sensors

MDPI

Article

Multi-Scale Deep Neural Network Based on Dilated Convolution for Spacecraft Image Segmentation

Yuan Liu [1,2], Ming Zhu [1,*], Jing Wang [1], Xiangji Guo [1,2], Yifan Yang [1,2] and Jiarong Wang [1]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; liuyuan18@mails.ucas.ac.cn (Y.L.); wangjing@ciomp.ac.cn (J.W.); guoxiangji18@mails.ucas.ac.cn (X.G.); yangyifan17@mails.ucas.ac.cn (Y.Y.); wangjiarong@cust.edu.cn (J.W.)
[2] School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zhuming@ciomp.ac.cn or zhu_mingca@163.com

Abstract: In recent years, image segmentation techniques based on deep learning have achieved many applications in remote sensing, medical, and autonomous driving fields. In space exploration, the segmentation of spacecraft objects by monocular images can support space station on-orbit assembly tasks and space target position and attitude estimation tasks, which has essential research value and broad application prospects. However, there is no segmentation network designed for spacecraft targets. This paper proposes an end-to-end spacecraft image segmentation network using the semantic segmentation network DeepLabv3+ as the basic framework. We develop a multi-scale neural network based on sparse convolution. First, the feature extraction capability is improved by the dilated convolutional network. Second, we introduce the channel attention mechanism into the network to recalibrate the feature responses. Finally, we design a parallel atrous spatial pyramid pooling (ASPP) structure that enhances the contextual information of the network. To verify the effectiveness of the method, we built a spacecraft segmentation dataset on which we conduct experiments on the segmentation algorithm. The experimental results show that the encoder+ attention+ decoder structure proposed in this paper, which focuses on high-level and low-level features, can obtain clear and complete masks of spacecraft targets with high segmentation accuracy. Compared with DeepLabv3+, our method is a significant improvement. We also conduct an ablation study to research the effectiveness of our network framework.

Keywords: semantic segmentation; deep learning; dilated convolution; multi-scale; DeepLabv3+

Citation: Liu, Y.; Zhu, M.; Wang, J.; Guo, X.; Yang, Y.; Wang, J. Multi-Scale Deep Neural Network Based on Dilated Convolution for Spacecraft Image Segmentation. *Sensors* **2022**, *22*, 4222. https://doi.org/10.3390/s22114222

Academic Editors: Bin Fan and Wenqi Ren

Received: 9 May 2022
Accepted: 25 May 2022
Published: 1 June 2022

1. Introduction

Since the beginning of the 21st century, human beings have accelerated the pace of space exploration, and space technology plays a vital role in it. Rapidly evolving computer vision and machine learning techniques have facilitated space technology development through applications in tasks such as collision avoidance self-navigation systems, spacecraft health monitoring, and asteroid classification [1,2]. The on-orbit formation and operation of large space equipment, such as space stations and the work and maintenance of spacecraft, cannot be completed without the support of space technology. Performing space missions often requires spacecraft images acquired by vision sensors as inputs, using computer vision technology to determine spacecraft position and attitude information, and then performing complex tasks, such as spacecraft rendezvous and docking, space equipment assembly on-orbit, spacecraft grasping and maintenance, and space debris removal [3–5]. In recent years, many spacecraft have been launched into space, the number of space exploration missions has increased, and the low Earth orbit has become crowded. Defunct satellites and space debris are growing and urgently need to be cleaned up [6]. Spacecraft positioning and other related issues have attracted more and more attention from researchers in space technology. The successful segmentation of spacecraft objects in the image is the key to

the accurate positioning of spacecraft. The fine and precise spacecraft mask obtained after segmentation is conducive to the keypoint detection of the object, which is crucial for the realization of vision-based attitude estimation [7]. This paper investigates this issue.

Spacecraft image segmentation is to separate the spacecraft object and background from the acquired image. Compared with object detection, spacecraft image segmentation can obtain more accurate position information of the object and obtain more precise object contour, which is more difficult to achieve. Considering the particularity of the environment of the on-orbit mission, the spacecraft image segmentation task needs to face more severe lighting conditions, such as low light, overexposure, and spacecraft reflection, than the general image segmentation task. The impact of complex background and noise on the mission also needs to be considered. All these problems put forward higher requirements for the robustness and accuracy of the segmentation algorithm. Spacecraft objects can be divided into cooperative and non-cooperative spacecraft in different types of missions [8]. Cooperative spacecraft may be positioned and docked using dedicated radio links, reference markers, backward reflectors, etc. Non-cooperative spacecraft may be unknown spacecraft that do not have cooperative conditions. Cooperative spacecraft are designed with cooperative targets for precise docking at the grasping position for the on-orbit assembly mission. However, before docking is performed, it is necessary to operate the manipulator appropriately to ensure that the cooperative marks of the spacecraft are within the field of view of the hand-eye camera. Using spacecraft image segmentation techniques, even if the target spacecraft is non-cooperative, the spacecraft position can be obtained as long as the appearance of the spacecraft is known. The European Space Agency (ESA) and Stanford University held a competition in 2019 to use supervised learning to estimate the position and attitude of a known spacecraft from an image [9]. This paper belongs to this research direction but only studies the positioning of known spacecraft through monocular images.

In the past decade, deep learning has developed rapidly. Image segmentation tasks have achieved far better results than traditional machine learning methods through deep neural networks. Several branches, such as semantic segmentation, instance segmentation, and panoptic segmentation, have been developed. Image segmentation for known spacecraft is well suited for processing using deep learning methods. Researchers can build a spacecraft dataset with labelled information based on specific spacecraft application scenarios and train the network until the model converges. During inference, the network can quickly make predictions on the images and derive spacecraft segmentation results. However, space data are sensitive, and it is difficult to establish a real space object dataset. Some publicly available segmentation datasets, such as the Pascal Visual Object Classes (Pascal VOC), the Microsoft Common Objects in COntext (COCO), and the Cityscapes, are unsuitable for validating spacecraft segmentation algorithms [10–12]. The publicly available spacecraft datasets, such as Spacecraft Pose Estimation Dataset (SPEED) and the Unreal Rendered Spacecraft on Orbit (URSO), are mainly used to solve the attitude problem and lack the annotation of segmentation information [4,9]. Therefore, based on SPEED and URSO, we selected many photorealistic spacecraft images for refined annotation and constructed a spacecraft image segmentation dataset. We do not use the same annotation method as the spacecraft component segmentation dataset established in Ref. [7]. Our research is more concerned with whether the entire spacecraft can be accurately segmented. At the same time, we distinguish different kinds of spacecraft in the dataset annotation. We use a semantic segmentation network to solve the problem of spacecraft image segmentation. We also design the internal structure model of a deep neural network.

Inspired by DeepLabv3+ [13] and DenseASPP [14], we propose a dilated convolution-based multi-scale neural network for spacecraft image segmentation, which we call SISnet. The network uses DeepLabv3+ as the basic framework and adopts the encoding and decoding form proposed by U-Net [15]. We use the spacecraft image segmentation dataset we made as the experimental dataset to verify the segmentation ability of the network model. Specifically, in the backbone network part, our network model uses dilated residual networks (DRN) [16] as the encoder for deep feature extraction. Our encoder utilizes dilated convolution, enabling

the network to increase the receptive field while maintaining image resolution. Compared with residual networks [17], our backbone network removes the residual structure and skip-connections in the last few layers, which improves the gridding artifacts. Inspired by FarSee-Net [18] and Bai et al. [19], we utilize two atrous space pyramid pooling (ASPP) modules in the network to form a parallel-ASPP structure. The high-level and low-level features extracted from the backbone network are fused in the parallel-ASPP structure to improve the segmentation effect of the model for multi-scale targets. In addition, inspired by SENet [20] and ECA-Net [21], we have added a Squeeze-and-Excitation Module (SEM) between the backbone network and the parallel-ASPP structure. SEM is a channel attention mechanism module, which enables the network to pay more attention to some feature maps by giving channel weights. The innovation of this paper is that, through dilated convolutional and the parallel-ASPP structure, our network enhances contextual information and mitigates the effects of gridding artifacts. At the same time, the channel attention mechanism is used to recalibrate the features, which improves the learning ability of the network. Our method is more robust to noise in the image and can segment more complete and clear spacecraft masks. We experimented with SISnet on the spacecraft image segmentation dataset we produced above and compared it with other deep learning neural network methods. We experimented with SISnet on a spacecraft image segmentation dataset we produced and compared it with other deep learning neural network approaches. Experiments show that our method achieves a higher level with better segmentation results. Our work and predicted mask effects are shown in Figure 1.

