DMANet: Dense Multi-scale Attention Network for Space Non-cooperative Object Pose Estimation

ZHANG Zhao, HU Yuhui, ZHOU Dong, WU Ligang, YAO Weiran, LI Peng*

School of Astronautics, Harbin Institute of Technology, Harbin 150001, P.R. China

(Received 4 December 2023; revised 30 December 2023; accepted 28 January 2024)

Abstract: Accurate pose estimation of space non-cooperative targets with a monocular camera is crucial to space debris removal, autonomous rendezvous, and other on-orbit services. However, monocular pose estimation methods lack depth information, resulting in scale uncertainty issue that significantly reduces their accuracy and real-time performance. We first propose a multi-scale attention block (MAB) to extract complex high-dimensional semantic features from the input image. Second, based on the MAB module, we propose a dense multi-scale attention network (DMANet) for estimating the 6-degree-of-freedom (DoF) pose of space non-cooperative targets, which consists of planar position estimation, depth position estimation, and attitude estimation branches. By introducing an Euler angle-based soft classification method, we formulate the pose regression problem as a classical classification problem. Besides, we design a space non-cooperative object model and construct a pose estimation dataset by using Coppeliasim. Finally, we thoroughly evaluate the proposed method on the SPEED+, URSO datasets and our dataset, compared to other state-of-the-art methods. Experiment results demonstrate that the DMANet achieves excellent pose estimation accuracy.

Key words: 6-degree-of-freedom (DoF) pose estimation; space non-cooperative object; multi-scale attention; deep learning; neural network

CLC number: TP391 **Document code:** A **Article ID:** 1005-1120(2024)01-0122-13

0 Introduction

Spacecraft pose estimation is used to obtain the position and orientation of space targets and provide security for on-orbit services, including maintenance, autonomous rendezvous and debris removal[1-5]. With the development of space technology, a large number of satellites have been launched. The total number of satellites in orbit has exceeded 6 700 as of January 2023[6]. Over time, spacecraft can encounter issues such as equipment degradation, fuel depletion, and accidental impacts, which lead the overall system to fail[7]. The repair of malfunctioning spacecraft requires the determination of their pose. In addition, those abandoned spacecraft occupy space orbits and generate space debris, which heavily threaten the safety of spacecraft in operation. Pose estimation can confirm the real-time position of space objects, and estimate their size and motion status, which is important to ensure the safety of spacecraft[8-11].

There are a variety of sensors used for pose estimation, including active vision sensors such as time of flight (ToF) cameras[12] and Lidars[13-14], as well as passive vision sensors such as stereo cameras[15-16] and monocular cameras[17]. Active vision sensors typically have high power consumption and are difficult to maintain, limiting their large-scale application in space. Stereo camera can precisely measure the target position, but it heavily relied on object texture and is vulnerable to environmental perturbations. Therefore, monocular cameras are wide-

*Corresponding author, E-mail address: ligangwu@hit.edu.cn.

How to cite this article: ZHANG Zhao, HU Yuhui, ZHOU Dong, et al. DMANet: Dense multi-scale attention network for space non-cooperative object pose estimation[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2024, 41(1): 122-134.

http://dx.doi.org/10.16356/j.1005-1120.2024.01.010

ly used to accomplish pose estimation tasks in space, because of its small size, low power consumption, and simple hardware requirement. However, monocular cameras cannot directly obtain 3D information without a depth estimation algorithm, and their accuracy is relatively poorer than those of the previous methods.

Traditional monocular solutions for pose estimation can be categorized into feature-based and template-based methods. Feature-based methods extract local features from the image and compare them with the features obtained on the 3D model to establish 2D-3D matches[18-20]. These algorithms are usually divided into two stages: Extracting local features and comparing them with 3D keypoints, then involving 2D-3D correspondences to solve geometric problems, e.g., obtaining a 6D position using the perspective-n-point (PnP) algorithm[21]. These algorithms can be run in real-time, but they require objects to have clear texture details, otherwise, feature matching is likely to fail.

Template-based methods first establish a template database offline from a 3D model of the object, then match the image with the database to obtain the best position estimate. These algorithms can obtain good results for texture-less objects[21-23], and the accuracy improves with the completeness of the template database[24]. However, they are sensitive to illumination and occlusion[25-26], and real-time performance is inversely proportional to the number of template elements[27].

Learning-based methods use the powerful representation capabilities of neural networks to estimate the position of a known instance or a class of objects. Depending on the algorithmic framework, these algorithms can be categorized as two stages[28] or end-to-end[26]. The former uses neural networks to project the 3D keypoints of the target onto the corresponding 2D image, then chooses a traditional algorithm to get the target pose. The latter uses the multi-branch network to directly regress the target pose. In recent years, a large number of accuracy and robust pose estimation algorithms have emerged from learning-based methods[29-30] which can accurately estimate the pose even under the conditions of occluded objects or cluttered backgrounds. Although these methods are effective, they need a lengthy and resource-intensive training process, necessitate extensive datasets with precise labeling, and their ability to generalize across different scenarios remains a subject for further evaluation.

Xiang et al.[31] designed a fully-convolutional neural networks (CNN), called PoseCNN, to acquire the target pose. They used one CNN with three branches to accomplish semantic labels, 3D translation and 3D rotation estimation seperately. The method provides a backbone network and a multi-branch structure; however, its accuracy needs to be further corrected[32]. Kehl et al.[33] proposed the single shot multi-box detector (SSD)-6D network, which extends the SSD detection framework[34] to accomplish 3D detection and 3D rotation estimation. The method decomposed the 3D rotational space into discrete viewpoints and in-plane rotations, used a neural network to obtain the target 2D bounding box, discretized the spatial rotations and solved the classification problem to obtain the 6D pose. Since the algorithm discretizes the 3D space and uses classification to solve the 6D pose, the accuracy needs to be improved.

