aerospace

Article

A Vision-Based Pose Estimation of a Non-Cooperative Target Based on a Self-Supervised Transformer Network

Quan Sun [1,2], Xuhui Pan [1], Xiao Ling [1], Bo Wang [1], Qinghong Sheng [1], Jun Li [1], Zhijun Yan [1], Ke Yu [2] and Jiasong Wang [3,*]

[1] College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; iamsunquan@126.com (Q.S.); xuhuipan@nuaa.edu.cn (X.P.); xlingsky@nuaa.edu.cn (X.L.); wangbo_nuaa@nuaa.edu.cn (B.W.); qhsheng@nuaa.edu.cn (Q.S.); jun.li@nuaa.edu.cn (J.L.); yanzj@nuaa.edu.cn (Z.Y.)

[2] Shanghai Electro-Mechanical Engineering Institute, Shanghai 200041, China; zjuxyk@163.com

[3] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

* Correspondence: wjssquid@163.com

Abstract: In the realm of non-cooperative space security and on-orbit service, a significant challenge is accurately determining the pose of abandoned satellites using imaging sensors. Traditional methods for estimating the position of the target encounter problems with stray light interference in space, leading to inaccurate results. Conversely, deep learning techniques require a substantial amount of training data, which is especially difficult to obtain for on-orbit satellites. To address these issues, this paper introduces an innovative binocular pose estimation model based on a Self-supervised Transformer Network (STN) to achieve precise pose estimation for targets even under poor imaging conditions. The proposed method generated simulated training samples considering various imaging conditions. Then, by combining the concepts of convolutional neural networks (CNN) and SIFT features for each sample, the proposed method minimized the disruptive effects of stray light. Furthermore, the feedforward network in the Transformer employed in the proposed method was replaced with a global average pooling layer. This integration of CNN's bias capabilities compensates for the limitations of the Transformer in scenarios with limited data. Comparative analysis against existing pose estimation methods highlights the superior robustness of the proposed method against variations caused by noisy sample sets. The effectiveness of the algorithm is demonstrated through simulated data, enhancing the current landscape of binocular pose estimation technology for non-cooperative targets in space.

Keywords: non-cooperative targets; stray light interference; vision-based pose estimation; self-supervised transformer network

check for updates

Citation: Sun, Q.; Pan, X.; Ling, X.; Wang, B.; Sheng, Q.; Li, J.; Yan, Z.; Yu, K.; Wang, J. A Vision-Based Pose Estimation of a Non-Cooperative Target Based on a Self-Supervised Transformer Network. *Aerospace* **2023**, *10*, 997. https://doi.org/10.3390/aerospace10120997

Academic Editor: M. Reza Emami

Received: 23 September 2023
Revised: 10 November 2023
Accepted: 23 November 2023
Published: 28 November 2023

1. Introduction

As human exploration and development of outer space advances, countries demand higher levels of space technology [1]. Some of the key challenges in the aerospace field are spacecraft rendezvous and docking, on-orbit capture and repair of malfunctioning satellites, and space debris removal [2]. These challenges require the ability to perform rendezvous, docking, and capture of non-cooperative targets [3]. However, this task depends on the relative pose measurement of non-cooperative targets, which is difficult to achieve due to the poor quality of space images. Space images often have low contrast and texture and are affected by stray light in space. Non-cooperative targets lack artificial markers and feature cursors for auxiliary measurement, making it hard to obtain geometric, grayscale, depth, and other information about the target surface [4]. Various factors limit the availability of samples, which poses problems and challenges for attitude measurement.

There are various methods to achieve the pose measurement of non-cooperative targets, depending on the sensors used. These methods include visual target measurement, scanning laser radar measurement, non-scanning three-dimensional laser imaging measurement [5], pose measurement method based on multi-sensor fusion [6], and so on. The visual measurement method uses a camera to obtain the target image. This method is simple and does not require complex structures or too many devices. It can measure the target with only a camera and a computer, but it requires high computing power. Binocular vision can calculate the target distance and real size using the principle of triangulation, which is more suitable for the pose measurement of space non-cooperative targets [7]. However, this method also requires that the pose estimation algorithm can detect and process image feature information. Moreover, the optical images are more vulnerable to stray light, which affects the recognition and detection of space targets and indirectly leads to the scarcity of data set samples.

Currently, deep learning methods have been applied to various fields beyond image recognition, and the Transformer model is a rising star in the field of non-cooperative target detection and recognition. After the introduction of the Transformer structure from natural language processing to computer vision, it has broken the limited receptive field constraint of CNN. It has gained significant attention due to its advantages, such as not requiring proposals like Faster R-CNN, not using anchors like YOLO, not needing centers or post-processing steps like NMS, as in CenterNet, and directly predicting detection boxes and classes. The Backbone, as a feature extraction network, primarily extracts relevant information from images for subsequent stages. The role of the Neck is to fuse and enhance the features extracted by the Backbone before providing them to the Head for detection. The Head employs the previously extracted features to predict the position and class of objects [8]. As a target detection method, DETR transformed Transformers into the field of object detection, opening up new research avenues [9]. YOLOS is a series of ViT-based object detection models with minimal modifications and inductive biases [10]. Additionally, DETR has various related variants. To address the slow convergence issue of DETR, researchers proposed Deformable DETR and TSP-FCOS and TSP-RCNN [11,12]. Deformable DETR uses deformable convolution to effectively solve the slow convergence and low detection accuracy for small objects in sparse spatial positioning. ATC primarily alleviates redundancy in the attention maps of DETR and the problem of feature redundancy as the encoder deepens. It is evident that the Transformer network in the Neck section has mature research solutions that can significantly enhance accuracy. Furthermore, in the context of non-cooperative target issues, appropriate modifications can prevent the loss of information when reading patch information. This approach can retain more feature information, considering the scarcity of information sources.

Therefore, this paper proposes the application of deep learning methods Transformer to the problem of spatial target feature detection and recognition while considering external interference factors present in space missions in Figure 1. It fully leverages the characteristics of CNN models to enhance accuracy and sensitivity to small-scale features. Through improved network structures, it further refines the algorithm for non-cooperative target pose estimation in scenarios with limited sample sizes. Validation and comparative experiments are conducted using satellite sample data generated in a virtual environment with strong noise interference, confirming the reliability of the proposed algorithm under conditions of limited data volume.

