaluation. *Acta Astronautica*, 192:258–275, 2022.

[167] Ke Wang, Bin Fang, Jiye Qian, Su Yang, Xin Zhou, and Jie Zhou. Perspective transformation data augmentation for object detection. *IEEE Access*, 8:4935–4943, 2019.

[168] Martin Engilberge, Haixin Shi, Zhiye Wang, and Pascal Fua. Two-level data augmentation for calibrated multi-view detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 128–136, 2023.

[169] Jinfan Zhou, Lixin Luo, Sungmin Eum, Heesung Kwon, and Jeong Joon Park. Generative spatiotemporal data augmentation. *arXiv preprint arXiv:2512.12508*, 2025.

[170] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8575–8584, 2021.

[171] Casimir Feldmann, Niall Siegenheim, Nikolas Hars, Lovro Rabuzin, Mert Ertugrul, Luca Wolfart, Marc Pollefeys, Zuria Bauer, and Martin R Oswald. Nerfmentation: Improving monocular depth estimation with nerf-based data augmentation. In *European Conference on Computer Vision*, pages 92–108. Springer, 2024.

[172] Shouwei Ruan, Yinpeng Dong, Hang Su, Jianteng Peng, Ning Chen, and Xingxing Wei. Towards viewpoint-invariant visual recognition via adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4709–4719, 2023.

[173] Yunhao Ge, Harkirat Behl, Jiashu Xu, Suriya Gunasekar, Neel Joshi, Yale Song, Xin Wang, Laurent Itti, and Vibhav Vineet. Neural-sim: Learning to generate training data with nerf. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022.

[174] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic

data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018.

[175] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Marek Martina, and Martin Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[176] Sergey Zakharov, Rareș Ambruș, Vitor Guizilini, Wadim Kehl, and Adrien Gaidon. Photo-realistic neural domain randomization. In *European Conference on Computer Vision*, pages 310–327. Springer, 2022.

[177] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2100–2110, 2019.

[178] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020.

[179] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

[180] Zi Wang, Minglin Chen, Yulan Guo, Zhang Li, and Qifeng Yu. Bridging the domain gap in satellite pose estimation: a self-training approach based on geometrical constraints. *IEEE Transactions on Aerospace and Electronic Systems*, 2023.

[181] Juan Ignacio Bravo Pérez-Villar, Álvaro García-Martín, and Jesús Bescós. Spacecraft pose estimation based on unsupervised domain adaptation and on a 3d-guided loss combination. In *European Conference on Computer Vision*, pages 37–52. Springer, 2022.

[182] Kiruki Cosmas and Asami Kenichi. Utilization of fpga for onboard inference of landmark localization in cnn-based spacecraft pose estimation. *Aerospace*, 7(11):159, 2020. doi: 10.3390/aerospace7110159.

[183] Vasileios Leon, George Lentaris, Dimitrios Soudris, Simon Vellas, and Mathieu Bernou. Towards employing fpga and asip acceleration to enable onboard ai/ml in space applications. In *2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC)*, pages 1–4. IEEE, 2022. doi: 10.1109/VLSI-SoC54400.2022.9939566.

[184] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[185] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[186] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[187] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[188] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[189] Alexandre Englebert, Olivier Cornu, and Christophe De Vleeschouwer. Poly-cam: high resolution class activation map for convolutional neural networks. *Machine Vision and Applications*, 35(4): 89, 2024.

[190] Alexandre Englebert, Olivier Cornu, and Christophe De Vleeschouwer. Backward recursive class activation map refinement for high resolution saliency map. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2444–2450. IEEE, 2022.

[191] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

[192] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7588–7597, 2019.

[193] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[194] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.

[195] Shintaro Hashimoto, Daichi Hirano, and Naoki Ishihama. 6-dof pose estimation for axisymmetric objects using deep learning with uncertainty. In *2020 IEEE Aerospace Conference*, pages 1–12. IEEE, 2020.

[196] Toma Takahashi, Yudai Furuta, and Toshinori Kuwahara. Two-stage cnn-based pose estimation for uncooperative axisymmetric spacecraft. In *2025 International Conference on Space Robotics (iSpaRo)*, pages 291–298. IEEE, 2025.

[197] Yanshun Chen, Weiqing Yan, Guanghui Yue, and Wujie Zhou. Multi-resolution hybrid explicit representation for novel view synthesis of dynamic scenes. *IEEE Transactions on Intelligent Vehicles*, 2024.

[198] Xilong Zhou, Bao-Huy Nguyen, Loïc Magne, Vladislav Golyanik, Thomas Leimkühler, and Christian Theobalt. Splat the net: Radiance fields with splattable neural primitives. *arXiv preprint arXiv:2510.08491*, 2025.

[199] Hadi Alzayer, Philipp Henzler, Jonathan T Barron, Jia-Bin Huang, Pratul P Srinivasan, and Dor Verbin. Generative multiview relighting for 3d reconstruction under extreme illumination variation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10933–10942, 2025.