Unlike the pose estimation of terrestrial objects, non-cooperative objects in space face more challenges, such as unclear texture and drastic illumination changes. Sharma et al.[35] first used CNNs for spacecraft pose estimation based on hard viewpoint classification, again using spatial discretization and transforming the pose estimation into a classification problem. They then proposed a spacecraft pose network (SPN) to accomplish the pose estimation[36]. This method uses a CNN with three branches to accomplish 2D bounding box detection, discrete coarse pose classification, and finer estimate regression. This paper also proposed a spacecraft pose estimation dataset (SPEED), including a total of 15 300 frames of AR generated and real satellites, for training and testing of the position estimation algorithms. Huang et al.[37] proposed a non-model-based monocular pose estimation network. They designed a CNN with three sub-networks to predict the relative pose, relative position, and de-

tect objects, respectively. Although this method can achieve end-to-end pose estimation, it neither considers multi-scale adaptation of space objects, nor acquires independent object depths, which are always the most important information of space missions.

In addition, Proença et al.[38] proposed a spacecraft pose estimation framework based on orientation soft classification. The method used ResNet as the backbone to obtain the relative position and relative attitude of the spacecraft through regression and probabilistic orientation soft classification, respectively. They also proposed a simulator URSO based on Unreal Engine 4 to generate training and evaluation data. Huang et al.[39] proposed an end-to-end attitude estimation network. The method used a multi-branch keypoints regression subnetwork to obtain the 2D keypoints locations, and a pose estimation subnetwork to estimate the target spacecraft pose based on the 2D keypoints and the corresponding 3D keypoints.

However, these methods use the same network or branch to estimate the spatial target positions and do not design individual branches and loss functions for the most important information, depth, in the spatial task. These methods therefore obtain large estimation errors and are unable to meet actual task requirements. In this paper, we propose the dense multi-scale attention network (DMANet) network for pose estimation of space non-cooperative targets, featuring the multi-scale attention mechanism and complex semantic representation. DMANet segments the 3D position estimation into two components: Planar positioning, encompassing the X and Y axes, and depth positioning, aligned along the Z-axis. Significantly, DMANet introduces a tailored loss function, specifically designed for the depth information. This novel approach proves to be more efficacious for spatially spatial tasks, offering a substantial improvement over conventional methodology.

The contributions of this paper are as follows:

(1) We design a feature extraction module, named multi-scale attention block (MAB), which combines multi-scale convolutional networks with channel attention mechanisms. Based on the MABs, we further propose the dense multi-scale attention (DMA) backbone, which can extract multi-scale information from the image of space non-cooperative objects.

(2) We propose the DMANet for non-cooperative object pose estimation, which consists of three branches that can extract information about the position, depth, and orientation of the object. For orientation estimation, we encode the labels of images by soft classification, dividing the geometric space into different subintervals, and transform the direct regression problem into a classification problem.

(3) To prove the effectiveness of DMANet, we construct a space non-cooperative object pose estimation dataset based on Coppeliasim. Then we perform pose estimation experiments on SPEED+ dataset[40], URSO datasets[38] and our dataset. Experimental results show that the proposed method can achieve high precision 6-degree-of-freedom (DOF) pose estimation for space non-cooperative objects.

1 Dense Multi-scale Attention Network

In this section, we introduce the MAB module and DMANet network in detail. Then, the soft classification method and corresponding loss function are clearly described, which formulate the pose regression problem as a classical classification problem.

1.1 Multi-scale attention block

Benefiting from the influence of the attention mechanism, we propose the MAB module, as shown in Fig.1. MAB is composed of multi-scale convolution and channel attention module, and its input is a preprocessed image feature map.

In the part of multi-scale convolution, we use kernels with the sizes of 1×1, 3×3, and 7×7 for convolution processing of input feature graphs, and then combine feature graphs of different scales in channel dimension and input them into channel attention module.

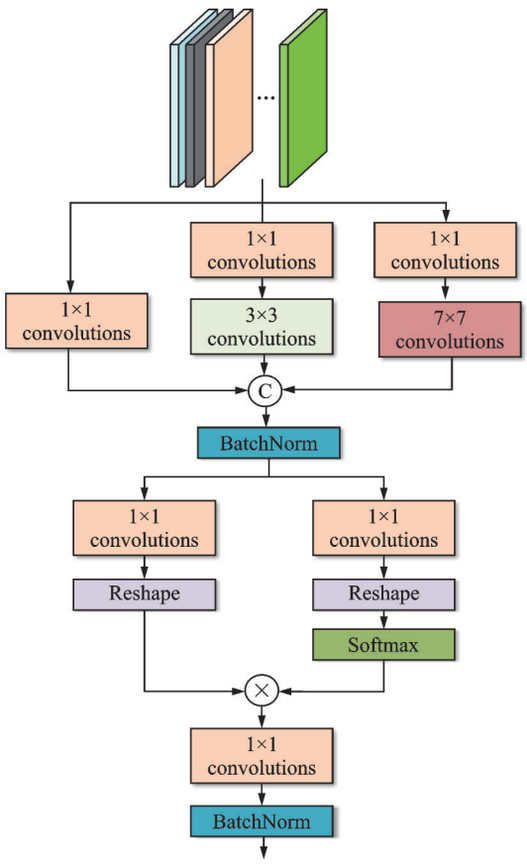


Fig.1 Multi-scale attention block

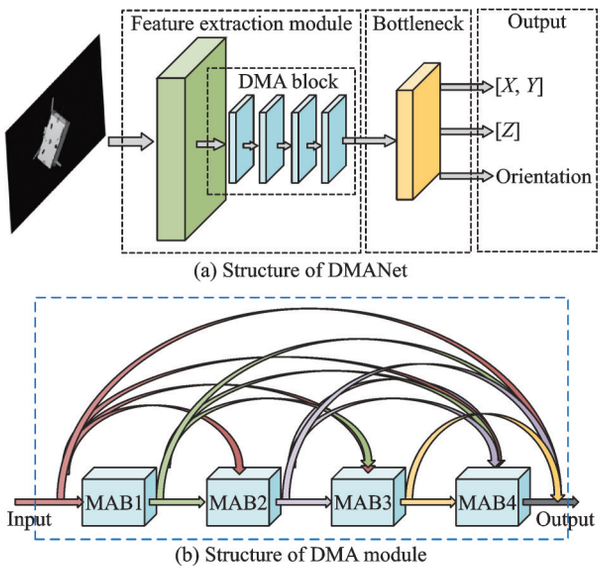


Fig.2 DMANet framework

Table 1 Architecture of DMA module

Layer	Channel (input)	Channel (output)	Activation
MAB1	64	64	ReLU
MAB2	128	64	ReLU
MAB3	192	64	ReLU
MAB4	256	64	ReLU

The channel attention mechanism is employed within deep learning to provide weights to different features in input data. This mechanism uses the inherent correlation between different channels of input to calculate the weightage or "attention" they should be given. We use the channel attention mechanism to extract multi-scale information more efficiently and subsequent experiments prove that the MAB module has a strong feature extraction capability.

