Attitude Estimate of On-orbit Spacecraft based on the U-Linked Network

Yejian Zhou, Weifeng Li, Yan Ma, Bingning Li, Zhengyu Wen, Lei Zhang

***Abstract*—Real-time attitude estimation of on-orbit spacecraft is a core task in various space applications. Most of the existing methods are based on long-term observation by high-resolution sensors, such as spaceborne cameras and ground-based radars. However, when the observation period is limited, it is difficult to obtain target instantaneous attitude information by these methods. To achieve instantaneous attitude estimation from a single camera image, a U-Linked network (ULNet) is proposed in this work. The prior structural constraints of key points are used to reflect the relationship between three-dimensional (3D) target attitude parameters and two-dimensional (2D) images. In this way, target attitude estimation can be solved through the feature point regression when the large-perspective image dataset can be built. The simulation results confirm the feasibility of the proposed method. Besides, the estimation performance of the proposed method also is investigated under different imaging observation conditions.**

***Index Terms*—Spacecraft monitoring, Dynamic Estimation, Deep Learning, Optical Image Interpretation.**

I. Introduction

INstantaneous attitude estimation is crucial to monitoring on-orbit spacecraft and has become a hot issue in space situation awareness applications, such as the rendezvous between cooperative spacecraft and the re-enter of defunct spacecraft. For most cooperative targets, these attitude parameters can be obtained from the state report information transmitted by themselves. However, the attitude estimation of uncooperative targets only can be achieved by using extra observation. Several methods are presented based on high-resolution imaging sensors, such as radar and camera[1], [2], [3], [4]. Most of the existing methods rely on multi-look or sequential observation strategy and fail when only one image is provided by a single sensor.

On the other hand, some exploratory methods are proposed to interpret target state parameters from the observation image using deep learning algorithms. Xie et al.[5] designed a component extraction network (CEN) to detect target components, such as body and solar wings, in ISAR images, and then the particle swarm algorithm was used to estimate the attitude and geometrical information. Park et al.[6] proposed a spacecraft pose network (SPN) for learning the nonlinear mapping between the target pose and its structure feature in the camera image. As key components are chosen to describe the spacecraft, this sort of estimation method is more like the top-down method in human posture estimation[7], [8], [9]. However, when the occlusion occurs, these components are difficult to be extracted from the images.

Inspired by the bottom-up method in human posture estimation, we propose a target instantaneous attitude estimation method based on the key points detection network. The key points distribution in two-dimensional (2D) camera images is regarded as an implicit mapping of target structures in three-dimensional (3D) world. In order to extract key points of the spacecraft components, the network structure of the classical UNet is refined[10]. According to the experience from the related works[11], [12], density convolution blocks are inserted between the down-sampling part and the up-sampling part, and the heatmap regression method is used for extraction of key points[13], [14]. Compared with the existing component extraction works, the proposed network is easy to be trained, and becomes robust to extract target feature when the component is obscured in the observation image. Besides, the proposed method has the potential for achieving real-time attitude estimation with the spaceborne equipment.

II. Target attitude estimation from the camera imagery

As shown in Fig.1, the camera coordinate system is built to describe the spacecraft attitude in the imaging moment. The W axis refers to the optic axis of the camera, which is perpendicular to the imaging plane. The U axis is image horizon axis, and the V axis

Yejian Zhou, Weifeng Li and Zhengyu Wen are with College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, P.R.China.

Yan Ma is with Beijing Institute of Tracking Telementry and Telecommunication, Beijing, 100094, P.R.China.

Bingning Li is with Xian Satellite Control Center, Xian 710071, P.R.China.

Lei Zhang is with School of Electronics and Communication Engineering, Sun Yat-sen University (Shenzhen Campus), Shenzhen, 518107, P.R.China.