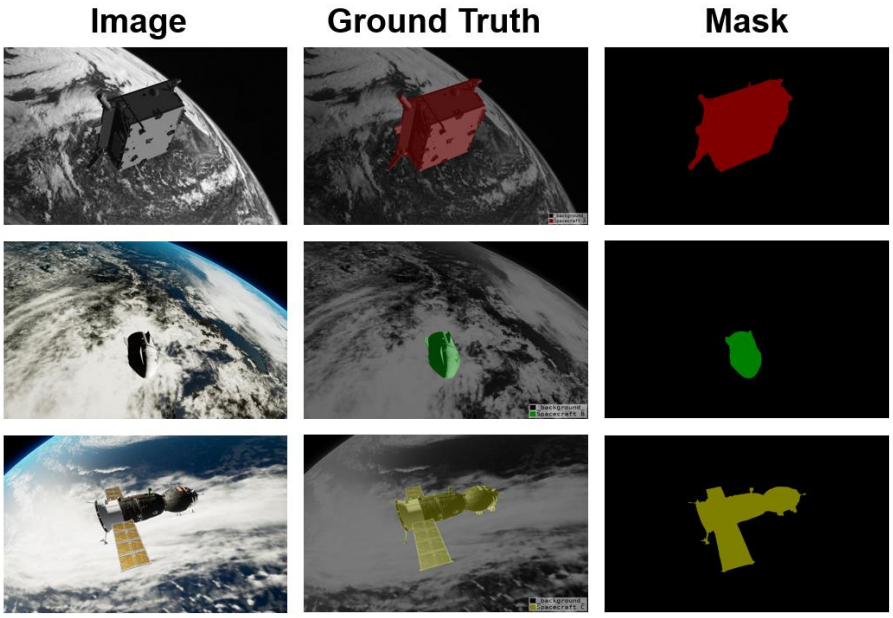

Figure 1. The display of our work.

The contributions of our paper are summarized as follows:

1. We designed an end-to-end segmentation network for spacecraft objects and produced a spacecraft image segmentation dataset to validate the algorithm.
2. We optimized the backbone network and used dilated convolution to increase the receptive field while maintaining the image resolution. With only 53 layers, our backbone achieves better results than ResNet-101 with more layers and Xception-65 with a larger number of parameters.
3. We added the channel attention mechanism to the segmentation network of the encoder–decoder structure to form an encoder + attention + decoder structure. Our network focuses on both high-level and low-level feature branches to improve the learning effect of the segmentation network.

4. We designed a parallel-ASPP structure. Using the superposition of different dilated convolutions, our network achieves multi-scale feature fusion with different-depth feature maps, enabling the network to segment the contours of spacecraft objects at different scales more clearly and completely.

The rest of the paper is organized as follows: in Section 2, we present the related work. In Section 3, we detail our proposed method. In Section 4, we detail the network training, comparison experiments, and ablation study of our approach. Quantitative and qualitative analyses are performed for the core benefits and performance of our network. In Section 5, we discuss the drawbacks of our method and the problems encountered during the research. Finally, in Section 6, we conclude the article.

2. Related Work

2.1. Image-Based Space On-Orbit Service Technology

In recent years, on-orbit service concepts, such as space debris removal, satellite acquisition, and spacecraft construction, have attracted more and more attention from academia and industry [9]. Space technology, such as spacecraft object detection and on-orbit service based on image processing, has gradually contributed to the research field of space object monitoring. The image processing technology is rapidly driving the progress of space technology. Sun et al. proposed an algorithm for adaptive real-time detection of faint optoelectronic geosynchronous orbit (GSO) debris [22] to detect a large amount of space debris in GSO that threatens the safety of spacecraft [23,24]. Through image adaptive fast registration and dilated difference algorithms, combined with mathematical morphology, threshold segmentation, and global nearest neighbor (GNN) multi-target tracking algorithms, Sun et al. achieved image background suppression, alignment, suspected target extraction, and multi-object tracking. This method can detect dim geostationary Earth orbit (GEO) and non-GEO debris in GSO. Khan et al. designed an image-based visual servo system [25]. They used the visual servo system for the first time to successfully track a super-orbital re-entry of a spacecraft and record its spectrum feature. They optimized the visual servo system by adding a simplified feedforward control dynamic model, and they verified the tracking performance of the system on the International Space Station (ISS).

Sharma et al. started research on monocular vision-guided on-orbit service technology in 2015 [26]. According to the explicit on-orbit service mission requirements of the United States and other countries [24,27], to approach space objects in an energy-efficient, safe, and accurate way, they began to explore relying on three-dimensional computer models and a single two-dimensional image for the initial pose estimation of space objects. They wanted to estimate the initial pose without using reference marks, without using any distance measurements or any prior relative motion information. Since then, the research in space object monitoring has also started to evolve from the detection and tracking of space objects to the precise acquisition of object position and attitude. This is very consistent with the direction of rigid body 6D pose estimation, which deep learning has rapidly developed in recent years. Almost all the pose estimation methods need to separate the object from the complex background, so most methods are segmentation-driven pose estimation. The precise mask of the object is crucial for subsequent processing, such as keypoint detection and pose discriminator [28].

In 2019, Stanford University and the ESA hosted the Satellite Pose Estimation Challenge, limiting input data to two-dimensional images and moving away from object 3D model data. This competition is very suitable for the application scenario where the object is a non-cooperative spacecraft without a target and the competition is also very challenging. Kisantal et al. proposed that the pose estimation of spacecraft could be divided into positioning and attitude determination, and the position error and attitude error were compared, respectively [9]. The classic monocular image-based space object positioning method iteratively solves the object position by extracting the shape context of the target, such as Harris corners, Canny edges, Hough transform, SIFT, SURF, and ORB features [29–34]. The Weak

Gradient Elimination (WGE) technique was introduced by Sharma et al. [35]. It uses simple geometric constraints to synthesize the detected features to separate the edge features of the spacecraft from the weak edge features of the background. This method greatly reduces the search space of the feature correspondence problem. A more effective approach in recent years has been the use of deep neural network methods. It implements feature extraction through convolutional neural networks and uses datasets for supervised training, allowing the algorithm to predict the position of spacecraft quickly. In 2019, Sharma et al. [36] used a convolutional neural network combined with an advanced object detection algorithm to detect the 2D boundary box of the spacecraft in the image. Pedro et al. [4] adopted the ResNet [17] architecture with pre-trained weights as the network backbone. They removed unnecessary pooling layers to preserve spatial feature resolution to compress CNN features. In 2021, Dung et al. [7] built the first dataset for space object detection and segmentation, which annotated different components of the spacecraft. After training the dataset using deep learning methods, they could implement the components recognition. Dung et al. conducted experiments with multiple state-of-the-art detection and segmentation networks. They benchmarked the dataset, but they lacked improvements to the network structure and included no analysis of the experimental results. Our study improves the network structure and optimizes the effect of our model through structures such as channel attention mechanism and dilated convolution, and we have further analyzed and discussed the spacecraft segmentation issue.

2.2. Semantic Segmentation Based on Deep Learning

Since the advent of AlexNet [37], deep neural networks have begun to dominate image classification tasks and have shown excellent classification performance. FCN [38] adapted the classification network to a fully convolutional network and made significant progress by applying the network to semantic segmentation by classifying pixels. In order to enhance contextual information aggregation to improve the effect of semantic segmentation, some variants of FCN-based models have been proposed. SegNet [39] adopts an encoder–decoder structure that utilizes low-level information to help refine segmentation masks. U-Net [15] concatenates the outputs of low-level layers with those of high-level layers for information fusion. DeepLab [40] and CRF-RNN [41] use the conditional random field as post-processing for structure prediction in scene parsing. DPN [42] implements semantic segmentation using Markov random fields. PSPNet [43] builds a pyramid structure and fuses middle-level and high-level semantic features to obtain multi-scale context information.