1.2 Backbone network

Based on the MAB, we propose the DMA backbone network, as shown in Fig.2. Four MABs are used in the DMA backbone network, and the input and output channels of each module are shown in Table 1. The model preserves and uses the feed-forward feature maps of all preceding layers, and passes its own feature maps to all subsequent layers.

The dense structure, benefiting from its inherent architecture, allows each layer to access the "knowledge" from its preceding layers. The wonderful advantage is that it does not only preserve and utilize these feed-forward feature maps, but also transfer its own feature maps to all successive layers. The benefits of such efficient connectivity are multi-fold. It mitigates the vanishing-gradient issue often faced in deep architectures, amplifies feature propagation, and significantly reduces the number of parameters, thus enhancing the model's efficiency.

In the DMA backbone network, the main purpose of using four MABs is to build a hierarchical network, ensuring that the features of each convolutional layer can be learned by subsequent parts of the network. When we choose the number of input and output channels, we mainly consider that the number of network layers and the number of characteristic channels is directly related to the computational complexity of network operations. In addition, the number of input channels of each MAB is determined by the dense structure, because the input channels of each MAB contain the output of the previous module. Reasonable allocation of the number of channels can avoid unnecessary calculation

burdens while retaining the necessary feature extraction capability.

1.3 Pose estimation overall network

We design a pose estimation network for space non-cooperative objects based on DMA backbone network called DMANet, which is shown in Fig.2(a).

Let $I_{\text{input}} \in \mathbf{R}^{H \times W \times C}$ represent the input image. First, we preprocess the input image to obtain the multi-channel feature map, which is represented by $I_{\text{pre}} \in \mathbf{R}^{H \times W \times P}$ ($P = 64$). Subsequently, the output after input $I_{\text{pre}} \in \mathbf{R}^{H \times W \times P}$ to the DMA feature extraction module is $I_{\text{DMA}} \in \mathbf{R}^{H \times W \times 5P}$.

Unlike other pose estimation networks, DMANet is a three-branch output structure composed of three fully connected layers. In our opinion, the depth information of targets captured by monocular cameras will affect the accuracy of the other two dimensions of targets. Therefore, we use one of the three branches to regression the distance of the target on the Z-axis, and design a separate loss function for it later. The remaining two branches will perform orientation estimate and position estimate in the other two dimensions.

1.4 Probabilistic orientation soft classification

Different from the method to directly regress the quaternion of the target, we hope to encode the orientation labels of the data set through soft classification, let the orientation estimation branch of DMANet output a probability mass function (PMF), and then compare it with the encoded labels. Thus, the quaternion regression process is transformed into a classification problem.

First, we convert the quaternion from the dataset $Q_{\text{gt}}^i = (q_1^i, q_2^i, q_3^i, q_4^i), i = 1, 2, \cdots, N$ to the Euler angle $E_{\text{gt}}^i = (\varphi_{\text{gt}}^i, \psi_{\text{gt}}^i, \theta_{\text{gt}}^i), i = 1, 2, \cdots, N$, where N represents the number of labels. To ensure the uniform distribution of the bins of orientations, we divide each attitude angle into M intervals on average, and the dimension of the image label is converted from $(1 \times 1 \times 1)$ to $(M \times M \times M)$.

Then we design a kernel function $K(a, b)$ to encode the label information with the generated orientation bins, as shown in Eq.(1), where σ is used to adjust the size of the output. To facilitate the calculation, the result of Eq.(1) needs to be normalized, as shown in Eq.(2).

$$K(a,b) = \mathrm{e}^{-\left(\frac{|\cos(a-b)|}{M\sigma^2}\right)} \tag{1}$$

$$N(a_j,b) = \frac{K(a_j,b)}{\sum_{k=1}^{M} K(a_k,b)} \tag{2}$$

The labels of the three attitude angles in the image are encoded, as shown in Eq.(3). The probability mass function of each label can be obtained. By comparing with the output of the orientation estimation branch of DMANet, the network can be effectively optimized.

$$N(e_i,e_{\text{gt}}) = N(\varphi_i,\varphi_{\text{gt}}) + N(\psi_i,\psi_{\text{gt}}) + N(\theta_i,\theta_{\text{gt}}) \tag{3}$$

After training the network, we decode the probability mass function of the network output by

$$\hat{e} = \arg\max_{e_o} \sum_{k}^{M \times M \times M} \omega_k \cos(e_o - e_p) \tag{4}$$

where e_p represents the element of the Euler angle generated when the label is encoded, that is, we could get the result of orientation estimation and $\omega_i = \{\omega_1, \omega_2, \cdots, \omega_{M \times M \times M}\}$ represents the output result of the orientation estimation branch of DMANet.

DMANet consists of three branches, and we design the loss function as

$$L_{\text{total}} = L_{xy} + L_z + \lambda_1 L_{\text{ori}} + \lambda_2 L_{\text{distr}} \tag{5}$$

where λ_1 and λ_2 are hyperparameters, and used to keep each part of the loss function the same order of magnitude. Each part of the loss function is described in detail below.

First of all, since the comparison between the probability mass function of the network output and the encoded label should consider the similar content, as well as the difference in the distribution of the two probabilities, we adopt the combination of L_{ori} and L_{distr} for orientation estimation

$$L_{\text{distr}} = -\sum_{i=1}^{N} P_{\text{p}}^i \ln(P_{\text{gt}}^i) \tag{6}$$

$$L_{\text{ori}} = \sum_{i=1}^{N} \|P_{\text{p}}^i - P_{\text{gt}}^i\|_2 \tag{7}$$

where P_{p}^i and P_{gt}^i denote the predicted probability mass function and the corresponding ground-truth,

respectively.

