Attitude Estimation of On-Orbit Spacecraft Based on the U-Linked Network

Yejian Zhou, Weifeng Li, Yan Ma, Bingning Li, Zhenyu Wen, and 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 space-borne 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 3-D target attitude parameters and 2-D 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*— Deep learning, dynamic estimation, optical image interpretation, spacecraft monitoring.**

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 reenter 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 multilook 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 and DAmico [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 topdown 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 bottomup method in human posture estimation, we propose a target instantaneous attitude estimation method based on the key points detection network. The key point distribution in 2-D camera images is regarded as an implicit mapping of target structures in 3-D 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 downsampling part and the upsampling 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 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:

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

where X_P represents the 2-D position of the key point projected on the imaging plane P, X_U represents the 3-D

Manuscript received 26 July 2022; revised 3 October 2022 and 24 October 2022; accepted 4 November 2022. Date of publication 9 November 2022; date of current version 18 November 2022. This work was supported by the National Natural Science Foundation of China under Grant 62101494. *(Corresponding author: Yejian Zhou.)*

Yejian Zhou, Weifeng Li, and Zhenyu Wen are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: yjzhou25@zjut.edu.cn).

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

Bingning Li is with the Xi'an Satellite Control Center, Xi'an 710071, China.

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

Digital Object Identifier 10.1109/LGRS.2022.3220764

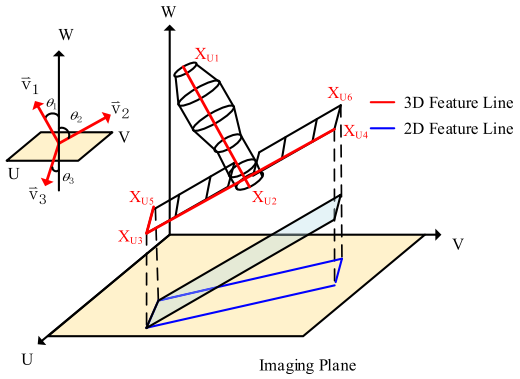

Fig. 1. Imaging geometry model of key points in the camera coordinate system.

position matrix of key points in the camera coordinate system, **K** is the known intrinsic parameter matrix of the camera, **R** is the rotation matrix of target pose, and $\vec{t}$ is the translation vector.

Algorithm 1 Target Pose Conversion

Require: X_{P1},X_{P2}, X_{P3},X_{P4},X_{P5},X_{P6}

$l_{U12} = \|X_{U1}(u', v', w') - X_{U2}(u', v', w')\|$,
$l_{U34} = \|X_{U3}(u', v', w') - X_{U4}(u', v', w')\|$,
$l_{U35} = \|X_{U3}(u', v', w') - X_{U5}(u', v', w')\|$

$L_{P1}(u, v) = X_{P1}(u, v) - X_{P2}(u, v)$,
$L_{P2}(u, v) = X_{P3}(u, v) - X_{P4}(u, v)$,
$L_{P3}(u, v) = X_{P3}(u, v) - X_{P5}(u, v)$

$\theta_1 = \arcsin(\frac{\|L_{P1}\|}{L_{U12}})$, $\theta_2 = \arcsin(\frac{\|L_{P2}\|}{L_{U34}})$,
$\theta_3 = \arcsin(\frac{\|L_{P3}\|}{L_{U35}})$

$\vec{v}_1 = (\frac{L_{P1}(u)}{l_{U12}}, \frac{L_{P1}(v)}{l_{U12}}, \cos\theta_1)^{\mathrm{T}}$,
$\vec{v}_2 = (\frac{L_{P2}(u)}{l_{U34}}, \frac{L_{P2}(v)}{l_{U34}}, \cos\theta_2)^{\mathrm{T}}$,
$\vec{v}_3 = (\frac{L_{P3}(u)}{l_{U35}}, \frac{L_{P3}(v)}{l_{U35}}, \cos\theta_3)^{\mathrm{T}}$

return $(\vec{v}_1, \vec{v}_2, \vec{v}_3)$

As the 2-D projection information on imaging plane can be directly obtained from the image, the 3-D pointing vector of target typical line structures can be determined if its depth is calculated [14], [15]. Furthermore, once three pointing vectors are estimated in different dimensions, the target pose is obtained in the 3-D world. Take the typical spacecraft TG-I for example. According to the symmetrical characteristic of the target structure in Fig. 1, three vectors $(\vec{v}_1, \vec{v}_2, \vec{v}_3)$ are adopted to reflect the attitude information of target typical line structures, the symmetrical axis of the body, the long edge of the solar wing, and the short edge of the solar wing. The depth of each line structure can be expressed with the ratio between its projection length in the 2-D image and true 3-D size. Therefore, the 3-D coordinates of six key points (from X_{U1} to X_{U6}) are measured, and their projection positions (from X_{P1} to X_{P6}) need to be extracted from the image. In this way, target attitude pointing vectors can be calculated. Details are given in Algorithm 1.