In semantic segmentation tasks, contextual information plays an important role in image understanding to improve segmentation quality. Dilated convolutions can increase the receptive field without losing information by inserting cavities of different rates into regular convolutions. Deeplabv2 [44] and Deeplabv3 [45] embed contextual information using an ASPP, which uses parallel dilated convolution with different rates to fuse multi-scale information and expand the receptive field. DeepLabv3+ [13] adds a decoder structure while using the Xception with depthwise separable convolution as the backbone. It refines the segmentation results and makes the boundary of the object segmented more clearly. Visin et al. [46] proposed using recurrent neural networks to retrieve correlations in the global space.

In recent years, attention mechanisms have begun to be applied to deep learning tasks. From machine translation to image classification, attention mechanisms have demonstrated excellent performance. The self-attention mechanism [47] has been pioneered to achieve good results in machine translation by extracting the global dependencies of the input. SENet [20] proposes a lightweight channel attention mechanism that adaptively recalibrates the channel feature maps by establishing interdependencies between channels through the "Squeeze-and-Excitation" (SE) block. BAM [48] and CBAM [49] design attention mechanisms for both channel and spatial dimensions in a similar way to achieve adaptive feature refinement. PSANet [50] makes the network segmentation mask more sensitive to

position and category information by the point-wise spatial attention module to adaptively aggregate contextual information for each point. DANet [51] expanded the self-attention mechanism in the segmentation task. It utilizes two attention mechanisms, which combine adaptive local features with their global dependencies to capture rich contextual relations. SKNet [52] designed a selective kernel unit that fuses the effective receptive fields of neurons of different sizes through branches of different kernel sizes. EMANet [53] proposed to describe the attention mechanism as an expectation-maximization method using the expectation-maximization attention module to perform mask estimation iteratively. Table 1 below shows a comparison of our proposed approach with several related methods. The first line indicates: whether to use the segmentation method, whether to adopt a deep learning strategy, whether the network design is carried out according to the characteristics of the spacecraft object, and whether relevant datasets have been established.

Table 1. The comparison table of related work.

Method	Segmentation	Deep Learning	Spacecraft Object	Dataset
Khan et al. [25]	no	no	yes	no
Sharma et al. [36]	no	yes	yes	Pose
Proenca et al. [4]	no	yes	no	Pose
Dung et al. [7]	yes	yes	no	Segmentation
DeepLabv3+ [13]	yes	yes	no	no
Ours	yes	yes	yes	Segmentation

3. Methods

The SISnet network proposed in this paper is an end-to-end self-supervised learning network. The network adopts the structure of encoder+ attention+ decoder, which can effectively improve the segmentation accuracy of spacecraft objects. The network uses dilated convolution encoder to extract deep features. Through the channel attention mechanism, the network calibrates the feature information and innovatively pays attention to the calibration of the low-level features branch. Then, SISnet fuses features through a parallel-ASPP structure to enhance contextual information. Finally, the decoder of the network decodes the fused feature maps. The monocular images are input to the SISnet, which outputs spacecraft mask images.

3.1. Overall Network Architecture

Our network is improved with DeepLabv3+ [13] as the baseline. Although DeepLabv3+ has achieved outstanding results in image segmentation, there are still some problems in the spacecraft image segmentation task, such as blurred object boundary segmentation, incomplete contour, inaccurate segmentation pixels, and image noise affecting the learning effect of the model. In order to solve these problems of DeepLabv3+, we have carried out a series of optimizations. First, in the backbone, we did not choose Xception-65 or ResNet-101, used by DeepLabv3+ for feature extraction. We use the DRN-D-54 [16] as a feature extraction network and achieve better results than Xception-65 or ResNet-101, with only about the same number of layers as ResNet-50. After the deep feature extraction in the backbone, the extracted features are divided into two different branches: high-level features and low-level features. Then, both branches should enter the SEM module of the channel attention mechanism to calibrate the channel feature response adaptively to improve the representation quality generated by the network. Finally, we improve the decoder structure. We add the second ASPP module. The two ASPP modules form a parallel ASPP structure. The high-level features enter the first ASPP module, and the low-level features enter the second ASPP module. In addition, the network still retains the low-level features branch without passing through ASPP module, which is fused together with the high-level and the low-level features passed through the ASPP module in the decoder structure. After a 3×3 convolution, the fused features are up-sampled by four times bilinear interpolation,

and, finally, the predicted spacecraft mask is obtained. We will describe the details of the improvements in the following sections. The overall network structure is shown in Figure 2.

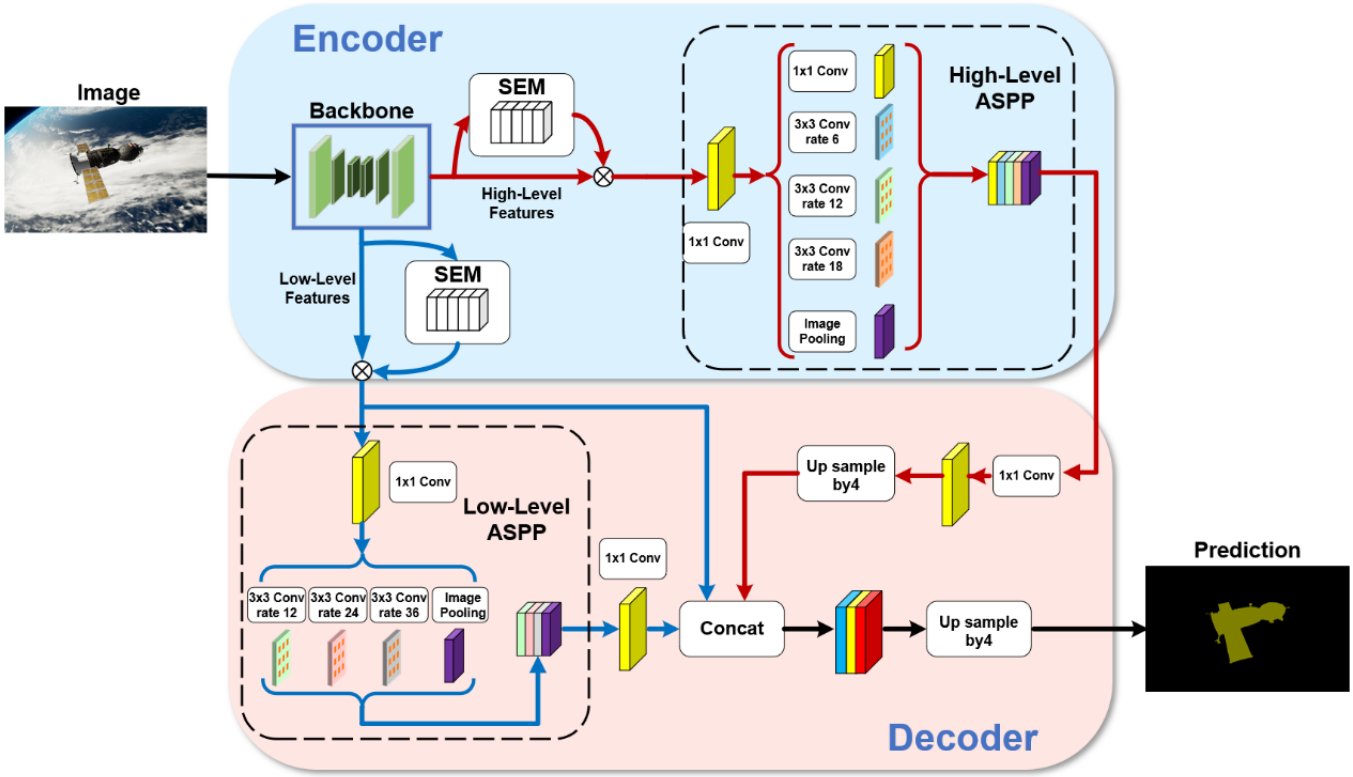

Figure 2. The structure of the semantic segmentation network for spacecraft images.

3.2. Network Backbone

In the backbone, we use an architecture DRN composed of dilated convolution and residual network as the encoder for feature extraction. The residual network achieves excellent performance on image classification tasks through structures such as multiple convolutional layer cascades and residual linking. The residual network achieves excellent performance on image classification tasks through structures such as numerous convolutional layer cascades and residual connections. Compared with the ordinary network, the residual network causes less information loss and a better learning effect and has become the preferred feature extraction tool for various neural networks. However, the residual network inevitably gradually reduces the resolution inside the convolutional network so that some spatial structures are no longer easy to distinguish. While image segmentation is a pixel-level image classification task, the loss of spatial acuity reduces the accuracy of image segmentation. DRN increases the receptive field of higher layers by setting dilated convolution to compensate for the reduction in the receptive field caused by replacing some down-sampling in the residual network. DRN enables the convolutional network to ensure higher resolution without changing the receptive field. Moreover, DRN removes some of the maximum pooling in the residual network that is no longer necessary and the residual connection at the back of the network. It improves the accuracy of image segmentation without changing the depth and complexity of the model.