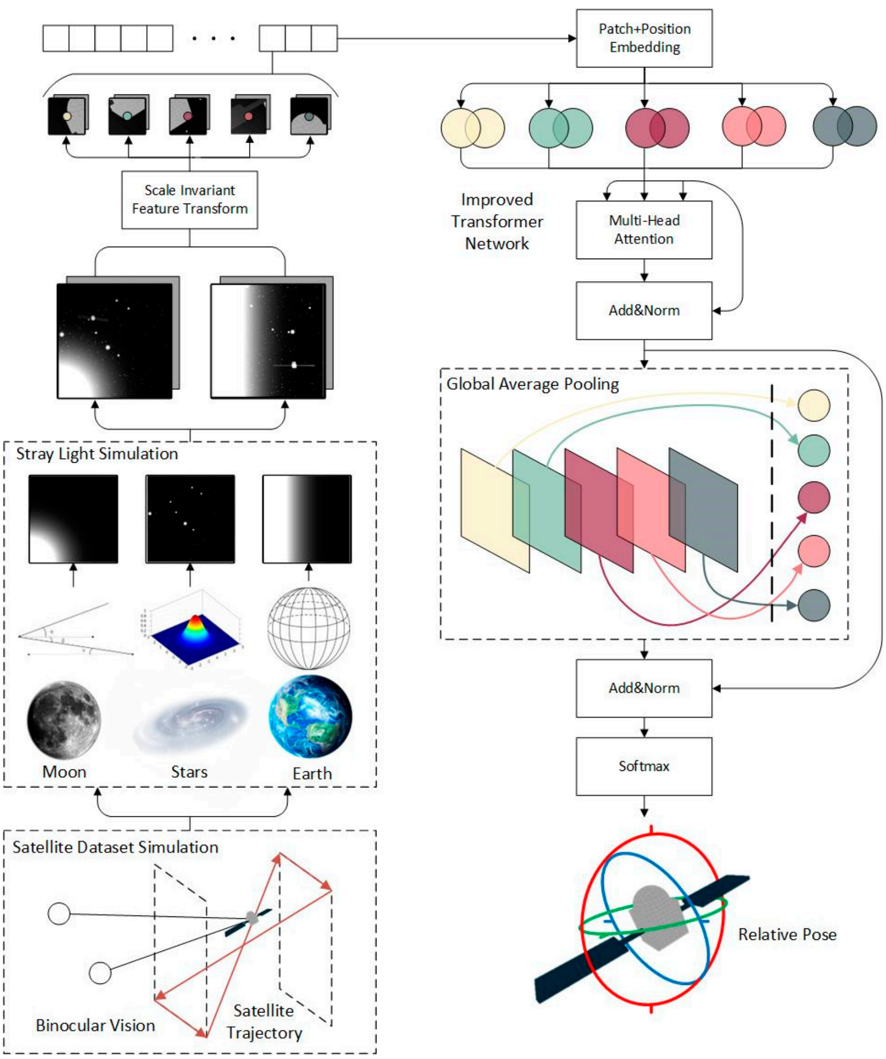


Figure 1. Overview of our Self-supervised Improved Transformer Network (STN). Our network model achieves the pose estimation capability of small sample data sets under the influence of stray light through self-supervised learning through changes in the network structure.

2. Related Work

The pose estimation methods can be mainly categorized into two categories: traditional methods and deep learning methods. For non-cooperative targets captured by binocular optical cameras, there have been many studies addressing the issue of stray light and small-sample training.

2.1. Traditional and Deep Learning Methods

To acquire target model information in noisy environments, some traditional research methods transform pose estimation problems into template matching problems, utilizing essential matrices for pose initialization. Pose calculation involves image filtering, edge detection, line extraction, and stereo matching. A three-dimensional model of non-cooperative micro and nanosatellites is reconstructed using a stereo vision system [13]. Subsequently, a method based on feature matching estimates the target's relative pose, followed by ground experiments to assess the algorithm's accuracy. Segal S et al. [14] employ the principles of binocular vision measurement and utilize an Extended Kalman Filter to track and observe target feature points, achieving pose measurement for non-cooperative spacecraft. Finally, a trial system for estimating non-cooperative target poses is constructed. However, non-cooperative images often vary in quality, and traditional methods suffer significant accuracy reduction with blurry or smoothly-edged targets, making them inadequate for complex non-cooperative target measurements. Despite proposing algorithms based on horizontal and vertical feature lines to derive fundamental matrices without using paired point information, the reliance on high-quality imagery contradicts the scarcity of suitable non-cooperative target image datasets. As a result, these methods face significant limitations in practical applications.

Deep learning methods do not depend on the target model, do not need manual feature design, and have better generalization abilities when the training data are adequate. Li K et al. [15] proposed a method that outperforms the heatmap and regression-based methods and improves the uncertainty prediction. Zhu Z et al. [16] suggested an algorithm that can effectively suppress interference points and enhance the accuracy of non-cooperative target pose estimation. Despond F T [17] used a novel convolutional model to estimate the relative x, y and attitude of the target spacecraft. Deep learning methods are more versatile and robust for different targets and scenarios than traditional methods and can be more effectively applied to non-cooperative pose estimation.

2.2. Small-Sample Training

To address the challenge of pose estimation for non-cooperative space targets with limited real samples, researchers have also turned to deep learning methods and conducted a series of studies. As the most mature image processing networks, neural network approaches have been widely employed in non-cooperative target pose estimation, forming the basis for numerous improved and optimized algorithms capable of addressing various scenarios. Pasqualetto Cassinis L et al. [18] present a fusion of convolutional neural network-based feature extraction and the CEPPnP (efficient Procrustes perspect-n-points) method, combined with Extended Kalman Filtering for non-cooperative target pose estimation. Hou X et al. [19] introduce a hybrid artificial neural network estimation algorithm based on dual quaternion vectors. Ma C et al. [20] propose a Neural Network-Enhanced Kalman Filter (NNEKF), innovatively improving filter performance using the virtual observation of inertial characteristics. Huan W et al. [21] employ existing object detection networks and keypoint regression networks to predict 2D keypoint coordinates, reconstructing a 3D model through multi-viewpoint triangulation and minimizing 3D coordinates with nonlinear least squares to predict position and orientation. Li Xiang et al. [22] designed a non-cooperative target pose estimation network based on the Google Inception Net model.