III. 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:

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

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 Algorithm 1.

A. 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 down-sampling block consists of two 3×3 convolutional and a 2×2 maximum pooling layers. During the downsampling, the number of feature channels is doubled to decrease the feature dimension and preserve valid features. By contrast, each upsampling 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.

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 downsampling layers and density convolution blocks, respectively. Then the feature map of this node can be determined according to the following equation (3), as shown at the bottom of the next page.

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, $-(\sigma/2) < h < (\sigma/2)$, $-(\sigma/2) < g < (\sigma/2)$.

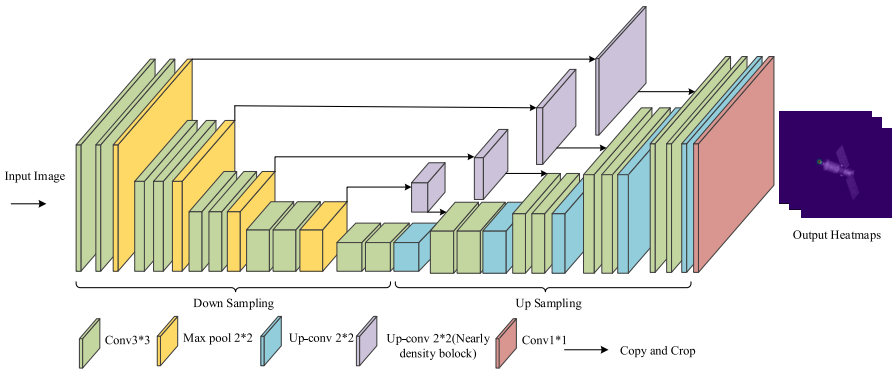

Fig. 2. Architecture of the proposed ULNet.

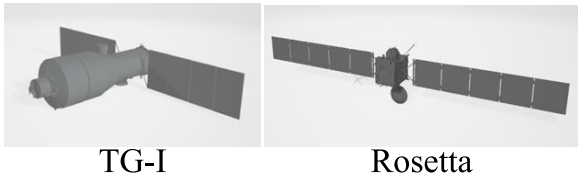

Fig. 3. 3-D models of the spacecraft.

When the position of the sliding window coincides with that of the ground truth of the key point, the output arrives at 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

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

where S_i^{gt} presents the ground truth of the heatmap of the ith key point.

IV. Experiments

To verify the feasibility of the proposed method, a simulation image dataset of two typical spacecraft, TG-I and Rosetta, is generated using the ray tracing algorithm [21], [22]. The 3-D 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 nonoccluded images are selected for training, and 50 images are selected as the test samples.

A. 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. The batch size is set at 4, the learning rate is 0.01, and the training gradient threshold is 0.0002. Both topdown and bottomup strategies are investigated. The topdown strategy means that the target component is extracted before the target feature description. By contrast, the bottomup strategy is to detect target key points directly according to the heatmap of each sample. The classical UNet is also adopted with the bottomup 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.

TABLE I
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

From Table I, it can be seen that the estimation error of the proposed method is less than 2°. Compared with the conventional topdown method, the proposed method works when some part of the component is occluded. Because, the bottomup 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 spacecraft, the shallow semantic features on the

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

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

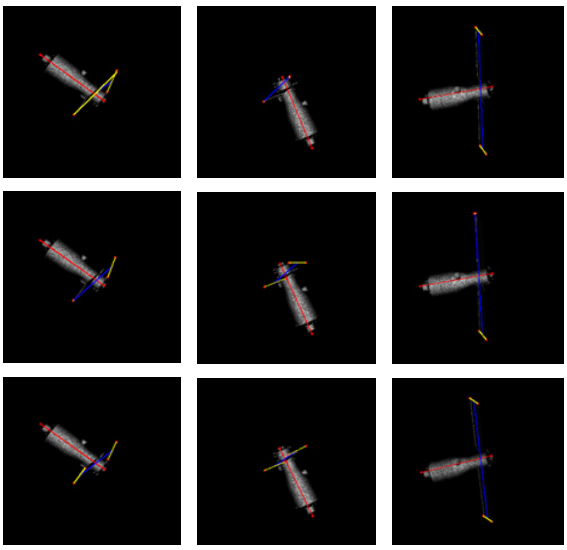

Fig. 4. Extraction results of three methods. [First row: (topdown) with the ULNet; second row: (bottomup) with the UNet; third row: (bottomup) with the ULNet].