There are several variants of DRN. Specifically, the structure we use is DRN-D-54. The structure diagram of the network is shown in Figure 3. We use a BatchNorm layer and a ReLU activation layer for each convolutional layer, forming a Conv-BN-ReLU group. DRN-D-54 is designed with three ordinary convolutional layers at the front of the network. In the middle of the network, similar to ResNet, DRN-D-54 has many bottleneck layers, some of which use dilated convolution. The bottleneck structure contains convolution with a kernel size of 1×1, which can change the network dimension more flexibly and reduce

the computation of the network [17]. Depending on whether the number of channels varies, the DRN-D-54 design has two different bottleneck layers. When the number of channels in the bottleneck layer changes, additional convolutional layers are connected at the residual connection. At the back of the network, there is a dilated convolutional layer and a standard convolutional layer. The low-level features have only passed through 12 convolution layers, while the high-level features pass through all convolution layers in the DRN-D-54 network.

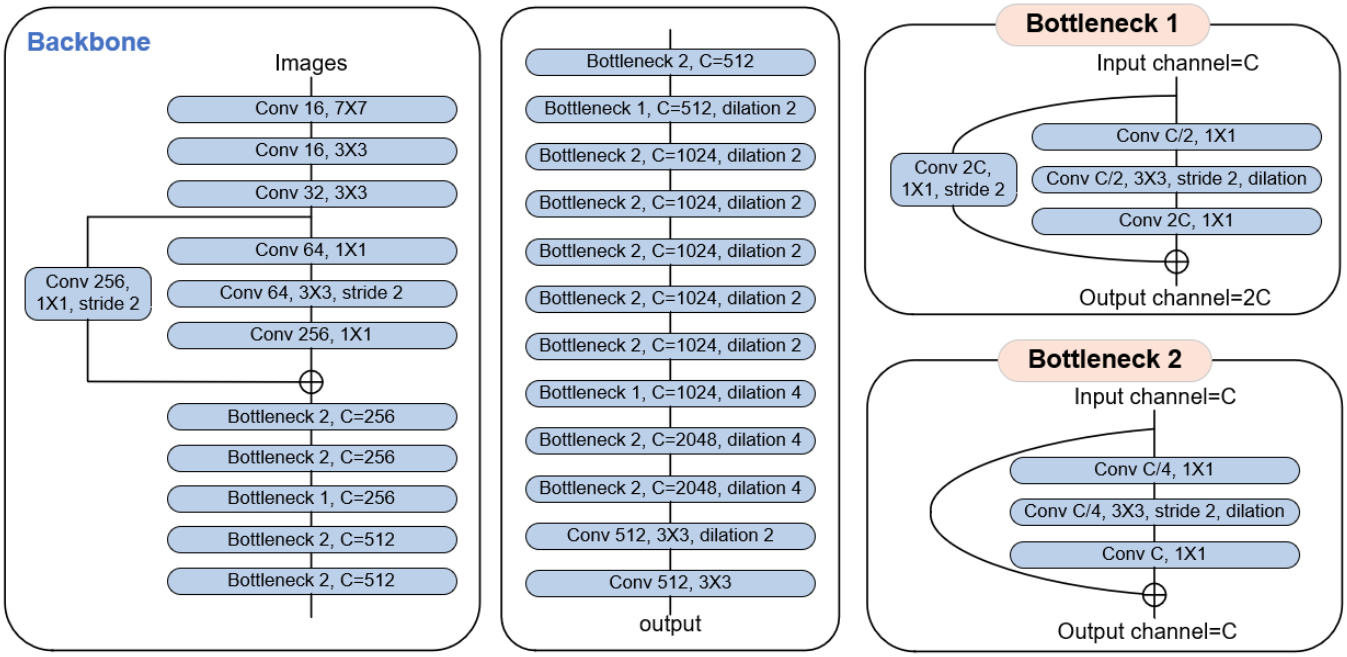

Figure 3. The structure of the network backbone. Each shape represented as a convolutional layer is actually a Conv-BN-ReLU group consisting of a convolutional layer, a BatchNorm layer, and a ReLU activation layer. The stride represents the step of the convolution. Dilation represents the rate adopted by dilated convolution. The letter C denotes the number of channels in the corresponding layer. 1×1, 3×3, and 7×7 are denoted as the kernel size of different convolution kernels.

Dilated convolution improves the resolution of the output feature map without reducing the receptive field of individual convolution kernels, which has been proved to improve the segmentation performance in many segmentation networks. However, not all the convolutions can be replaced by dilated convolutions to achieve good results because dilated convolution may lead to gridding artifacts. As shown in Figure 4, assuming that the original image has only a point pixel, there are nine discrete pixel blocks in the feature map after dilated convolution. This gridding artifacts phenomenon will make the feature map relatively rough, showing a fine point-like distribution so that the segmentation results are not fine enough. Especially in the case of noise in the image, it will greatly affect the final segmentation result. Because there is a large amount of dilated convolution in Xception-65, there is noise in our data. In experiments, Xception-65 is not easier to converge, and it performs worse segmentation than many shallower networks or networks without dilated convolution. In order to remove the negative effects of these gridding artifacts, the degridding operation must be performed in the network. In ResNet, maxpooling at the front end of the network leads to output high-frequency high-amplitude activation values, and these high-amplitude activations are then easily propagated down by the later convolution and form the gridding artifacts after the dilated convolution. Therefore, DRN-D-54 removes it at the front end of the network. To avoid the residual connections that superimpose the gridding artifacts of the previous layer onto the next layer, the structure of DRN-D-54 last two layers is relatively simple. Compared to ResNet, we remove the residual connections from the last two layers and set a low dilation rate to help reduce the gridding artifacts. Therefore, the penultimate dilation rate is set smaller than the previous

dilation rate [16]. Performing multiple dilated convolutions in succession may aggravate the gridding artifacts. Because the feature map after feature extraction of the backbone has to go into the attention and decoder structures, we do not use the dilation convolution in the last layer. These can make our backbone have a smoother feature map and affect our segmentation results. The experimental part verified that DRN-D-54 as a backbone achieved better segmentation results with fewer layers than Xception-65 and ResNet-101.

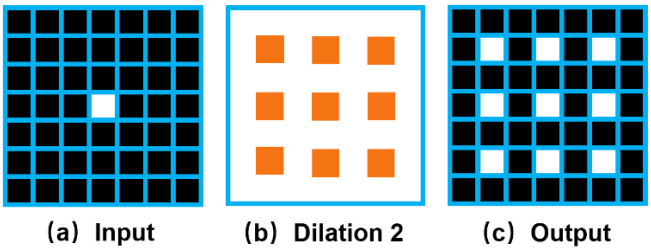

Figure 4. The gridding artifacts are caused by dilated convolution. (**a**) represents the input of a point pixel. (**b**) represents a layer of dilated convolution. (**c**) represents the output after dilated convolution. A single point pixel is mapped into multiple.

3.3. Squeeze-and-Excitation Attention Module

To enhance the effect of feature extraction, we add an attention mechanism in the network. Inspired by SENet [20], we designed the Squeeze-and-Excitation Module (SEM) in the network. SEM uses channel attention to enable the network to obtain the importance of each feature channel by learning automatically. Its schematic diagram is shown in Figure 5. We calibrate the high-level and the low-level features through SEM and then enter the decoder structure at the back of the network. We verify that attention modulation on two branches of different depths has better results than focusing only on the high-level features.