Manuscript received XX XX, 2022. This work was supported by the National Natural Sciences Foundation of China (Grant No. 62101494). (Corresponding author: Yejian Zhou, email: yjzhou25@zjut.edu.cn;)

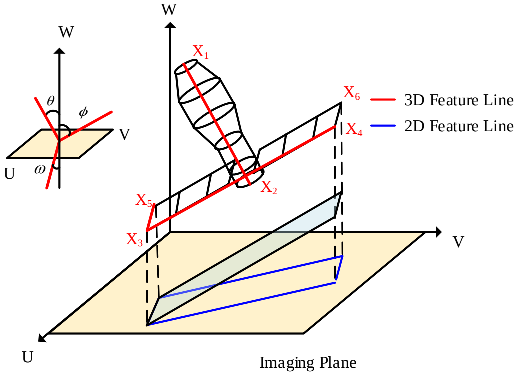

Figure 1: The imaging geometry model of key points in the camera coordinate system.

is image vertical axis. When the target trajectory and the observation geometry are determined according to the target tracking data, the estimated target attitude can be converted to the target orbit system, like earth-centered inertial system.

According to the standard pinhole model, the imaging of a key point on the spacecraft can be expressed with the following equation.

$$\left(X_P, 1\right)^{\mathrm{T}} = \mathbf{K}\left(\mathbf{R}, \vec{t}\right)\left(X_U, 1\right)^{\mathrm{T}} \tag{1}$$

where X_P represents the 2D position of the key point projected on the imaging plane P, X_U represents the 3D position matrix of key points in the camera coordinate system, $\mathbf{K}$ is the known intrinsic parameter matrix of the camera, $\mathbf{R}$ is the rotation matrix of target pose and $\vec{t}$ is the translation vector.

The depth of one-dimensional line structure, consisting of two key points, can be expressed with the ratio between its projection length in the 2D image and true size in the 3D world. Take the typical spacecraft TG-I for example. Six key points are chosen according to the symmetrical characteristic of the target structure, shown in Fig. 1. Among them, X_{U1} and X_{U2} are the endpoints of the body axis, while X_{U3}, X_{U4}, X_{U5}, and X_{U6} are the vertexes of the solar wing. As target 2D projection information on imaging plane (UOV) can be obtained directly, three attitude parameters (θ, ϕ, ω) are used to describe the intersection angles between three attitude vectors of these three line structures and the imaging plane, respectively. Details are given in the Algorithm 1. The $point\_distance(\cdot)$ is the Euclidean distance calculation between two points [14], [15], the 3D position of six points can be determined according priori target model, and the projection position of them should be extracted from the 2D image.

III. The key points extraction network

In order to extract key points of the spacecraft in the camera image, the U-Linked network (ULNet) is designed based on the UNet framework. The processing is described in the following equation:

Algorithm 1 Target pose conversion

```
Require: $X_{P1}$,$X_{P2}$, $X_{P3}$,$X_{P4}$,$X_{P5}$,$X_{P6}$
Ensure: $(\theta, \phi, \omega)$
  $L_{U12} = point\_distance(X_{U1}, X_{U2})$
  $L_{P12} = point\_distance(X_{P1}, X_{P2})$
  $L_{U34} = point\_distance(X_{U3}, X_{U4})$
  $L_{P34} = point\_distance(X_{P3}, X_{P4})$
  $L_{U35} = point\_distance(X_{U3}, X_{U5})$
  $L_{P35} = point\_distance(X_{P3}, X_{P5})$
  $\theta = arccos(L_{P12}/L_{U12})$
  $\phi = arccos(L_{P34}/L_{U34})$
  $\omega = arccos(L_{P35}/L_{U35})$
  return $(\theta, \phi, \omega)$
```

$$Y = Est(\Psi(I); \vartheta) \tag{2}$$

where ϑ represents the parameters of the network $\Psi(\cdot)$.

Once the point position vector Y is predicted by the ULNet, the position coordinates of these six points are substituted in the attitude angle calculation of the Algorithm1 .