In view of the effect of depth information on position estimation, it is unreasonable to use absolute error to measure the position error and size of non-cooperative objects. Therefore, we design a loss function, as shown in Eq.(8), which is linearly correlated with the target depth information.

$$L_{xy}=\sum_{i}^{N}\frac{\|(x,y)_{\text{pre}}^{i}-(x,y)_{\text{gt}}^{i}\|}{\|z_{\text{gt}}\|} \tag{8}$$

A branch of DMANet contains a regression of the Z-axis information, and we measure the branch output by the 2-norm as

$$L_{z}=\sum_{i}^{N}\|z_{\text{pre}}^{i}-z_{\text{gt}}^{i}\|_{2} \tag{9}$$

2 Experimental Results and Analysis

To demonstrate the effectiveness of our proposed approach, we conducted experiments on the SPEED+ dataset[40], URSO dataset[38] and our dataset. In addition, we analyzed the results of our experiments and also conducted ablation study to show the effectiveness of our method.

2.1 Datasets

The SPEED+ dataset was proposed by the ESA in 2021. To mimic the visual features and lighting variations of the spaceborne images, besides 60 000 synthetic images, SPEED+ also used the Testbed for Rendezvous and Optical Navigation (TRON) facility to take 9 531 photos of the spacecraft model, which showed the spacecraft in various attitudes. URSO dataset is constructed based on simulation platform Unreal Engine 4. The dataset includes Soyuz and Dragon spacecraft models with geometric shapes imported from the 3D model library. The synthetic scenes consider both the light intensity of the space environment and the influence of the Earth's complex background. The 5 000 viewing points in low Earth orbit were used to take random images of the spacecraft at different times and camera's orientation.

In addition, we have constructed a space pose estimation dataset by using virtual physics engine, Coppeliasim, which contains 1 000 images of a space non-cooperative object. These images are split in frame-wise by portion of 4∶1, which means that the train set contains 800 images and test set includes the rest. The simulation environment is shown in Fig.3.

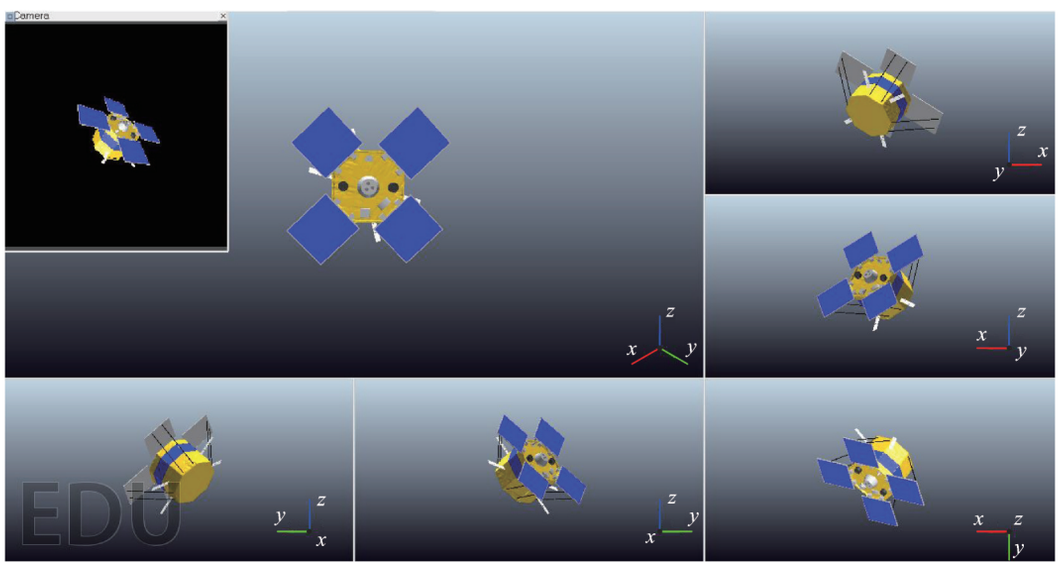


Fig.3 Schematic diagram of Coppeliasim simulation environment

The non-cooperative object, measuring 2.9 m×2.9 m×1.2 m, is a spacecraft model programmed via Python code to continually alter its position and orientation within the view of a vision sensor. We obtain the object image sequence by using the vision sensor in Coppeliasim, the perspective angle

(α_x, α_y) and resolution (W, H) can be obtained in Coppeliasim. In this work, we set $\alpha_x = \alpha_y = 75°$ and $W = H = 224$. The coordinate system is defined as shown in the lower right corner of each subfigure in Fig.3, where red represents the X-axis, green the Y-axis, and blue the Z-axis.

2.2 Implementation and training details

DMANet was trained on a Tesla P100 GPU and 2.60 GHz Intel(R) Xeon(R) Gold 6132 based on the deep learning framework Pytorch. Unless otherwise specified, we set $M = 6$ and $\lambda_1 = 10, \lambda_2 = 5$ during training. In addition, to prevent the network from producing local optimal solutions, the learning rate would gradually decline with training steps increasing. We have trained 50 epochs in total, and set the initial learning rate = 0.01, batch size=32, and the image of the dataset was resized to (224,224).

2.3 Evaluation metric

Jensen-Shannon divergence is a measure of how similar two probability distributions are. We use it to judge the probability mass function of the network output. $p(x)$ and $q(x)$ represent the probability quality function of the network output and the encoded label, respectively.

$$\mathrm{JSD} = \frac{1}{2}\sum p(x)\ln\left(\frac{p(x)}{p(x)+q(x)}\right) + \frac{1}{2}\sum q(x)\ln\left(\frac{q(x)}{p(x)+q(x)}\right) \tag{10}$$

In addition, we adopted the pose estimation evaluation index proposed by the European Space Agency (ESA) for space targets[41]. The orientation error and position error are calculated as

$$\begin{aligned} E_q &= 2\arccos(|<\hat{q}, q_{gt}>|) \\ E_l &= \|\hat{l} - l_{gt}\|_2 \end{aligned} \tag{11}$$

2.4 Experimental results

We trained the model on the SPEED+, URSO, and our dataset seperately. The loss in the training process is shown in Fig.4. It can be seen that the loss decreased significantly in the first 20 epochs of training. In addition, we provided the position error curves of the DMANET network for the X, Y, and Z axes during the training of our proposed dataset.