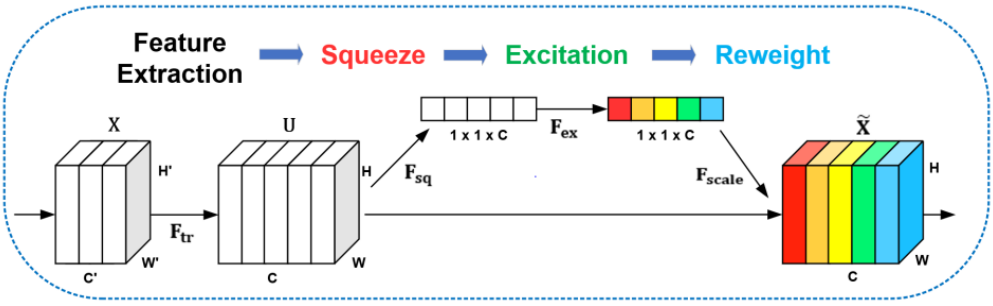

Figure 5. The schematic diagram of SEM.

SEM is a computing unit that can divide into three computing operations: Squeeze, Excitation, and Reweight, as shown in Figure 5. For a given image input $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$, the input is transformed into a feature map $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ through the feature extraction process of $\mathbf{F}_{tr}$. $\mathbf{F}_{tr}$ can represent a convolution operation. $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_C]$ is represented as the set of convolution kernels. $\mathbf{v}_C$ represents the parameter of the c-th kernel. $\mathbf{x}^s$ denotes the s-th output. $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_C]$ as the output of this process. The transformation of the feature map can be expressed by the following formula:

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s \tag{1}$$

Here, $*$ represents the convolution operation. $\mathbf{v}_c = [\mathbf{v}_c^1, \mathbf{v}_c^2, \ldots, \mathbf{v}_c^{C'}], \mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{C'}]$ and $\mathbf{u}_c = \mathbb{R}^{H \times W}$. $\mathbf{v}_c^s$ represents a single channel in $\mathbf{v}_c$, corresponding to the input $\mathbf{X}$. After the feature map $\mathbf{U}$ is generated, the network officially enters the SEM. The first is the Squeeze operation. SEM uses global average pooling to compress the feature graph $\mathbf{U}$ along the spatial dimension $H \times W$, turning each two-dimensional feature channel into a real number.

In this operation, the input of $H \times W \times C$ is transformed into the output $\mathbf{z} \in \mathbb{R}^C$ of $1 \times 1 \times C$ through the process of $\mathbf{F}_{sq}$. $\mathbf{z}$ is the set of real numbers corresponding to channel number C. z_c represents the c-th element of $\mathbf{z}$, and its calculation formula is as follows:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathrm{u}_c(i, j) \tag{2}$$

This real number can be regarded as having a global receptive field, and the dimension of the output matches the number of feature channels of the input. It characterizes the global distribution of the response over the feature channels and makes the global receptive field available for the layers close to the input. Next is the excitation operation. SEM uses two fully connected layers to fuse the feature map information of each channel and then uses the ReLU function and the sigmoid function so that the network can learn the dependencies between channels through end-to-end training. That enables full capture of channel-wise dependencies. SEM takes the result $\mathbf{z}$ obtained by the Squeeze operation as input and transforms it into s through the process of F_{ex}. The formula is as follows:

$$\mathbf{s} = F_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{3}$$

Here, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ represent two fully connected layers. The dimension of $\mathbf{z}$ is $1 \times 1 \times C$ and becomes $1 \times 1 \times \frac{C}{r}$ after the first full connection layer $\mathbf{W}_1$. r denotes the scaling parameter of the fully connected layer, which can reduce the amount of computation. δ denotes the ReLU function and does not change the dimension of the output. After the second fully connected layer $\mathbf{W}_2$, the output dimension becomes $1 \times 1 \times C$. Finally, the output $\mathbf{s}$ is obtained by the sigmoid function. $\mathbf{s}$ is the weight of feature maps in U, which is obtained by learning the fully connected layer and nonlinear layer. Finally, there is the reweight operation. The network considers the weight $\mathbf{s}$ of the excitation output as the importance of each feature channel after feature selection and then weights it to the previous feature map $\mathbf{U}$ by a channel-wise multiplication operation to obtain the output $\tilde{\mathbf{X}} = \left[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_C\right]$. The weight completes the recalibration of the original features in the channel dimension. For the c-th $\tilde{\mathbf{x}}_c$ element in $\tilde{\mathbf{X}}$, SEM obtained the final $\tilde{\mathbf{x}}_c$ by multiplying the two-dimensional matrix $\mathbf{u}_c$ with each of the corresponding weights s_c through the process of $\mathbf{F}_{scale}$, as shown in the following equation.

$$\tilde{\mathbf{x}}_{\boldsymbol{c}} = \mathbf{F}_{scale}(\mathbf{u}_{\boldsymbol{c}}, \mathbf{s}_{\boldsymbol{c}}) = \mathbf{s}_{\boldsymbol{c}} \mathbf{u}_{\boldsymbol{c}} \tag{4}$$

We added SEM between the DRN and the parallel-ASPP structure. In the Squeeze stage, the SEM module transforms the dimension of the input feature map from $H \times W \times C$ to $1 \times 1 \times C$ through global pooling. From the perspective of image resolution, both the high-level and the low-level features will be processed to the same size. Therefore, even with lower resolution feature maps, the SEM module can learn the channel-wise weights. In terms of module structure, SEM and DRN constitute the SE-DRN module, which guarantees the quality of feature extraction from the input image $\mathbf{X}$ to the output feature map $\tilde{\mathbf{X}}$, as shown in Figure 6. The overall calculation of SEM is very small, and the input dimension is not changed. The channel attention mechanism enhances the useful feature channels, weakens the redundant feature channels, and can significantly improve the accuracy of noisy data. The SEM enables our network to have better adaptability to strong noise and highly redundant space data.

3.4. Parallel-ASPP

To further capture and fuse image features at different scales, high-level and low-level features go into the parallel-ASPP structure we designed. The multiple dilation rates set by ASPP can help the network capture multi-scale contextual information, proven effective

in both DeepLabv3 [45] and DeepLabv3+ [13]. However, DeepLabv3+ still leads to the problem of blurred segmentation boundaries. Inspired by Bai et al. [19], we added a second ASPP module to the decoder structure to further capture the multiscale contextual information of low-level features. Moreover, inspired by separable convolution [54] and Zhang et al. [18], we similarly change ASPP to factorized ASPP. The feature map goes through a separable convolution with a kernel size of 1×1, then into a parallel dilated convolution structure with different rates. This separated structure can be more convenient for us to adjust the number of channels of the network and can reduce the amount of calculation. In our SISnet, instead of duplicating the ASPP module, we put the high-level and low-level features into two different ASPP modules, as shown in Figure 7 below.

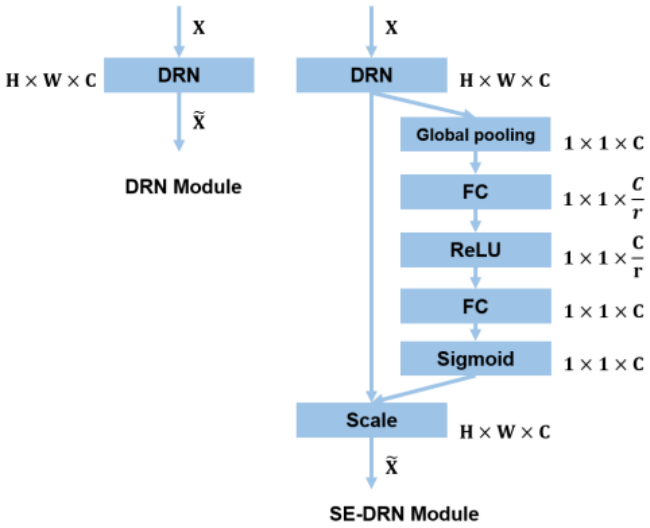

Figure 6. The dimensional change graph of the original DRN (**left**) and the SE-DRN module (**right**).

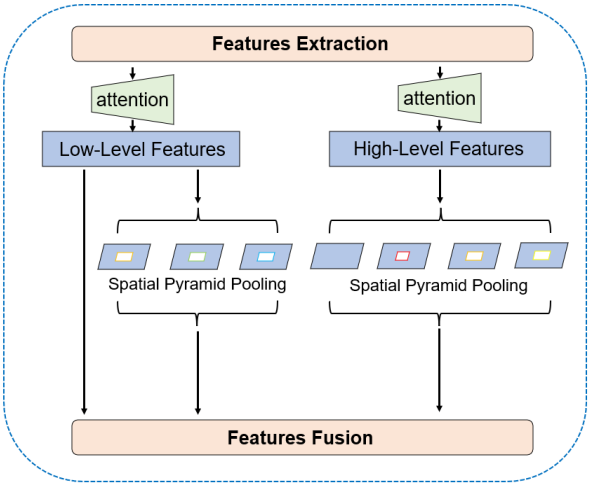

Figure 7. The schema of the parallel-ASPP structure.