Applications of the proposed MEGNN-based method to PHM 2010 milling TCM dataset and experiments demonstrate it outperforms three DL-based methods (CNN, AlexNet, ResNet) under small samples [23]. Pan T et al. [24] proposed a generative adversarial network (GAN), which is considered a promising way to solve the problem of small samples. Ma et al. [25] proposed a face recognition method based on sparse representation of deep learning features. This method first extracts face features using deep CNN and then classifies the obtained face features by sparse representation. Experiments prove that this method has higher recognition accuracy, which can improve by 6–60% compared with traditional methods, can effectively cope with the interference caused by intra-class changes, such as lighting, pose, expression, and occlusion, and has a greater advantage when encountering small sample problems. Despite the application of deep learning methods to space target scenarios, their efficacy is still hampered by the scarcity of actual samples, often relying on simulation datasets for training, leaving room for improvement in accuracy and methodology.

2.3. Stray Light

During the process of collecting space signals using optical sensors, non-target light information is captured in the form of stray light, and such interference is challenging to completely suppress or eliminate. Correlation methods can only reduce the impact of stray light interference [26]. For complex space environments, many studies have also incorporated methods for handling unique spatial noise. Yang Ming et al. [27] address the issue of significant lighting and Earth background effects on non-cooperative spacecraft attitude measurement in space, proposing an end-to-end attitude estimation method based on convolutional neural networks with AlexNet and ResNet architectures. Compared to using regression methods alone for attitude estimation, this approach effectively reduces the average absolute error, standard deviation, and maximum error of attitude estimation. Synthetic images used for network training adequately consider factors such as noise and lighting in orbit. Additionally, Sharma S et al. [28] introduce the SPN (spacecraft pose network) model, which trains the network using grayscale images. The SPN model consists of three branches, with the first using a detector to detect the boundary boxes of the target in the input image and the other two branches using regions within the 2D boundary boxes to determine the relative pose. The improvement in accuracy methods also brings up another issue: the scarcity of samples in space target data. To address the problem of small samples in space target data, the dataset of the target is built using Unity3d2019 [29] software. To fully simulate the space lighting environment, the brightness of simulated sunlight in the environment is randomly set, starry background noise is randomly added, and data normalization is performed for data enhancement. Jiang Zhaoyang et al. [30] designed a dual-channel neural network based on VGG and DenseNet architectures to locate the pixel corresponding to feature points in the image and provide their corresponding pixel coordinates, proposing a neural network pruning method to achieve network lightweighting. Addressing the interference of space lighting and the issue of small samples, Sharma S et al. [31] present a monocular image-based pose estimation network. Phisannupawong T et al. and Chen B et al. [32,33] achieve 6-DOF pose estimation for non-cooperative spacecraft using pre-trained deep models. Despite Sonawani S et al. [34] being the first to create a dataset for non-cooperative targets using a semi-physical simulation platform, overall, there has not been extensive research into algorithms that simultaneously handle stray light and small sample sizes.

3. Materials and Methods

The data simulation in this paper is based on the Unity3d2019 [29] platform, creating a simulated dataset of decommissioned space satellites. By modifying their position and pose parameters, visible light simulated images from binocular vision are generated while recording the target's position and pose parameters to create the simulation dataset. Furthermore, three simulation methods—stellar magnitude analysis, latitude and longitude projection, and point spread function—are employed to construct target simulation images under the influence of stray light in space. For the intelligent detection and recognition of target features, SIFT feature points are utilized to stably describe local textures and shapes. Leveraging scale and rotation invariance, it emphasizes capturing fine details in small-scale features, effectively handling changes in target pose and partial occlusion situations, and swiftly detecting target features. This approach facilitates the rapid and accurate identification of target features for intelligent detection. The feature points on the input side, taken as patches, are projected into fixed-length vectors and fed into the Transformer. In subsequent encoders, based on the Transformer network as the foundational model, modifications are made to the feedforward neural network. The fully connected layer is replaced with a global average pooling layer from the CNN network, harnessing CNN's advantages to capture local features and enhance model generalization. Direct output is obtained from the encoder layer, yielding the position and pose information of the satellite in the photo.

3.1. Satellite Dataset Simulation

The software Unity3d2019 allows users to create and render realistic three-dimensional models and animations for various purposes, such as design visualization, games, and movies. It supports all standard 3D model formats, and it is convenient to capture virtual views with specific angles for a 3D model via it. Thus, it is suitable for us to build the synthetic dataset via it. The imagery in this paper features a dark background with an added point light source, creating a metallic reflection effect on the satellite's surface. The parameters for simulating the binocular camera are outlined in Table 1. During the modeling process, factors such as the target's size, shape, structure, and material are taken into consideration. For the simulation, the target object is modeled as a cube with sides measuring 4 m in length, topped with a hemisphere that stands 5.5 m above the cube's apex. The solar panels are set to a length of 18 m and a width of 6 m.

Table 1. Simulation of binocular camera parameters.

Parameter	Left Camera	Right Camera
focal length	20.8 mm	20.8 mm
full field of view	60°	60°
sensor size	36 mm × 24 mm	36 mm × 24 mm
pixel numbers	1024 pixel × 1024 pixel	1024 pixel × 1024 pixel
baseline	2000 mm	2000 mm
simulation unit length	1000 mm	1000 mm

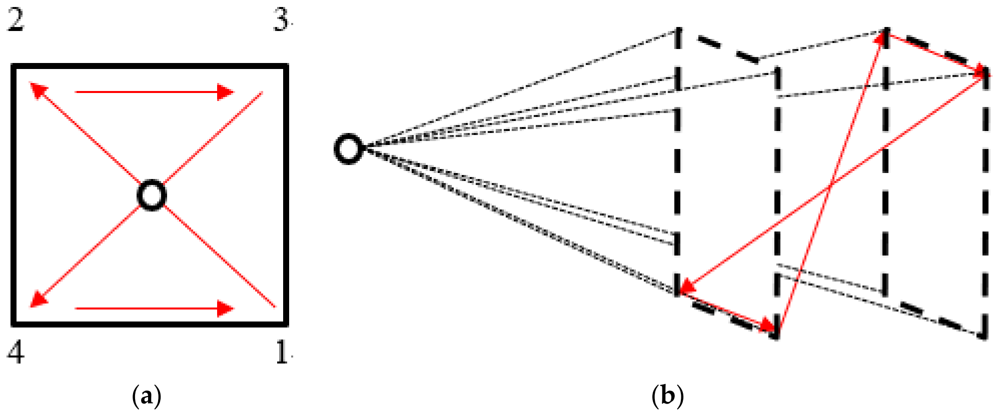
The satellite follows the described rotation and movement pattern, simultaneously moving and rotating while recording position and orientation information.