A. The architecture of the ULNet

As shown in Fig.2, the ULNet consists of the encoder and decode part, and these two parts are connected by a series of adjacent density convolutional blocks [11], [12] . The encoder part aims to extract the target structure feature from the image, while the decoder part is expected to compress these features. Some details are below.

1) Downsampled convolutional blocks and upsampled convolutional blocks: In this work, each downsampling block consists of two 3×3 convolutional and a 2×2 maximum pooling layers. During the down-sampling, the number of feature channels is doubled to decrease the feature dimension and preserve valid features. By contrast, each up-sampling block is composed of 2×2 deconvolution layers. Both these two kinds of blocks are activated by the ReLU function. In this work, there are 24 convolutional layers used to make up the encoder and decoder parts.

2) Nearly density convolutional block: Different from the long link strategy used in the classical U-net framework, the short link strategy is adopted to avoid the semantic difference loss in the proposed network [17], [18], [19], [20]. A nearly dense convolution block is inserted to connect each up-sampling block and its relative down-sampling block. In this way, the regression performance of the proposed is improved, and the convergence is accelerated during the network training.

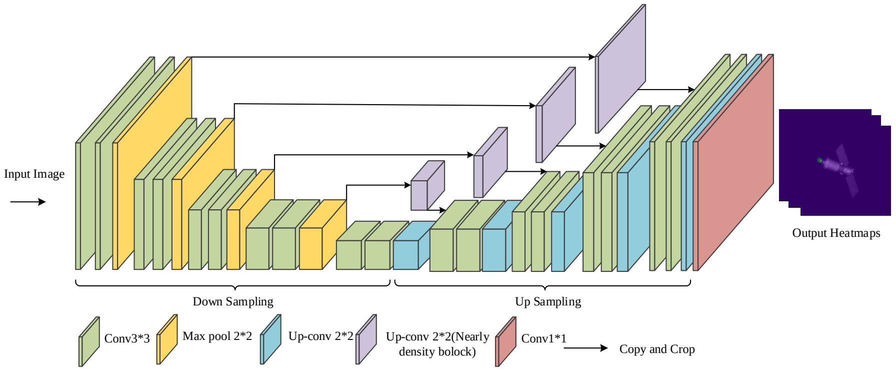

Figure 2: The architecture of the proposed ULNet.

$$b^{i,j} = \begin{cases} Cov(b^{i-1,j}) & , \quad j = 0 \\ Cov(Mg(Mg(b^{i,0})^{j-1}, Us(b^{i+1,j-1}))) & , \quad j > 0 \end{cases} \tag{3}$$

where the function $Cov(\cdot)$ represents the convolution activated by the ReLU function, $Us(\cdot)$ represents the upsampling, and $Mg(\cdot)$ represents the channel merging.

Use $b^{i,j}$ to represent the output of the convolutional block $B^{i,j}$. The superscripts i and j are the serial number of down-sampling layers and density convolution blocks, respectively. Then the feature map of this node can be determined according to Eq.(3).

B. Heatmap regression and network training

The heatmap-based approach is to calculate the confidence of each image pixel whether it is a key point. Compared with the conventional feature regression methods, the heatmap-based method performs better in the task of extracting key points[14], [15] . For a certain feature point, the heatmap $\mu_i(h,g)$ is the Gaussian distribution around its position (h_i^*, g_i^*). Therefore, a Gaussian window function $S_i(h,g)$ is designed to estimate the position of the key points as following.

$$S_i(h,g) = e^{-\frac{(h+h_i)^2+(g+g_i)^2}{2\sigma^2}} \tag{4}$$

where (h_i, g_i) is the central position of the sliding window in the heatmap, σ is the size of the sliding window, which is generally proportional to the heatmap size,$-\frac{\sigma}{2} < h < \frac{\sigma}{2}$,$-\frac{\sigma}{2} < g < \frac{\sigma}{2}$.