The ASPP module entered by the high-level features, after going through a 1×1 separate convolution, goes through a 1×1 standard convolution in parallel, three 3×3 dilated convolutions with dilation rates of [6,12,18], and a global average pooling layer. The module superimposes the parallel feature maps, then connects them by 1×1 convolution, and, finally, the feature maps are enlarged by four times bilinear interpolation up-sampling. In the ASPP module entered by the low-level features, the overall structure remains the same, and only changes are made in the parallel internal layers. It removes the 1×1 standard convolution and expands the dilation rate of the three-layer dilated convolution [12,24,36], which expands the receptive field of low-level features. Moreover, we

retain a low-level features branch that does not enter the ASPP module and use the feature maps from SEM to perform subsequent feature fusion directly. This branch is equivalent to not entering two standard convolutional layers, reducing unnecessary computation. Finally, the feature maps of these three branches are combined for feature fusion. The fused features are fine-tuned by 3×3 convolution and then four times up-sampling using a bilinear interpolation method to predict the spacecraft mask. This asymmetric ASPP structure makes the segmentation boundary more complete and the semantic information clearer. These two ASPP modules form a parallel effect and fuse features together in the subsequent network, so we call it parallel-ASPP.

3.5. Loss Function

The loss function used in the SISnet network is the cross-entropy loss function. The formula is as follows:

$$L = -\sum_{i=1}^{N} y_i log\hat{y}_i + (1 - y_i) log(1 - log\hat{y}_i) \quad (5)$$

Here, L is the training loss, N is the number of samples, y_i is the real sample label, and $\hat{y}_i$ is the prediction label. The smaller the L value, the closer the prediction label is to the real sample label, and the more accurate the network segmentation result is.

4. Experiments

This section evaluates our proposed SISnet on the spacecraft segmentation dataset we labelled. In Section 4.1, we introduce the dataset and annotation details we used. In Section 4.2, we introduce the implementation details of our network and evaluation indicators. In Section 4.3, we compare our proposed method with representative methods in the segmentation domain: U-Net [15], HRNet [55], DeepLabv3+ [13], PSPNet [43], EMANet [53]. In Section 4.4, we conducted a comparative experiment on the backbone and a comparative experiment on the parallel-ASPP structure. In Section 4.5, we conducted an ablation study to demonstrate the effectiveness of our backbone, attention module, and decoder improvements.

4.1. Dataset

Currently, there are very few datasets that can be used for spacecraft image segmentation. We used the photorealistic spacecraft images provided in SPEED [9] and URSO [4], annotated by the LabelMe tool, and obtain fine masks of spacecraft targets. Based on this, we established a spacecraft segmentation dataset. The dataset contains three types of spacecraft, Tango, Dragon, and Soyuz. The appearance of each spacecraft is shown in Figure 8 below. All three spacecraft are morphologically different and representative. Choosing these three types of spacecraft as the dataset for training, rather than a single spacecraft, improves the variety of objects in our dataset and the difficulty of network learning while allowing us to demonstrate better generalization of our network.

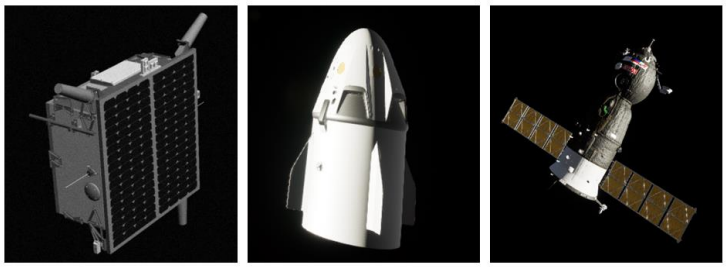

Figure 8. The appearance diagram of three types of spacecraft.

During the labelling process, we marked different types of spacecraft with different colors. In addition, three objects are represented by the category names of Spacecraft A, Spacecraft B, and Spacecraft C. Their categories and color settings are shown in Table 2 below. The dataset consists of 600 monocular images with image sizes of 1280 × 960 and 1920 × 1200. Spacecraft A and Spacecraft B each have 150 images in the dataset, and Spacecraft C has 300 images. Since the morphology of Spacecraft C is more complex and often asymmetrical in the image, a larger amount of data were specially set for Spacecraft C. The effect of image annotation is shown in Figure 9 below. We divided the dataset into the training set, test set, and validation set in a 6:2:2 ratio. Finally, there are 360 training set images, 120 test set images, and 120 validation set images. The ratio of the three object categories in each subset is 1:1:2, which maintains the same data distribution. Among them, the training set and the validation set participate in the training process of the model. The validation set is used to avoid training overfitting and to determine the learning rate during training. We performed mask prediction and evaluation metrics on the test set to evaluate the training results of different models. Meanwhile, we used the resize operation to set the image size to 512 × 512 pixels in the network.

Table 2. Classification number, name, and color table.

Number	Spacecraft	Category Name	RGB Value	Color
1	Tango	Spacecraft A	(128, 0, 0)	
2	Dragon	Spacecraft B	(0, 128, 0)	
3	Soyuz	Spacecraft C	(128, 128, 0)	
4	–	Background	(0, 0, 0)	

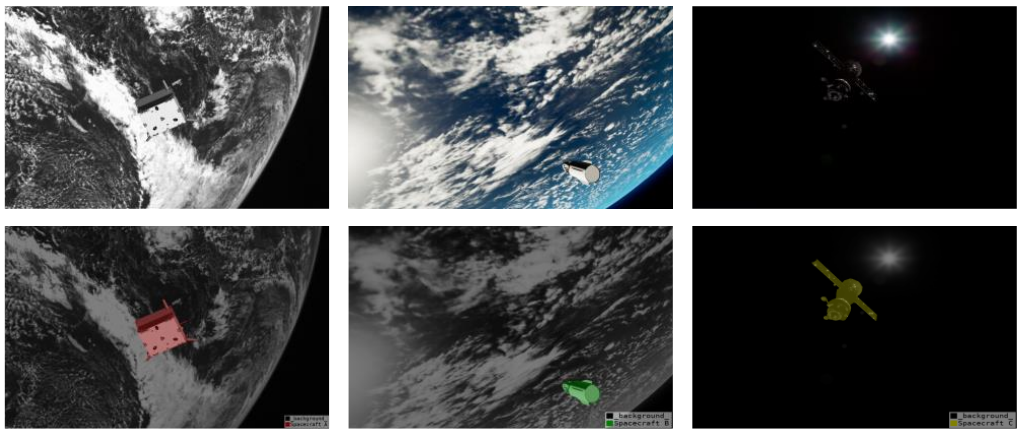

Figure 9. Diagram of the dataset. The first column is the original image, and the second column is the image displayed in an overlapping form after annotation.

In remote sensing datasets, the size of the same type of object does not change much, so many deep learning models are anchor-based networks. The network can converge faster. However, even if it is the same object in the spacecraft picture, the size of the object may vary greatly due to different distances, so the spacecraft object has multi-scale characteristics. In addition, compared with the targets in the KITTI dataset, the spacecraft objects may appear in the image in any pose because of their motion characteristics. If the spacecraft structure is complex, then the changes reflected in the image may be more, which is the multi-attitude characteristics of the spacecraft data.

The background of the spacecraft is space, so the image will inevitably be affected by the illumination generated or reflected by various stars and spacecraft, resulting in large noise. Real data may even be much noisier than synthetic data. Spacecraft A is derived from the SPEED [9]. In order to improve the realism of the synthetic data, all images of Spacecraft A have been subjected to Gaussian blur processing ($\sigma = 1$) and zero mean, Gaussian white

noise processing with ($\sigma^2 = 0.0022$). Due to the lighting conditions, there are often problems of underexposure or overexposure. The reflection on the surface of the spacecraft and the difficulty of identifying the structural parts of the spacecraft due to insufficient light will increase the difficulty of accurate target segmentation. Considering the influence of those mentioned above, multi-scale, multi-attitude, noise, and illumination imbalance, as well as the common complex backgrounds and weak targets in some public datasets, we selected the data when establishing the spacecraft segmentation dataset. Among them, there are 80 small-scale spacecraft images and 81 images of spacecraft in poor lighting conditions. There are 70 images of spacecraft with complex backgrounds. Our dataset contains more hard examples. The proportion of hard examples in the dataset reaches 38%. Some hard examples are shown in Figure 10 below. It can be seen that our spacecraft segmentation task is challenging and hard. Our dataset is a small-sample dataset. However, compared with public datasets commonly used for image segmentation, such as the Pascal VOC 2012, the number of images of our various objects is of the same magnitude. At this order of magnitude, our network can achieve segmentation well.