The model rotates along the roll, pitch, and yaw axes, in turn, to capture the target image from different orientations, which facilitates the recovery of the attitude information and the assessment of the object's spatial state in the later stage. Here is the rotation mode:

1. Maintain α and β angles unchanged, and rotate γ angle by 5 degrees each time, completing a full rotation;
2. Keep α angle constant, increase β angle by 5 degrees, and rotate γ angle by 5 degrees each time, completing a full rotation;
3. Maintain α angle constant, increase β angle by another 5 degrees, and rotate γ angle by 5 degrees each time, completing a full rotation;
4. Continue until β angle completes a full rotation, and increase α angle by 5 degrees;
5. Repeat steps 1, 2, 3, and 4 while also increasing β angle by 5 degrees each time, completing a full rotation.

The model's position changes correspond to the image's spatial coverage and depth variation, which helps to recover the target's location information in the subsequent process. It also provides three-dimensional information for the reconstruction of binocular vision. Here is movement pattern:

1 -> 2 -> 3 -> 4, with 1 moving from near to far, 2 moving horizontally only, 3 moving from far to near, and 4 moving horizontally only. The front view is depicted in Figure 2a, while the side view is illustrated in Figure 2b, with the camera represented by a circle and the red arrows indicating the satellite's trajectory.



(**a**) (**b**)

Figure 2. (**a**) Trajectory simulation front view; (**b**) side view of trajectory simulation.

In the end, this paper generates a total of 373,248 images using binocular output with dimensions of 72 × 72 × 72. Simultaneously, it outputs simulated satellite pose information comprising position coordinates (3) and rotation angles represented by quaternions (4).

3.2. Stray Light Simulation

Stray light interference poses a significant challenge in space-based optical image processing, leading to a degradation in image quality and accuracy. This paper employs an effective approach for image simulation, enabling the generation of synthetic images that incorporate stray light effects. These simulated images serve for the research and evaluation of algorithms in the presence of stray light interference.

3.2.1. Moonlight Simulation

Moonlight typically manifests in imaging as a bright region gradually spreading outward from a central point. Depending on the current CCD sensor's capacity to suppress moonlight, an influence on imaging is considered when the moonlight angle is less than 30 degrees. Let σ represent the angle between the observation platform's line of sight $r_{1,m}$ and the line $r_{2,m}$ connecting the observation platform to the center of the Moon, known as the moonlight angle, as illustrated in Figure 3. For an observation platform orbiting the Earth, the Moon can be treated as a point light source. It generates a Gaussian distribution function's diffusive effect with the corresponding image position as its center.

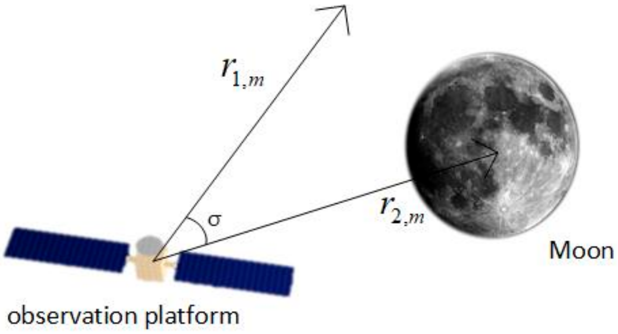


Figure 3. Moonlight angle diagram. When the moonlight angle between the Moon position assumed by the simulation and the observation position is less than 30°, the image produced by the existing CCD sensor will be affected by stray light.

3.2.2. Earth Atmosphere Radiation

The influence of atmospheric light on space-based visible light imaging primarily stems from the fact that the nadir point of the observation platform is illuminated by the Sun, causing the bright background, after undergoing atmospheric scattering, to enter the observation platform's field of view, leading to localized areas of brightness or the presence of interference patterns. Based on the current CCD sensor's capability to suppress atmospheric light, assuming the off-axis angle is less than 22 degrees, atmospheric light's impact on imaging is considered. Let θ represent the angle between the observation platform's line of sight $r_{1,e}$ and the tangent line $r_{2,e}$ to the atmospheric boundary from the observation platform, also known as the off-axis angle, as depicted in Figure 4.

Due to the Earth's proximity to the observation platform, it cannot be simplified as a point light source or parallel light; it must be treated as a surface light source. Hence, it is necessary to divide the Earth's surface into grids, as shown in Figure 5. In this case, the problem becomes equivalent to multiple point light sources generating Gaussian distribution functions' diffusive effects at corresponding positions in the image.

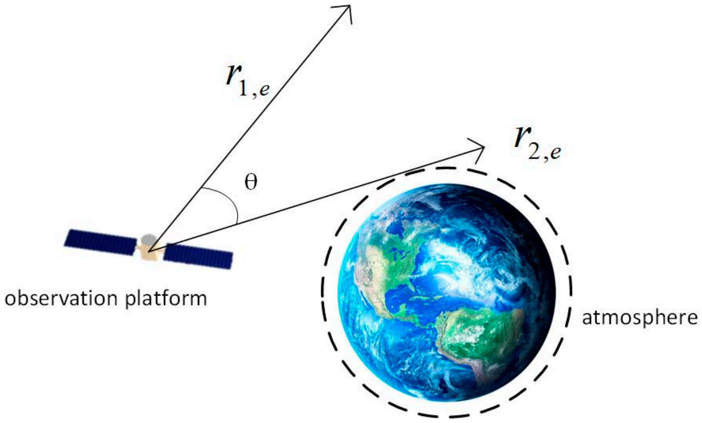


Figure 4. Off-axis angle diagram. When the off-axis angle between the earth position assumed by the simulation and the observation position is less than 22°, the image produced by the existing CCD sensor will be affected by stray light.