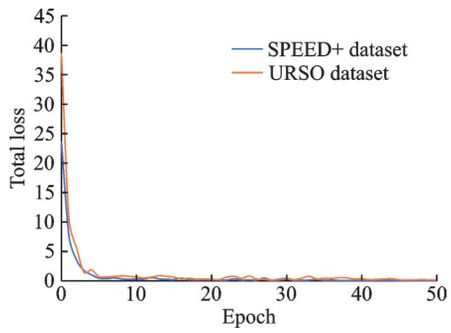


Fig.4 Graph plot of total loss in the training step in the SPEED+ and URSO datasets

In the process of DMANet training, we performed a position estimation on the test set for each epoch trained to verify the training effect, and the network weight were not be updated during the test. As shown in Fig.5, it can be seen that after training to 60 epochs, the network converged, and the position error of the target in the X-axis was estimated to be between 0.03 m and 0.08 m, the position error in the Y-axis was between 0.02 m and 0.05 m, and the position error in the Z-axis was between 0.01 m and 0.02 m.

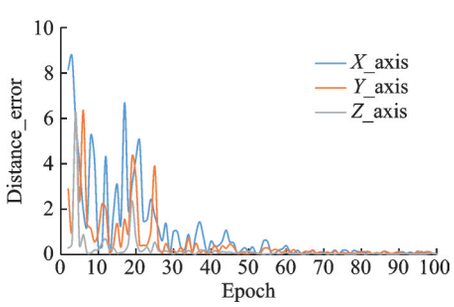


Fig.5 Distance error in the training phase in our dataset

We conducted experiments on lightbox images and sunlamp images in the SPEED+ dataset, and part of the experimental results are shown in Figs.6 (a,b) are from the lightbox test set, and Figs.6(c, d) are from the sunlamp test set. The Euler angle estimated by the decoded probability mass function of the trained network output and the Euler angle of the label are represented by the red line and the blue line, respectively, in the polar coordinate diagram. In addition, we also conducted tests on the URSO dataset. Different from the SPEED+ dataset, the

URSO dataset consists of RGB images, and we converted the images to grayscale. The experimental effects of orientation estimation are shown in Fig.7. Finally, we conducted pose estimation experiments on our proposed dataset. Part of the experimental results are shown in Fig.8. Due to the relatively simple background of our dataset, pose estimation accuracy is high.

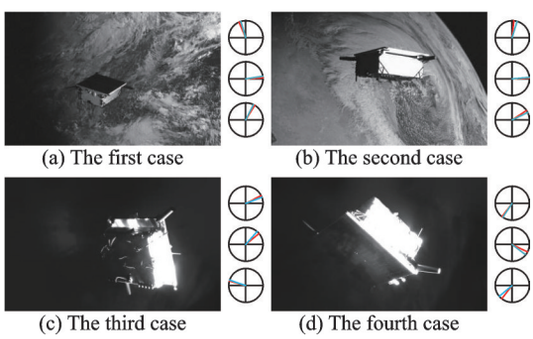


Fig.6 Cases from SPEED+ testing sets with predicted and ground truth orientations

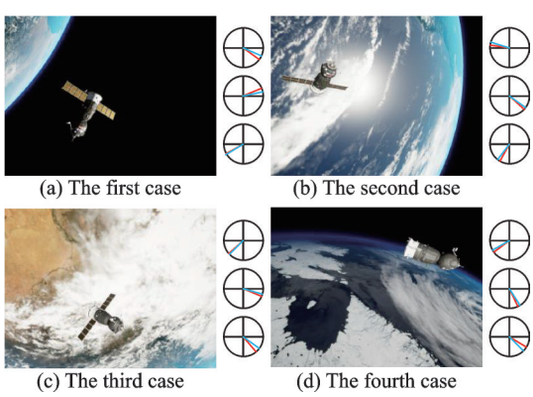


Fig.7 Cases from URSO dataset with predicted and ground truth orientations

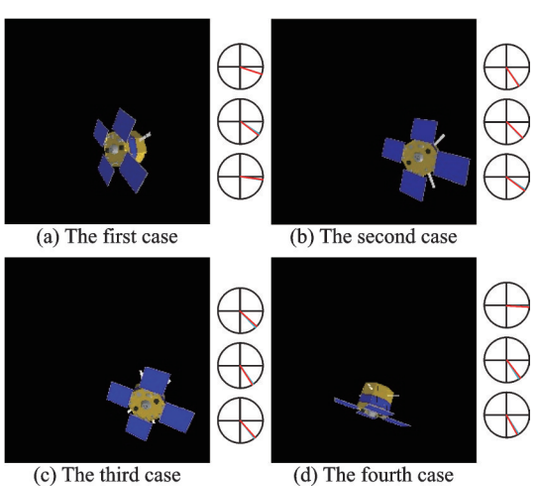


Fig.8 Cases from our dataset with predicted and ground truth orientations

The evaluation results of position estimation and orientation estimation to SPEED+ dataset and URSO dataset are shown in Table 2. It can be seen that the accuracy of DMANet is remarkable, and it has the ability of high-precision pose estimation. Furthermore, Table 3 shows ESA Pose Estimation 2021 Lightbox final scores. It can be seen that compared with other advanced methods, our method can guarantee the accuracy of both position and orientation estimation. This is due to the design of a separate depth information branch in our DMANet to obtain the location information of space non-cooperative targets.