When the position of the sliding window coincides with that of the ground truth of key point, the output arrives the maximum. So, the heatmap decoding aims to find the position of the maximum in the predicted heatmap. And the mean square error (MSE) function is adopted as the loss function to describe the difference between the ground truth and the regression result.

$$Loss = \frac{1}{6}\sum_{i=1}^{6} \left\| S_i^{gt}(h,g) - S_i(h,g) \right\|^2 \tag{5}$$

where S_i^{gt} presents the ground truth of the heatmap of the i-th key point.

IV. Experiments

To verify the feasibility of the proposed method, a simulation image dataset of two typical spacecrafts, TG-I and Rosetta, is generated using the ray tracing algorithm[21], [22] . The 3D models of these two targets are given in Fig.3. A set of imaging views and target relative attitudes is chosen to generate 1200 optical images for the diversity of the observation views. Both horizontal and vertical resolutions are set to 0.05 m, and the size of each image is 512×512. Among these simulation images, 400 non-occluded images are selected for training, and 50 images are selected as the test samples.

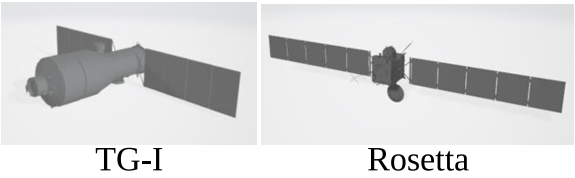

Figure 3: The 3D Models of the spacecraft

A. The attitude estimation of the typical spacecraft

In the first experiment, the proposed ULNet is used to predict the attitude parameters of the simulated images in the test dataset. Batch size is set at 4, learning rate is 0.01, and the training gradient threshold is 0.0002. Both top-down and bottom-up strategies are investigated. The top-down strategy means that the target component is extracted before target feature description. By contrast, the bottom-up strategy is to detect target key points directly according to the

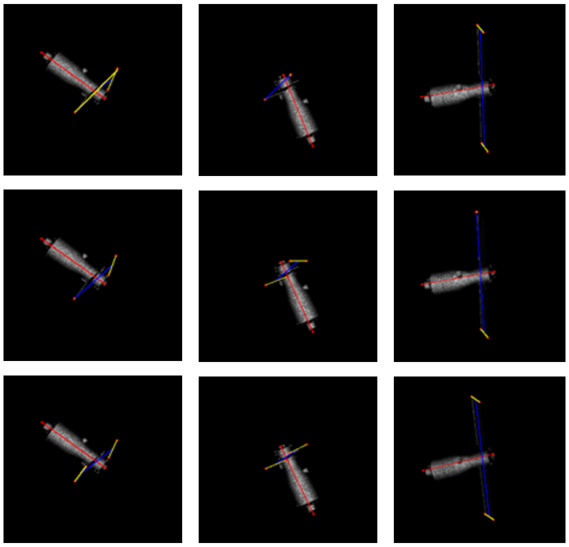

Figure 4: The extraction results of three methods. (First row: By top-down with the ULNet; Second row: By bottom-up with the UNet; Third row: By bottom-up with the ULNet)

Table I: The numerical comparison of the attitude estimation for TG-I.

Method	Error(degrees)	Time(s)
”top-down” with ULNet	(6.6,6.1,7.0)	0.54
”bottom-up” with UNet	(5.6,5.3,6.0)	1.20
”bottom-up” with ULNet	(1.5,1.3,1.7)	1.26

heatmap of each sample. The classical UNet is also adopted with the bottom-up method as the comparison method. Three frames are randomly selected from the test dataset, and the key point extraction results of them are depicted in Fig.4. The prediction results of the body and solar wing are marked in red and blue, respectively. The visional comparison shows that the prediction result of the proposed method performs better than that of the other methods in all three samples. In order to reflect the prediction performance in a quantitative way, the prediction average errors of the attitude parameters are also listed in Table I.