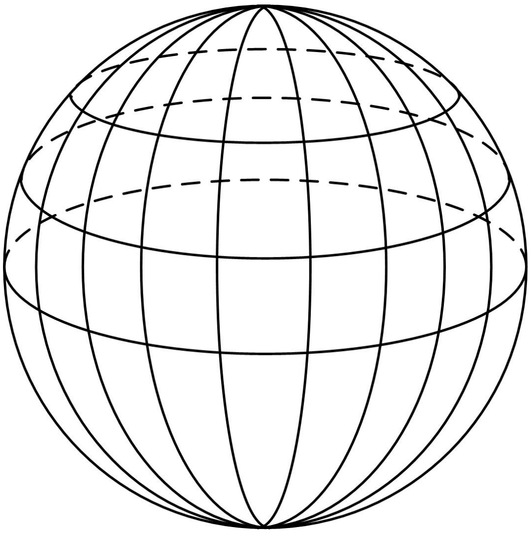


Figure 5. Grid division of the Earth's surface. The surface light source generated by the ground air light can be approximately equivalent to multiple-point light sources based on the grid.

3.2.3. Background Starlight Simulation

Background starlight points are determined using a random function to generate varying sizes of bright points, similarly employing the diffusive effect of Gaussian distribution functions. The Gaussian point spread function describes an optical system's resolution capability for point sources. Any point source will form an enlarged image point due to diffraction after passing through an optical system.

$$f(x,y) = \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right] \quad (1)$$

By measuring the system's point spread function [35], it becomes possible to more accurately extract image information, as shown in Equation (1). The simulation results of moonlight, atmospheric light, and starlight are illustrated in Figure 6.

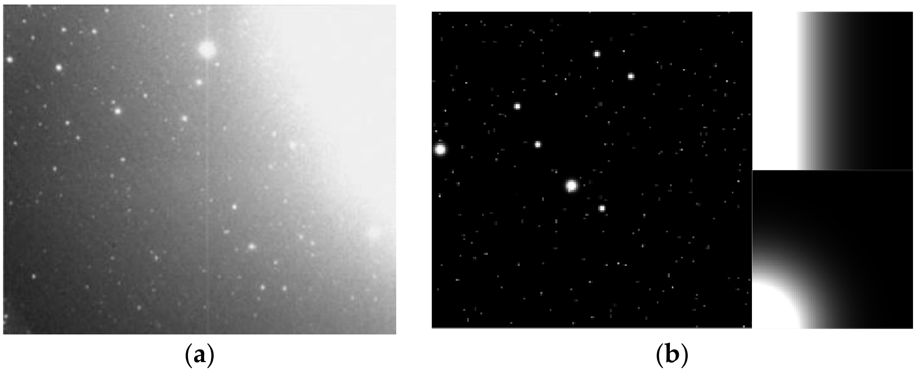


Figure 6. (**a**) Image on orbit; (**b**) simulation result.

3.3. Improved Transformer Network

Regarding the simulated images, to mitigate the effects of stray light, SIFT feature detection is applied. Gaussian images are generated at different scales, and their differences yield the DoG images. Leveraging the detection of local extrema, we can focus on fine detail features, thereby minimizing the impact of blurred noise on the images. Ultimately, the image's feature points are obtained, and the top 5 pairs of feature matching points with optimal accuracy are selected, constituting the data for a pair of photographs.

Transformer networks commonly employ absolute position encoding, assigning a unique positional code to each patch, thus lacking translation invariance. Local perception units utilize SIFT operators as a substitute for spatial convolution, introducing scale and rotation invariance of the operator into the Transformer module. The 5-point algorithm refers to how to obtain the essential matrix between the two images and then decompose the corresponding rotation matrix when the internal parameters of the camera are known and the 5 sets of image corresponding points between the two images are known. A method for translating vectors. The 5-point algorithm was proposed by David Nister [36] in 2004 and has become a widely used method for image-based three-dimensional reconstruction. Therefore, this article selects 5 pairs of points as input quantities and calculates pose parameters through rotation matrices and translation vectors. Five sets of feature detection results are utilized as input layers for the patches, as depicted in Figure 7. In this study, the image's feature point pairs are divided into patches, and each patch is projected into a fixed-length vector before being fed into the Transformer. The subsequent encoder operations mirror those of the original Transformer, effectively transforming the visual problem into a sequence problem. Similarly, the algorithm requires the incorporation of positional encoding without altering the vector dimensions.

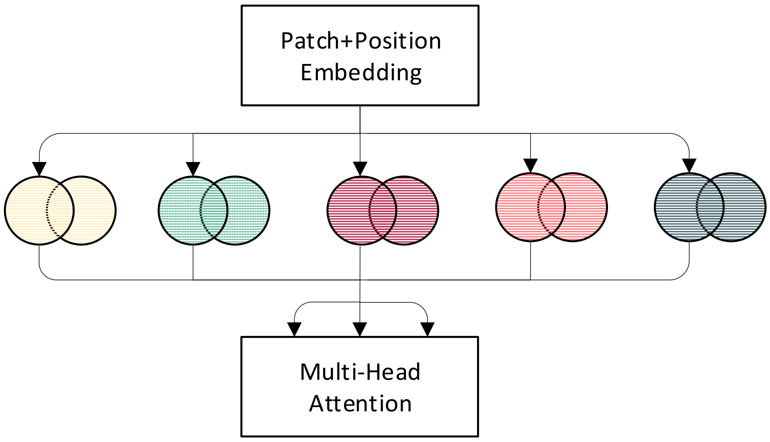


Figure 7. Model input. The input of the model is 5 sets of feature point pairs obtained by the feature extraction algorithm, which contains the feature information of the image.