Table 2 Evaluation results of DMANet on the SPEED+, URSO, and our dataset

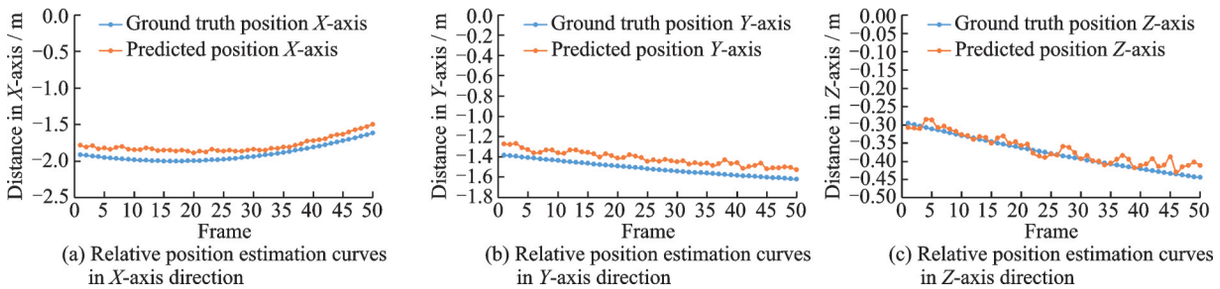
Dataset	E_t	E_q	JSD
Lightbox (SPEED+)	0.206	2.07	0.073
Sunlamp (SPEED+)	0.191	2.75	0.128
URSO	0.498	5.79	0.091
Our dataset	0.120	1.12	0.062

Table 3 ESA Pose Estimation 2021 Lightbox final scores

Algorithm	E_t	E_q
KARI_hhojeon [40]	0.395	**1.723**
Ozroto [40]	0.289	2.203
DMANet	**0.206**	2.07

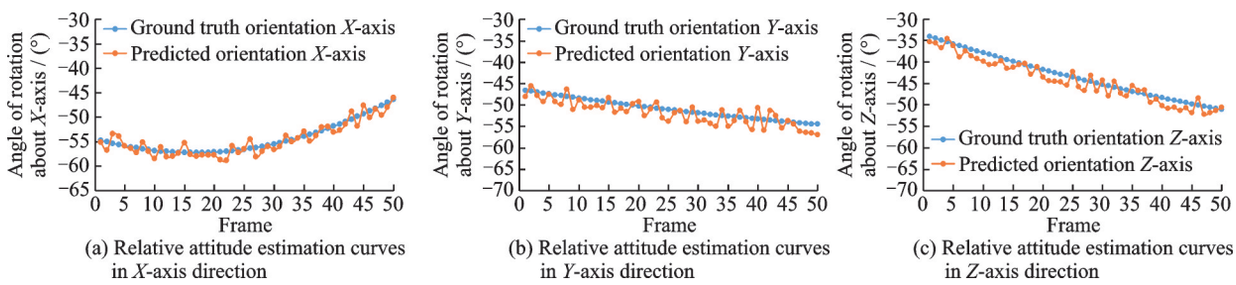
Through Coppeliasim, we simulated the process of continuous position and orientation changes of space non-cooperative objects in space, recorded the truth data and the output results of the DMANet network, and intercepted 50 sets of data in the test set (images from the 900th to the 950th frames). Relative position curves (X, Y, Z) and relative orientation estimation curves of space non-cooperative targets rotating around (X, Y, Z) axes are drawn, as shown in Figs.9, 10, and relative orientation estimation error curves are shown in Fig.11.

It can be seen that in terms of position estimation, in the direction of the X-axis, Y-axis, and Z-axis, our estimation error is about 0.1 m, 0.08 m, and 0.02 m, respectively, and the position estimation on Z-axis is the most accurate, which is consistent with our network structure. In terms of attitude



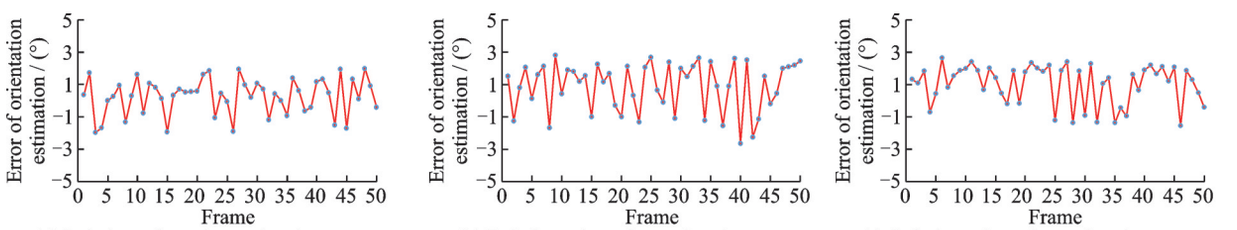
(a) Relative position estimation curves in *X*-axis direction (b) Relative position estimation curves in *Y*-axis direction (c) Relative position estimation curves in *Z*-axis direction

Fig.9 Curves of relative position estimation for 50 frames of image



(a) Relative attitude estimation curves in *X*-axis direction (b) Relative attitude estimation curves in *Y*-axis direction (c) Relative attitude estimation curves in *Z*-axis direction

Fig.10 Curves of relative attitude estimation for 50 frames of images



(a) Relative orientation estimation error curve in *X*-axis direction (b) Relative orientation estimation error curve in *Y*-axis direction (c) Relative orientation estimation error curve in *Z*-axis direction

Fig.11 Curves of relative orientation estimation error for 50 frames of images

estimation, we also estimated the rotation angles around the *X*-axis, *Y*-axis and *Z*-axis. As can be seen from Figs.10(a,b,c), the error of our estimation of the angles is 1°—2°.

2.5 Ablation study

In order to prove the validity of the DMA structure, we conducted a series of ablation experiments. It should be noted that both the ablation study and the hyperparameter experiment are based on the SPEED+ dataset.

First, we eliminated the dense structure in DMA and directly connect four MABs in serial, the network was named w.o. skip-connection. Then we removed the channel attention mechanism and multi-scale convolution from the MAB module, and obtained w.o. channel attention and w.o. multi-scale convolution, respectively. The results of the ablation study are shown in Table 4.

Table 4 Results of ablation experiments

Method	E_t	E_q	JSD
DMANet	**0.206**	**2.07**	**0.073**
w.o. skip-connection	0.263	3.26	0.161
w.o. channel attention	1.12	4.08	0.229
w.o. multi-scale convolution	1.02	3.13	0.090

We still used the evaluation indexes proposed in Section 2.3, and it can be seen that the vanilla DMANet achieved the best pose experimental results obtained by using the DMA backbone, both for the position estimation and the orientation estimation. The most important factor affecting the pose estimation accuracy was the channel attention module. In conclusion, such experimental results fully prove the effectiveness of DMA feature extraction.