TABLE II
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

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 pretrained 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.

B. Extension Experiments in Different Observation Conditions

In the third experiment, similar-resolution images are generated to investigate their extensiveness in different imaging conditions. The image resolutions are set to 0.06 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°. 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

TABLE III
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)

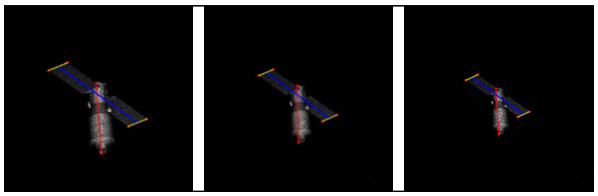

Fig. 5. Feature extraction results of different resolution samples. [(Left to right) Origin resolution; 0.06-m resolution; 0.07-m resolution.]

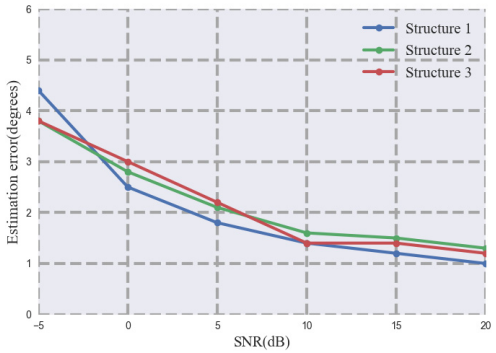

Fig. 6. Attitude estimation errors in different SNR conditions.

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.

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 ULNet 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] C.-Y. Huo, H.-C. Yin, X. Wei, X.-Y. Xing, and L. Man, "Attitude estimation method of space targets by 3D reconstruction of principal axis from ISAR image," *Proc. Comput. Sci.*, vol. 147, pp. 158–164, Jan. 2019.

[2] Y. Wang, F. Yuan, H. Jiang, and W. Chen, "High precision and fast estimation of position and attitude measurement for space targets," *Optik*, vol. 148, pp. 76–84, Nov. 2017.
[3] Y. Zhou, L. Zhang, and Y. Cao, "Dynamic estimation of spin spacecraft based on multiple-station ISAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2977–2989, Apr. 2020.
[4] Y. Zhou, L. Zhang, Y. Cao, and Y. Huang, "Optical-and-radar image fusion for dynamic estimation of spin satellites," *IEEE Trans. Image Process.*, vol. 29, pp. 2963–2976, 2020.
[5] P. Xie, L. Zhang, Y. Ma, Y. Zhou, and X. Wang, "Attitude estimation and geometry inversion of satellite based on oriented object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[6] T. H. Park and S. D'Amico, "Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap," 2022, *arXiv:2203.04275*.
[7] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.
[8] M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, and R. Cucchiara, "Compressed volumetric heatmaps for multi-person 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7204–7213.
[9] X. Ji, Q. Fang, J. Dong, Q. Shuai, W. Jiang, and X. Zhou, "A survey on monocular 3D human pose estimation," *Virtual Reality Intell. Hardw.*, vol. 2, no. 6, pp. 471–500, Dec. 2020.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[12] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2736–2746.
[13] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7093–7102.
[14] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13264–13273.
[15] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*, vol. 201. Englewood Cliffs, NJ, USA: Prentice-Hall, 1998.
[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.
[18] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[19] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 386–391.
[20] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient ConvNet descriptor pyramids," 2014, *arXiv:1404.1869*.
[21] A. Appel, "Some techniques for shading machine renderings of solids," in *Proc. Spring Joint Comput. Conf. AFIPS (Spring)*, 1968, p. 37.
[22] T. Whitted, "An improved illumination model for shaded display," in *Proc. ACM SIGGRAPH Courses (SIGGRAPH)*, 2005, p. 4.