While Transformer networks exhibit notable flexibility and transferability due to their modest inductive biases, they can encounter limitations in scenarios with limited data. On the other hand, convolutional neural networks (CNNs) possess strong inductive bias capabilities, including local sensitivity and translation invariance, which grant them high performance in low-data scenarios. In light of these considerations, this study focuses on improving the feedforward neural network within the Transformer architecture by replacing it with a global average pooling layer from the CNN network. Through this fusion, the strengths of both CNNs and Transformers are harnessed, overcoming their respective limitations and transcending the constraints of receptive fields.

The enhanced feedforward network, compared to the traditional feedforward network in the Transformer, employs global average pooling in lieu of fully connected layers, thus retaining spatial and semantic information captured by the preceding layers, as depicted in Figure 8. By stacking multiple decoder layers, the improved Transformer model efficiently incorporates contextual information during output sequence generation, resulting in significant performance gains across various sequence-to-sequence tasks. However, the accumulation of errors in the decoder layers is addressed by transitioning to linear layer output, mitigating the emphasis on sequence information and focusing on enhancing the model's accuracy.

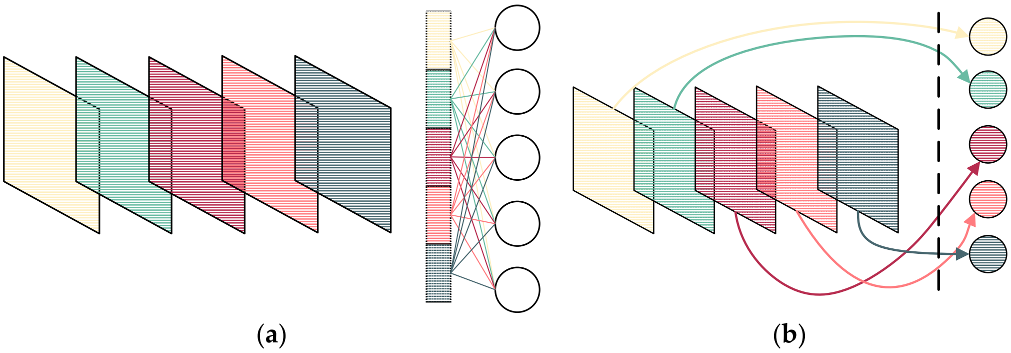


Figure 8. (**a**) Fully connected layer; (**b**) global average pooling.

4. Results

To acquire a spatial dataset under different poses and lighting conditions, this study employed Unity3D to generate images of decommissioned satellites with varying scales, poses, and operational scenarios, as illustrated in Figure 8a. This data set is a dark background specular reflection sample taken by a binocular camera and output according to the rotation mode. The three rotation angles are rotated by 5° to obtain a pair of binocular satellite images, forming a total of 72 × 72 × 72 (373,248) groups with a resolution of 1024 × 1024 motion satellite data set. A size of 1024 × 1024 is frequently used for sample images. For the grid size, there are three rotation angles with the same range from 0 to 360 degrees; each is resampled with 5 degrees as an interval; thus, it is 360/5 = 72. The interval of 5 degrees is chosen practically since too small an interval leads to high computation, and too large an interval results in low accuracy. Each sample image includes both position coordinates (three values) and rotation angles represented as quaternions (four values). To account for the effects of atmospheric and lunar illumination, Gaussian point spread functions were utilized to simulate scattered light interference, generating disturbance images, as depicted in Figure 8a,b.

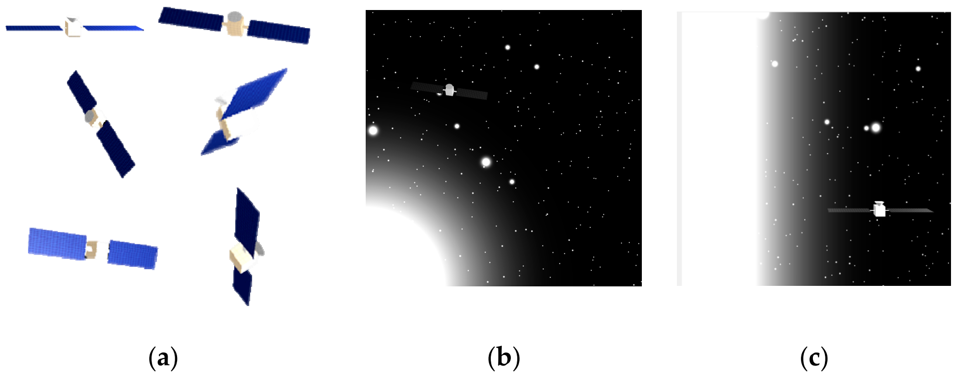
The improved Transformer model was employed to perform inference validation on the validation dataset, and the results, as indicated in Table 2, demonstrate favorable training and detection outcomes. The following table presents a comparison of position and pose errors obtained using different methods for localization and attitude estimation on the simulated dataset. This research aims to evaluate the performance of these methods in a simulated environment, thereby offering a reference for selecting appropriate techniques for real-world positioning and navigation systems. In this table, pose accuracy is presented in the form of errors, representing the average distance between the estimated and actual poses. It can be observed that the model presented in this paper maintains commendable accuracy in both position and pose errors.

Table 2. Pose estimation accuracy results.

Applied Model	Position Error Epos (%)	Attitude Error Eatt (°)
Model in this paper	0.958688	4.388
Transformer	1.163288	5.872
Fast R-CNN	1.080765	5.975
Yolov5	1.134575	5.504

Figure 9 presents a performance comparison of different pose estimation methods under varying sample sizes. This study aims to assess the influence of different sample quantities on pose estimation results and compare the stability and accuracy of different methods across different sample counts. The *x*-axis in the graph represents the sample quantity, while the *y*-axis represents a comprehensive performance metric for pose estimation, which combines both positions and pose errors as defined by the formula in Equation (2).

$$E = \sqrt{\frac{\sum_{i=1}^{N}\left[f(E_i) - E_{\mathrm{pos}}\right]^2 + \sum_{i=1}^{N}\left[f(E_i) - E_{att}\right]^2}{2N}} \tag{2}$$



(**a**) (**b**) (**c**)

Figure 9. (**a**) Target modeling. (**b**) Moonlight simulation. (**c**) Earth-atmosphere light simulation.