2.6 Hyperparameter experiment

In the process of training DMANet, we found

that when encoding image labels, different quantity direction dimensions would affect the experimental accuracy, so we set $M=3, 4, 6, 12$, and the experimental results are shown in Table 5. It can be clearly seen that when $M=6$, the prediction accuracy is relatively high, and when M is too small, the coding classification category is too small, and the accuracy will be low. Plus, the network structure will become more complex with the increase of M, which is not conducive to the improvement of the pose estimation accuracy.

Table 5 Experimental results of different numbers of bins

Bins	E_q	JSD
3	4.36	0.165
4	2.75	0.121
6	**2.07**	**0.073**
12	2.44	0.098

In addition, it can be seen that the designed loss function contains four weight values as hyperparameters. Setting different weight values can not only balance the order of magnitude of the network loss function but also cause the network to pay different attentions to different branches. We set λ_1 and λ_2 to different values for comparative analysis, as shown in Table 6. As we see, the best experimental results can be obtained when $\lambda_1=10, \lambda_2=5$.

Table 6 Experimental results of different values of λ_1 and λ_2

λ_1	λ_2	E_1	E_q	JSD
1	1	0.692	3.306	0.090
5	1	0.447	3.923	0.182
10	1	0.231	4.120	0.155
1	5	0.606	3.436	0.133
5	5	0.390	2.367	0.114
10	5	**0.206**	2.073	0.073
1	10	0.743	3.680	0.168
5	10	0.374	2.627	0.064
10	10	0.806	**1.830**	**0.057**

3 Conclusions

We propose a DMANet model based on DMA backbone network for 6-DOF pose estimation of space non-cooperative objects and use the Euler angle-based soft classification method to transform the regression problem into a classification problem. We build a virtual simulation environment by Coppeliasim, and construct a space non-cooperative object pose estimation dataset. Finally, we perform evaluation experiments on SPEED+, URSO, and our dataset to prove that our proposed DMANet can achieve high-precision pose estimation. We prove the effectiveness of DMA module through ablation studies. However, The DMANet architecture proposed in this paper contains a large number of parameters that need to be trained. We intend to design a more lightweight model and improve the generalization ability of the network in future studies.

References

[1] TAYLOR B, AGLIETTI G, FELLOWES S, et al. Remove debris mission, from concept to orbit[C]// Proceedings of SmallSat 2018—32nd Annual AIAA/USU Conference on Small Satellites. [S.l.]: AIAA, 2018: 1-10.

[2] WANG C, NIE H, CHEN J, et al. Design of a non-cooperative target capture mechanism[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2019, 36(1): 146-153.

[3] CASSINIS L P, FONOD R, GILL E. Review of the robustness and applicability of monocular pose estimation systems for relative navigation with an uncooperative spacecraft[J]. Progress in Aerospace Sciences, 2019, 110: 100548.

[4] FORSHAW J L, AGLIETTI G S, NAVARATHINAM N, et al. RemoveDEBRIS: An in-orbit active debris removal demonstration mission[J]. Acta Astronautica, 2016, 127: 448-463.

[5] ZHOU D, SUN G, SONG J, et al. 2D vision-based tracking algorithm for general space non-cooperative objects[J]. Acta Astronautica, 2021, 188: 193-202.

[6] USC satellite data base [DB/OL]. (2023-05-00). https://www.ucsusa.org/resources/satellite-database

[7] CHEN S, LI J, XIE Y, et al. Approaching intention prediction of orbital maneuver based on dynamic Bayesian network[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2023, 40(4): 460-471.

[8] OPROMOLLA R, FASANO G, RUFINO G, et al. A review of cooperative and uncooperative spacecraft

pose determination techniques for close-proximity operations[J]. Progress in Aerospace Sciences, 2017, 93: 53-72.

[9] GARCIA A, MUSALLAM M A, GAUDILLIERE V, et al. LSPNet: A 2D localization-oriented spacecraft pose estimation neural network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 2048-2056.

[10] ZHOU D, SUN G, LEI W, et al. Space noncooperative object active tracking with deep reinforcement learning[J]. IEEE Transactions on Aerospace and Electronic Systems, 2022, 58(6): 4902-4916.

[11] ZHOU D, SUN G, ZHANG Z, et al. On deep recurrent reinforcement learning for active visual tracking of space noncooperative objects[J]. IEEE Robotics and Automation Letters, 2023, 8(8): 4418-4425.

[12] SUN D, HU L, DUAN H, et al. Relative pose estimation of non-cooperative space targets using a TOF camera[J]. Remote Sensing, 2022, 14(23): 6100.

[13] OPROMOLLA R, FASANO G, RUFINO G, et al. Uncooperative pose estimation with a LIDAR-based system[J]. Acta Astronautica, 2015, 110: 287-297.

[14] AGHILI F, KURYLLO M, OKOUNEVA G, et al. Fault-tolerant position/attitude estimation of free-floating space objects using a laser range sensor[J]. IEEE Sensors Journal, 2010, 11(1): 176-185.

[15] DAVIS J, PERNICKA H. Proximity operations about and identification of non-cooperative resident space objects using stereo imaging[J]. Acta Astronautica, 2019, 155: 418-425.

[16] PESCE V, LAVAGNA M, BEVILACQUA R. Stereovision-based pose and inertia estimation of unknown and uncooperative space objects[J]. Advances in Space Research, 2017, 59(1): 236-251.

[17] LIU J, HE S, TAO Y, et al. Realtime RGB-based 3D object pose detection using convolutional neural networks[J]. IEEE Sensors Journal, 2019, 20(20): 11812-11819.

[18] KONISHI Y, HANZAWA Y, KAWADE M, et al. Fast 6D pose estimation from a monocular image using hierarchical pose trees[C]//Proceedings of Computer Vision-ECCV 2016: 14th European Conference. Amsterdam. The Netherlands: Springer International Publishing, 2016: 398-413.

[19] LI X, CAI Y, WANG S, et al. Learning category-level implicit 3D rotation representations for 6D pose estimation from RGB images[C]//Proceedings of 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). Dali, China: IEEE, 2019: 2310-2315.

[20] ZAKHAROV S, SHUGUROV I, ILIC S. Dpod: 6D pose object detector and refiner[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 1941-1950.

[21] ZHAO W, ZHANG S, GUAN Z, et al. 6D object pose estimation via viewpoint relation reasoning[J]. Neurocomputing, 2020, 389: 9-17.