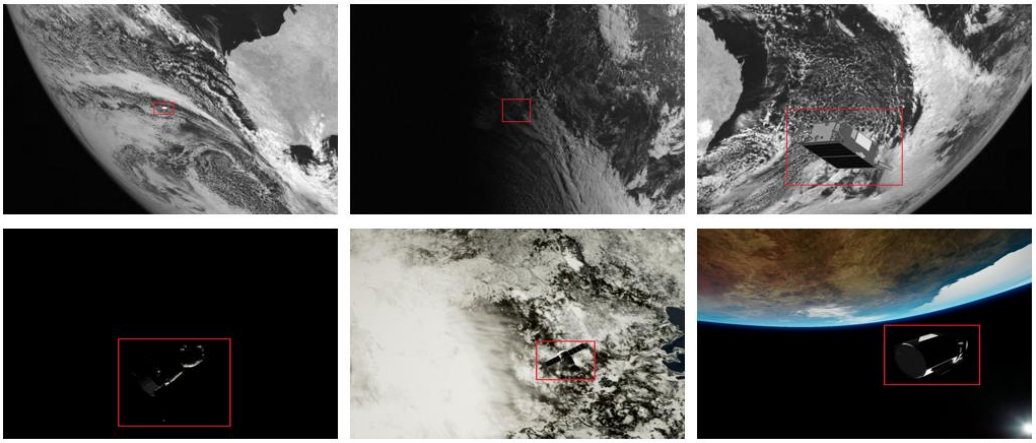

Figure 10. Diagram of hard examples in the dataset. Spacecraft objects are marked with red bounding boxes.

4.2. Implementation Details

The experimental environment is Intel(R)Core^{TM}i9 − 9900K CPU@3.60GHz, running memory 16 G, Ubuntu 18.04, 64-bit operating system. We use CUDA10.1(NVIDIA Corporation, Santa Clara, CA, USA), CuDNN 7.6.5 (NVIDIA Corporation, Santa Clara, CA, USA), python 3.7. We performed network training on an NVIDIA GEFORCE RTX2080Ti GPU. We trained our network with the open-source deep learning framework PyTorch. We used the Adam optimizer to update the weights of the neural network iteratively. The initial learning rate is set to 0.01 for network training. Other training parameters are shown in Table 3 below:

Table 3. Training parameters.

Parameter	Value	Parameter	Value
learning rate	0.01	SEM rate	16
epoch	200	batch size	2
High-ASPP rate	[1,6,12,18]	Low-ASPP rate	[12,24,36]
momentum	0.9	weight decay	0.0005
output stride	8	crop size	512 × 512

The SEM rate represents the scaling parameter of the SEM fully connected layer. The High-ASPP rate and Low-ASPP rate represent the dilation rate of parallel ASPP modules. The output stride represents the output stride of the encoder structure. We maintained consistent and comparable hardware and software parameters in each training and testing

experiment. To focus on comparing network performance and exploring the impact of data on the network from experiments, we did not use data augmentation to expand our dataset further. Figure 11 shows the loss decline of our SISnet and our baseline method during training. It can be seen that, as the number of iterations increases, loss decreases rapidly. In the middle and later stages of training, our network loss fluctuates less and tends to stabilize more quickly.

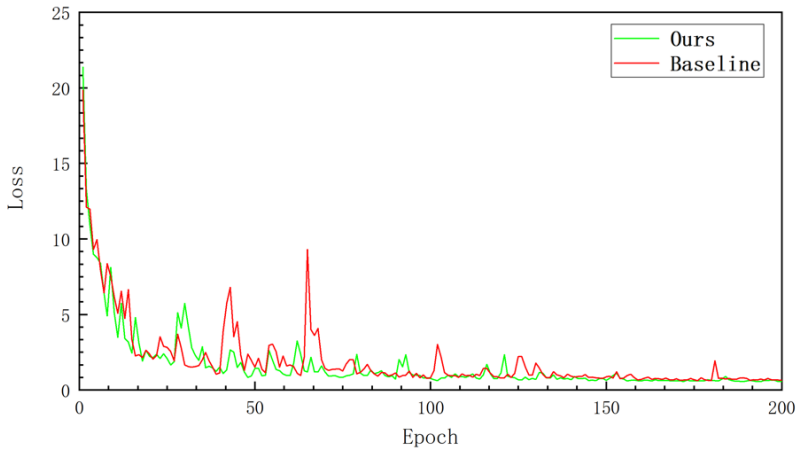

Figure 11. Network training loss curve.

4.3. Network Comparative Experiments

Under the condition of the same training environment and training parameters, we conducted comparative experiments on various semantic segmentation networks. Figure 12 shows the prediction results of the network. From top to bottom are the original image, U-Net, HRNet, DeepLabv3+, EMANet, PSPNet, our proposed method, and the final truth image.

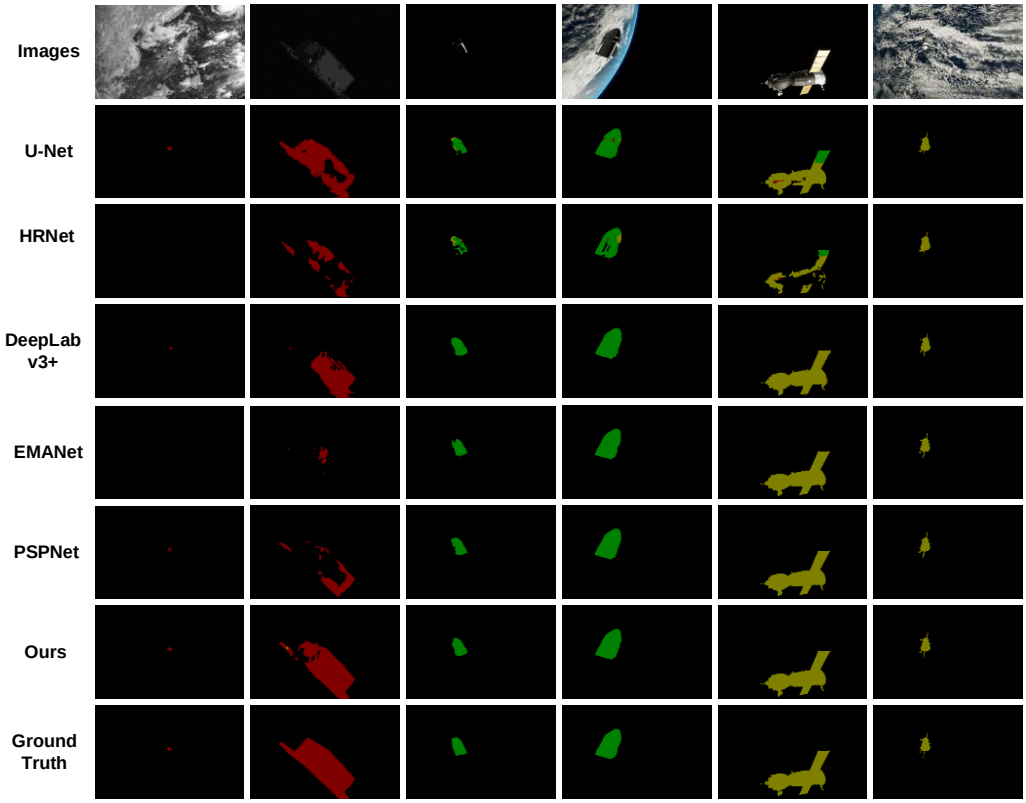

Figure 12. Experimental results and comparison with other methods.