Each pose estimation method is distinctively marked with different colors or line styles in the chart to facilitate a visual comparison of their performance under varying sample sizes. With an increase in sample quantity, the following trends are observable: For most pose estimation methods, an increase in sample quantity generally leads to gradual improvement in pose estimation performance. Fewer samples may result in unstable estimates, whereas greater sample sizes enhance both the accuracy and stability of the estimates. Different pose estimation methods may exhibit varying performance across different sample counts. The proposed method in this paper demonstrates excellent performance with fewer samples, while other methods might require a larger sample size to achieve optimal performance in Figure 10.

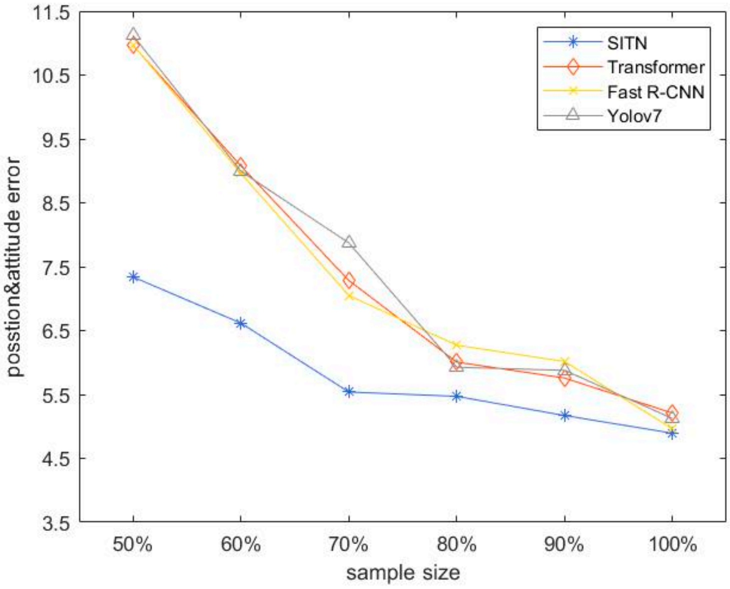


Figure 10. The accuracy change diagram of the model under the change of sample number.

5. Conclusions

This paper proposes a hybrid network model that combines Transformer and CNN for pose estimation of non-cooperative targets under binocular camera settings. The objective is to address the challenge of maintaining accuracy in the presence of sparse image samples due to scattered light effects. The model capitalizes on the strengths of both network architectures, capturing local and global information to enhance the network's representational capacity. Through experimentation, it is demonstrated that the network exhibits high-precision detection capabilities in comparison to various alternatives. Furthermore, the network maintains its computational efficiency even with changes in data volume, achieving the desired enhancements proposed in this study and confirming the effectiveness of the algorithm.

6. Future Work

This paper proposes a novel pose estimation algorithm for non-cooperative targets in stray light interference environments. It addresses the challenge of identifying non-cooperative targets without auxiliary markers, which is difficult for traditional algorithms. The algorithm suppresses noise through network input and maintains a certain accuracy of pose estimation even with reduced training data. The main contribution and significance of this paper is that it provides a new idea and method for visual pose estimation of non-cooperative targets, as well as an effective solution for the application of deep learning in small sample problems. This research has important implications and inspirations for the fields of space exploration, satellite rendezvous and docking, space debris cleanup and other fields.

However, this research also has some limitations and shortcomings that need to be further improved and extended in future work. One of them is that the data set used in this paper was obtained by simulation, which may lack generalization and fail to cope with more complex real environments. Therefore, future work needs to collect and construct more realistic data to verify and improve the robustness and adaptability of the algorithm. Another one is that this research may not be real-time enough to meet the real-time requirements of non-cooperative target on-orbit operations because it uses deep learning and other methods. Therefore, future work needs to optimize and accelerate the running speed and efficiency of the algorithm to adapt to higher real-time requirements.

In addition, this research also has some possibilities and potential for extension and expansion. For example, future work can explore and verify different network structures and parameters to improve the performance and accuracy of the algorithm. Alternatively, future work can use or develop more self-supervised learning techniques to overcome the difficulties of small sample problems while also increasing the interpretability and credibility of the algorithm. Alternatively, future work can compare and integrate the algorithm with other visual pose estimation methods to achieve better pose estimation effects and a wider range of application scenarios.

Author Contributions: Conceptualization, X.P. and B.W.; methodology, X.L.; software, J.L.; validation, X.P. and X.L.; formal analysis, B.W., Q.S. (Qinghong Sheng) and Z.Y.; investigation, Q.S. (Quan Sun); resources, K.Y.; data curation, J.W.; writing—original draft preparation, X.P., X.L., J.L., B.W., Q.S. (Qinghong Sheng) and Z.Y.; writing—review and editing, Q.S. (Quan Sun), K.Y. and J.W.; visualization, X.P.; supervision, X.L.; project administration, B.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Laboratory of Science and Technology on Space Microwave under Grant HTKJ2022KL504017 and HTKJ2021KL504013; the National Natural Science Foundation of China under Grant 42271448.