[22] LIU F, FANG P, YAO Z, et al. Recovering 6D object pose from RGB indoor image based on two-stage detection network with multi-task loss[J]. Neurocomputing, 2019, 337: 15-23.

[23] ZHU Y, WAN L, XU W, et al. ASPP-DF-PVNet: Atrous spatial pyramid pooling and distance-filtered PVNet for occlusion resistant 6D object pose estimation[J]. Signal Processing: Image Communication, 2021, 95: 116268.

[24] DABBOUR A A, HABIB R, SAII M. Object pose estimation in monocular image using modified FD-CM[J]. Computer Science, 2020, 21(1): 97-112.

[25] DO T T, CAI M, PHAM T, et al. Deep-6D pose: Recovering 6D object pose from a single RGB image[EB/OL]. (2018-02-28). http://arxiv. org/pdf/1802.10367.pdf.

[26] LIU J, HE S. 6D object pose estimation without PNP[EB/OL]. (2019-02-05). http://arxiv.org/pdf/1902.01728.pdf.

[27] LIU Y, ZHOU L, ZONG H, et al. Regression-based three-dimensional pose estimation for texture-less objects[J]. IEEE Transactions on Multimedia, 2019, 21(11): 2776-2789.

[28] RAD M, LEPETIT V. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 3828-3836.

[29] HU Y, HUGONOT J, FUA P, et al. Segmentation-driven 6d object pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 3385-3394.

[30] ZUO G, ZHANG C, LIU H, et al. Low-quality rendering-driven 6D object pose estimation from single RGB image[C]//Proceedings of 2020 International

Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE, 2020: 1-8.

[31] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes[EB/OL]. (2017-11-01). http://arxiv.org/pdf/1711.00199.pdf.

[32] BESL P J, OSAMU H. A method for registration 3-D shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 239-256.

[33] KEHL W, MANHARDT F, TOMBARI F, et al. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1521-1529.

[34] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of Computer Vision—ECCV 2016: 14th European Conference. Amsterdam, The Netherlands: Springer International Publishing, 2016: 21-37.

[35] SHARMA S, BEIERLE C, D'AMICO S. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks[C]//Proceedings of 2018 IEEE Aerospace Conference. Big Sky, USA IEEE, 2018: 1-12.

[36] SHARMA S, D'AMICO S. Neural network-based pose estimation for noncooperative spacecraft rendezvous[J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(6): 4638-4658.

[37] HUANG H, ZHAO G, GU D, et al. Non-model-based monocular pose estimation network for uncooperative spacecraft using convolutional neural network[J]. IEEE Sensors Journal, 2021, 21(21): 24579-24590.

[38] PROENÇA P F, GAO Y. Deep learning for spacecraft pose estimation from photorealistic rendering[C]//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020: 6007-6013.

[39] HUANG H, SONG B, ZHAO G, et al. End-to-end monocular pose estimation for uncooperative spacecraft based on direct regression network[J]. IEEE Transactions on Aerospace and Electronic Systems, 2023, 59(5): 5378-5389.

[40] PARK T H, MÄRTENS M, LECUYER G, et al. SPEED+: Next-generation dataset for spacecraft pose estimation across domain gap[C]//Proceedings of 2022 IEEE Aerospace Conference (AERO). Big Sky, USA: IEEE, 2022: 1-15.

[41] KISANTAL M, SHARMA S, PARK T H, et al. Satellite pose estimation challenge: Dataset, competition design and results[J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(3): 4083-4098.

Authors Mr. **ZHANG Zhao** received the B.S. degree in automation from Shandong University of Science and Technology in 2019 and the M.S. degree in control science and engineering from Harbin Institute of Technology in 2021. He is currently working toward the Ph.D. degree in the Department of Control Science and Engineering, Harbin Institute of Technology. His research interests include space non-cooperative object image fusion, pose estimation, and deep learning.

Prof. **WU Ligang** received the B.S. degree in automation from the Harbin University of Science and Technology, in 2001 and the M.E. degree in navigation guidance and control and the Ph.D. degree in control theory and control engineering from the Harbin Institute of Technology, in 2003 and 2006, respectively. His current research interests include switched systems, stochastic systems, computational and intelligent systems, sliding-mode control, and advanced control techniques for power electronic systems.

Author contributions Mr. **ZHANG Zhao** designed the study, complied the models, conducted the analysis, interpreted the results and wrote the manuscript. Mr. **HU Yuhui** helped to collate the data of the experimental results and reference research, wrote part of the manuscript. Dr. **ZHOU Dong** provided guidance on the ideas and writing of the paper, analyzed the experimental results, and modified the model. Prof. **WU Ligang** directed the research content of the paper, developed research ideas for the paper, and discussed the research background. Dr. **YAO Weiran** participated in the discussion of the experimental model and standardized the writing of the paper. Mr. **LI Peng** sorted out the formula and adjusted the format of the paper. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Bei)

DMANet: 针对空间非合作目标位姿估计的密集多尺度注意力网络

张 钊，胡瑀晖，周 栋，吴立刚，姚蔚然，李 鹏

（哈尔滨工业大学航天学院，150001 哈尔滨，中国）

摘要：利用单目相机对空间非合作目标进行准确的姿态估计对于空间碎片清除、自主交会和其他在轨服务至关重要。然而，单目姿态估计方法缺乏深度信息，导致尺度不确定性问题，大大降低了其精度和实时性。本文首先提出了一种多尺度注意块（Multi-scale attention block，MAB），从输入图像中提取复杂的高维语义特征。其次，基于MAB模块，提出了空间非合作目标6自由度位姿估计的密集多尺度注意网络（Dense multi-scale attention network，DMANet），该网络由平面位置估计、深度位置估计和姿态估计3个分支组成，通过引入基于欧拉角的软分类方法，将位姿回归问题表述为经典分类问题。此外，设计了空间非合作目标模型，并利用Coppeliasim构建了姿态估计数据集。最后，与其他最先进的方法相比，在SPEED＋、URSO数据集和本文数据集上全面评估了所提出的方法。实验结果表明，该方法具有较好的姿态估计精度。

关键词：六自由度位姿估计；空间非合作目标；多尺度注意力机制；深度学习；神经网络