From Table I, it can be seen that the estimation error of the proposed method is less than 2 degrees. Compared with the conventional top-down method, the proposed method works when some part of the component is occluded. Because, the bottom-up method only extracts the key points, which still can be easily extracted in this situation. The comparison result with the UNet illustrates the advantage of the additional density convolution block introduced in the proposed method. As for most artificial spacecrafts, the shallow semantic features on the same side of the solar wing are similar. As a result, the long connection strategy during the feature fusion will cause the loss of deep semantic features in the UNet framework. And this shortcoming is fixed by the proposed ULNet. To illustrate its feasibility for other spacecraft, the pre-trained ULNet is retrained with dozens of samples of Rosetta. The averaged estimation error of the Rosetta test dataset is listed in Table II. It confirms that the proposed algorithm can work in the observation of most spacecraft.

Table II: The numerical comparison of the attitude estimation for Rosetta.

Method	Error(degrees)	Time(s)
”top-down” with ULNet	(7,6.6,6.8)	0.60
”bottom-up” with UNet	(6.6,5.3,7.3)	1.16
”bottom-up” with ULNet	(2.2,2.0,3.0)	1.19

B. *The extension experiments in different observation conditions*

In the third experiment, similar-resolution images are generated to investigate its extensiveness in different imaging conditions. The image resolutions are set to 0.06 m and 0.07 m in the new test datasets, respectively. Ten frame samples are chosen to calculate the average error of the attitude estimation in each dataset. A visional comparison of the extraction results is given in Fig.5, which reflects the feasibility of the trained ULNet for similar-resolution samples. The numerical comparison is listed in Table III. When the horizontal and vertical resolution is adjusted to 0.07 m, the bias of the attitude vector is close to 5 degrees. It is still acceptable in practical applications. The reason for this phenomenon can be explained by the multiple down-sampling layers adopted in the ULNet. Under this network framework, the feature map of each pixel is saved, so that large-scale information in the feature map can be recovered by the up-sampling processing. And the adjacent density convolutional blocks reduce the information loss during the resampling processing.

Table III: The estimation errors of different imaging resolution samples.

Method	Error(degrees)
Origin resolution	(1.5,1.3,1.7)
0.06-m resolution	(2.4,2.2,2.6)
0.07-m resolution	(3.8,3.8,3.5)

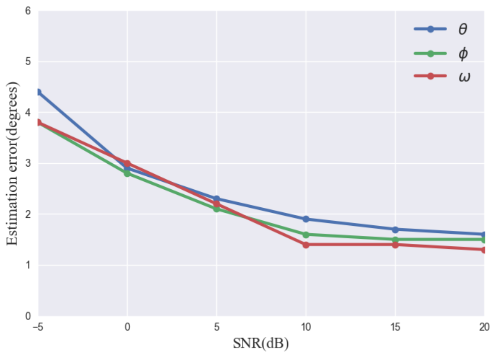

Figure 6: Attitude estimation errors in different SNR conditions.

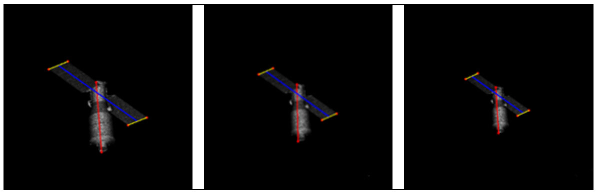

Figure 5: The feature extraction results of different resolution samples. (From left to right: Origin resolution; 0.06-m resolution; 0.07-m resolution)

Besides, in order to investigate its robustness, the proposed method is performed in different Gaussian noise conditions. As shown in Fig.6, when the signal-to-noise ratio (SNR) is higher than 10 dB the proposed method performs well. When the SNR of the image is lower than 0 dB, the estimation precision of the proposed method reduces.

V. Conclusion

In this letter, we propose a new method for real-time attitude estimation of on-orbit spacecraft. The target on-orbit attitude parameters are connected with key point features in a single image, and the U-Linked network is designed for automatic feature extraction. Simulation experiments confirm the feasibility and robustness of the proposed method with two typical spacecraft in different observation conditions. In practical applications, it has the potential for achieving real-time attitude estimation with the spaceborne camera.