Data Availability Statement: Data sharing not applicable. No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bao, W. Research status and development trend of aerospace vehicle control technology. *Acta Autom. Sin.* **2013**, *39*, 697–702. [CrossRef]
2. Liang, B.; Du, X.; Li, C.; Xu, W. Advances in space robot on-orbit servicing for non-cooperative spacecraft. *Jiqiren (Robot)* **2012**, *34*, 242–256. [CrossRef]
3. Li, R.; Wang, S.; Long, Z.; Gu, D.U. Monocular visual odometry through unsupervised deep learning. *arXiv* **2017**, arXiv:1709.06841.
4. Hao, G.; Du, X. Research status of optical measurement of space non-cooperative target pose. *Prog. Laser Optoelectron.* **2013**, *50*, 246–254.
5. Yu, L.; Feng, C.; Ying, W. Spacecraft relative pose measurement technology based on lidar. *Infrared Laser Eng.* **2016**, *45*, 0817003.
6. Feng, C.; Wu, H.; Chen, B. Pose parameter estimation between spacecraft based on multi-sensor fusion. *Infrared Laser Eng.* **2015**, *44*, 1616–1622.
7. Kendall, A.; Grimes, M.; Cipolla, R. Convolutional networks for real-time 6-DOF camera relocalization. *arXiv* **2015**, arXiv:1505.07427.
8. Zhu, X.; Jiang, Q. Research on UAV image target detection based on CNN and Transformer. *J. Wuhan Univ. Technol.* **2022**, *44*, 323–331.
9. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
10. Fang, Y.; Liao, B.; Wang, X.; Fang, J.; Qi, J.; Wu, R.; Niu, J.; Liu, W. You only look at one sequence: Rethinking transformer in vision through object detection. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26183–26197.
11. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
12. Sun, Z.; Cao, S.; Yang, Y.; Kitani, K.M. Rethinking transformer-based set prediction for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3611–3620.
13. Terui, F.; Kamimura, H.; Nishida, S. Motion estimation to a failed satellite on orbit using stereo vision and 3D model matching. In Proceedings of the 2006 9th International Conference on Control, Automation, Robotics and Vision, Singapore, 5–8 December 2006; pp. 1–8.
14. Segal, S.; Carmi, A.; Gurfil, P. Vision-based relative state estimation of non-cooperative spacecraft under modeling uncertainty. In Proceedings of the 2011 Aerospace Conference, Big Sky, MT, USA, 5–12 March 2011; pp. 1–8.
15. Li, K.; Zhang, H.; Hu, C. Learning-Based Pose Estimation of Non-Cooperative Spacecrafts with Uncertainty Prediction. *Aerospace* **2022**, *9*, 592. [CrossRef]
16. Zhu, Z.; Xiang, W.; Huo, J.; Yang, M.; Zhang, G.; Wei, L. Non-cooperative target pose estimation based on improved iterative closest point algorithm. *J. Syst. Eng. Electron.* **2022**, *33*, 1–10. [CrossRef]
17. Despond, F.T. Non-Cooperative Spacecraft Pose Estimation Using Convolutional Neural Networks. Ph.D. Thesis, Carleton University, Ottawa, ON, Canada, 2022.
18. Pasqualetto Cassinis, L.; Fonod, R.; Gill, E.; Ahrns, I.; Gil Fernandez, J. Cnn-based pose estimation system for close-proximity operations around uncooperative spacecraft. In Proceedings of the AIAA Scitech 2020 Forum, Orlando, FL, USA, 6–10 January 2020; p. 1457.

19. Hou, X.; Yuan, J.; Ma, C.; Sun, C. Parameter estimations of uncooperative space targets using novel mixed artificial neural network. *Neurocomputing* **2019**, *339*, 232–244. [CrossRef]
20. Ma, C.; Zheng, Z.; Chen, J.; Yuan, J. Robust attitude estimation of rotating space debris based on virtual observations of neural network. *Int. J. Adapt. Control Signal Process.* **2022**, *36*, 300–314. [CrossRef]
21. Huan, W.; Liu, M.; Hu, Q. Pose estimation for non-cooperative spacecraft based on deep learning. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 3339–3343.
22. Li, X. Design of Spatial Non-Cooperative Target Pose Estimation Algorithm Based on Deep Learning. Master's Thesis, Harbin University of Technology, Harbin, China, 2019.
23. Zhou, Y.; Zhi, G.; Chen, W.; Qian, Q.; He, D.; Sun, B.; Sun, W. A new tool wear condition monitoring method based on deep learning under small samples. *Measurement* **2022**, *189*, 110622. [CrossRef]
24. Pan, T.; Chen, J.; Zhang, T.; Liu, S.; He, S.; Lv, H. Generative adversarial network in mechanical fault diagnosis under small sample: A systematic review on applications and future perspectives. *ISA Trans.* **2022**, *128*, 1–10. [CrossRef] [PubMed]
25. Li, Y. Research and application of deep learning in image recognition. In Proceedings of the 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), Shenyang, China, 21–23 January 2022; pp. 994–999.
26. Xu, Z. Research on Stray Light Suppression and Processing Technology of Space-Based Space Target Detection System. Ph.D. Thesis, University of Chinese Academy of Sciences (Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences), Beijing, China, 2021. [CrossRef]
27. Yang, M. Research on Multi-Mode Intelligent Reconstruction Algorithm for Spatial Non-Cooperative Targets Based on Deep Learning. Master's Thesis, Harbin University of Technology, Harbin, China, 2019.
28. Sharma, S.; D'Amico, S. Pose estimation for non-cooperative rendezvous using neural networks. *arXiv* **2019**, arXiv:1906.09868.
29. Unity Technologies. *Unity*; [Computer Software]; Unity Technologies: San Francisco, CA, USA, 2019.
30. Jiang, Z. Non-Cooperative Spacecraft Monocular Vision Pose Measurement Method Based on Deep Learning. Ph.D. Thesis, Harbin Institute of Technology, Harbin, China, 2021. [CrossRef]
31. Sharma, S.; Beierle, C.; D'Amico, S. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. In Proceedings of the 2018 IEEE Aerospace Conference, Big Sky, MT, USA, 3–10 March 2018; pp. 1–12.
32. Phisannupawong, T.; Kamsing, P.; Torteeka, P.; Channumsin, S.; Sawangwit, U.; Hematulin, W.; Jarawan, T.; Somjit, T.; Yooyen, S.; Delahaye, D.; et al. Vision-based spacecraft pose estimation via a deep convolutional neural network for noncooperative docking operations. *Aerospace* **2020**, *7*, 126. [CrossRef]
33. Chen, B.; Cao, J.; Parra, A.; Chin, T.J. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 2816–2824.
34. Sonawani, S.; Alimo, R.; Detry, R.; Jeong, D.; Hess, A.; Amor, H.B. Assistive relative pose estimation for on-orbit assembly using convolutional neural networks. *arXiv* **2020**, arXiv:2001.10673.
35. Wang, Y.; Niu, Z.; Wang, D.; Huang, J.; Li, P.; Sun, Q. Simulation algorithm for space-based optical observation images considering influence of stray light. *Laser Optoelectron. Prog.* **2022**, *59*, 0229001.
36. Nister, D. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 756–770. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.