References

[1] Chaoying Huo, Hongcheng Yin , and et al., ”Attitude estimation method of space targetsby 3d reconstruction of principal axis from ISAR image”. *Procedia computer science*, 147:158-164, 2019.

[2] Yan Wang, Feng Yuan, Hong Jiang, and et al., ”high precision and fast estimation of position and attitude measurement for space targets”. *Optik*, 148:76-84, 2017.

[3] Yejian Zhou, Lei Zhang, and Yunhe Cao, ”Dynamic estimation of spin spacecraft based on multiple-station ISAR images”. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2977-2989, 2019.

[4] Yejian Zhou, Lei Zhang, Yunhe Cao, and et al. ”Optical-and-Radar Image Fusion for Dynamic Estimation of Spin Satellites”. *IEEE Transactions on Image Processing*, 2020, 29:2963-2976.

[5] Pengfei Xie, Lei Zhang, Yan Ma, and et al., ”Attitude estimation and geometry inversion of satellite based on oriented object detection”. *IEEE Geoscience and Remote Sensing Letters* , 19:1-5, 2022.

[6] Tae Ha Park and Simone DAmico, ”Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap”. arXiv preprint arXiv:2203.04275, 2022.

[7] Alexander Toshev and Christian Szegedy. ”Deeppose: Human pose estimation via deep neural networks”. *the IEEE conference on computer vision and pattern recognition*, pages 1653-1660, 2014.

[8] Matteo Fabbri, Fabio Lanzi, Simone Calderara, and et al., ”Compressed volumetric heatmaps for multi-person 3d pose estimation”. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204-7213, 2020.

[9] Xiaopeng Ji, Qi Fang, Junting Dong, and et al., ”A survey on monocular 3d human pose estimation”. *Virtual Reality Intelligent Hardware*,2(6):471-500, 2020.

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

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. ”Fully convolutional networks for semantic segmentation”. *the IEEE conference on computer vision and pattern recognition*, pages 3431-3440, 2015.

[12] Hang Zhang, Chongruo Wu, Zhongyue Zhang, and et al., ”Resnest: Split-attention networks”. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2736-2746, 2022.

[13] Feng Zhang, Xiatian Zhu, Hanbin Dai, and et al., ”Distribution-aware coordinate representation for human pose estimation”. *the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093-7102, 2020.

[14] Zhengxiong Luo, Zhicheng Wang, Yan Huang, and et al., ”Rethinking the heatmap regression for bottom-up human pose estimation”. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13264-13273, 2021.

[15] Emanuele Trucco and Alessandro Verri. ”Introductory techniques for 3-D computer vision”, volume 201. *Prentice Hall Englewood Cliffs*, 1998.3, 4.

[16] Richard Hartley and Andrew Zisserman.” Multiple view geometry in computer vision”. Cambridge university press,2003. 3, 4, 13.

[17] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and et al., ”Unet++: A nested u-net architecture for medical image segmentation”. *In Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3-11. Springer, 2018.

[18] Xingyi Zhou, Dequan Wang, and Philipp Krahenb uhl., ”Objects as points”. arXiv preprint arXiv:1904.07850, 2019.

[19] Erjin Zhou, Haoqiang Fan, Zhimin Cao, and et al., ”Extensive facial landmark localization with coarse to-fine convolutional network cascade”. *the IEEE international conference on computer visionworkshops*, pages 386-391, 2013.

[20] Forrest Iandola, Matt Moskewicz, Sergey Karayev, and et al., ”Densenet: Implementing efficient convnet descriptor pyramids”. arXiv preprint arXiv:1404.1869, 2014.

[21] Arthur Appel. ”Some techniques for shading machine renderings of solids”. *spring joint computer conference 1968*, pages 3745, 1968.

[22] Turner Whitted. ”An improved illumination model for shaded display”. In ACM Siggraph 2005 Courses, pages 4. 2005.