Overall, our network shows better segmentation results on these three types of objects, and the masks are smoother and fuller. To measure the segmentation effect of the network more accurately, we choose mean intersection over union (MIoU) as the evaluation metric and use the test set for quantitative analysis. When the value of MIoU is closer to 1, it indicates that the segmentation result is more accurate. The calculation formula of MIoU is as follows:

$$\mathrm{MIoU} = \frac{1}{k-1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{6}$$

In the above formula, k represents the number of categories, including the background, i represents the true value of the pixel category, and j represents the pixel prediction result. p_{ii} represents the number of correctly classified pixels, and p_{ij} represents the total number of pixels for which i is predicted to be j. p_{ji} represents the total number of pixels for which j is predicted to be i. During training, we set up checkpoints to ensure that the highest MIoU weights are recorded. Meanwhile, our training MIoU can reach more than 94% after 200 epochs of training. We quantitatively evaluated the above networks on the test set, and the results are shown in Table 4 below.

Table 4. The MIoU results of network comparative experiments on the test set.

Method	Spacecraft A	Spacecraft B	Spacecraft C	MIoU
U-Net [15]	64.42	83.51	90.13	79.35
HRNet [55]	40.04	32.59	56.9	43.18
DeepLabv3+ [13]	83.15	81.26	88.43	84.28
PSPNet [43]	83.61	81.88	93.84	86.44
EMANet [53]	74.38	89.12	94.18	85.89
Ours	93.11	87.62	92.87	91.20

It can be seen in Table 4 that EMANet uses ResNet-101 as the backbone and uses the expectation-maximization attention module to effectively improve the segmentation results. The combination of a deeper feature extraction network and the attention mechanism makes EMANet have better results on the test set for Spacecraft B and Spacecraft C. However, for Spacecraft A, which is noisy and has less color information, the result is not so good. Compared to U-Net and HRNet, our MIoU is higher because we have a deeper network. Compared with our baseline method, DeepLabv3+, we have greatly improved the IoU of all classes.

4.4. Structure Comparative Experiments

In this paper, we also experimented with the different backbone networks of the encoder. While keeping the baseline as DeepLabv3+, we replaced different backbones to compare the effect of different feature extraction networks in the spacecraft segmentation issue. We conducted experiments with Mobilenet-v2 [56], ResNet-50 [17], ResNet-101 [17], Xception-65 [13], Densenet121 [57], and DRN-D-54 [16]. The experimental results are shown in Table 5 below. Although Xception-65 performs well in some image segmentation tasks, its applicability to spacecraft objects is relatively poor due to noise and the gridding artifacts. Mobilenet-v2 is also a backbone using depthwise separation convolutions and achieved good results. ResNet-50 [D] and ResNet-101 [D] represent the dilated convolution versions of ResNet. ResNet-50 [D] and ResNet-101 [D] replace a small number of standard convolutional layers in the network with dilated convolutions with the dilation rate = 2 or 4. They increase the receptive field but do not cause serious gridding artifacts. ResNet-101 [D] achieved higher scores for the segmentation of Spacecraft A with a larger field of perception and deeper network depth. Meanwhile, because DRN-D-54 removes some residual connections, the focus on low-level features is not as good as ResNet-101 [D], so the segmentation effect of the network for Spacecraft A can be further improved when the decoder structure focuses on the low-level features. From the MIoU metric, ResNet-50 [D]

and ResNet-101 [D] show that the network adds dilation convolution and demonstrate that dilation convolution is still a boost for the spacecraft segmentation task. To achieve higher segmentation results, we have to pay attention to maintaining the depth of the network and enhancing the contextual information while avoiding the gridding artifacts. Therefore, we adopted the encoder structure based on DRN-D-54. Experimental results show that our encoder has a good segmentation effect on all objects while achieving the highest MIoU.

Table 5. The MIoU results of backbone comparative experiments on the test set.

Backbone	Spacecraft A	Spacecraft B	Spacecraft C	MIoU
Mobilenet-v2 [56]	74.02	84.29	91.78	83.37
ResNet-50 [17]	81.01	77.1	90.09	82.74
ResNet-50 [D]	78.17	82.74	88.42	83.11
ResNet-101 [17]	77.64	82.71	88.7	83.02
ResNet-101 [D]	83.15	81.26	88.43	84.28
Xception-65 [13]	67.61	72.5	88.34	76.15
Densenet121 [57]	51.62	64.97	83.38	66.66
DRN-D-54 [16]	78.56	86.64	91.46	85.55

Based on the above encoder experiments, while keeping the backbone network as DRN-D-54, we conducted experiments of adding a second ASPP module to the decoder structure. We let the low-level features in the baseline method directly enter the second ASPP module and then fuse with the high-level features passed through the ASPP module. The low-level features further aggregate contextual information through this process. During the experiments, we set different dilation rates for the second ASPP module. We achieved better segmentation results with fewer layers and larger dilation rates, as shown in Table 6 below. In the second experiment, we added a second ASPP structure that is identical to the first ASPP module. The segmentation results have been improved compared to the first experiment without adding the second ASPP module. However, in the third experiment, we removed the convolutional layer with the rate of 18, and the effect decreased instead because the low-level features require a larger receptive field to help improve the overall network effect. Therefore, we adjusted the convolutional layers in the parallel structure, as shown in the fourth experiment. We increased the dilation rate of convolution with all three layers only, which further improved the segmentation results. In the design of the final decoder, we also directly fused the low-level features through the channel attention with the features through the parallel ASPP structure to ensure the effect of attention modulation. We will show the experimental results in the ablation study.

Table 6. The MIoU results of the second ASPP module comparative experiments on the test set.

Number	Low-ASPP Rate	Spacecraft A	Spacecraft B	Spacecraft C	MIoU
1	-	78.56	86.64	91.46	85.55
2	[1,6,12,18]	80.88	86.77	91.98	86.54
3	[1,6,12]	70.31	87.77	86.66	81.58
4	[12,24,36]	85.66	87.25	91.02	87.98

4.5. Ablation Study

This section evaluates different network variables and analyses the reasons that affect network performance. We used DeepLabv3+ as the baseline, and the backbone of the baseline uses ResNet-101 with dilated convolution. We conducted improvement experiments with different network structures. (1) Baseline + A: indicates that the SEM attention module is embedded into the baseline method. (2) Baseline + B: indicates that DRN-D-54 replaces the backbone of the baseline method. (3) Baseline + D: indicates that our improved decoder structure is used. We first conducted experiments with single-variable improvement. Then, we conducted three groups: Baseline + A + B, Baseline + A + D, and Baseline + B + D.

Finally, the experiment of our proposed SISnet method, namely Baseline + A + B + D, is conducted. We performed the quantitative evaluation on the test set for each experiment, and the experimental results are shown in Table 7 below.

Table 7. The MIoU results of the ablation study on the test set.

Method	Spacecraft A	Spacecraft B	Spacecraft C	MIoU
Baseline	83.15	81.26	88.43	84.28
Baseline + A	84.87	82.74	88.11	85.24
Baseline + B	78.56	86.64	91.46	85.55
Baseline + D	91.87	83.83	90.61	88.77
Baseline + A + B	84.6	86.07	88.96	86.54
Baseline + A + D	93.25	84.15	91.5	89.63
Baseline + B + D	93.28	87.2	92.56	91.01
Baseline + A + B + D	93.11	87.62	92.87	91.20

As can be seen from Table 6, the segmentation results of our method compared to the baseline have been steadily improved. The DRN-D-54 backbone we used is optimized for the gridding artifacts and has a strong feature extraction ability. It can achieve better results than ResNet-101 with fewer layers. The channel attention mechanism we designed in the network can also improve the metric by about 1% by assigning channel weights. By enhancing the contextual information of low-level features, our decoder structure significantly improves the segmentation results of Spacecraft A, which is worth significantly improving the final MIoU. The ablation study validates the improvements we made. These improvements can improve the accuracy of spacecraft semantic segmentation to varying degrees. In terms of MIoU metrics, our approach achieves the best segmentation results.

5. Discussion

We also found some problems in our research, briefly explained here. Due to the small test set of our dataset, if some image segmentation results are not good, it may cause relatively large changes in the MIoU value. Whether they are real data or realistic synthetic data, we must pay attention to the problem of noise in the network design because, in real tasks, noise is unavoidable. Our Spacecraft A is more regular in geometric structure and has salience in the image. The segmentation results of Spacecraft A are significantly improved when the effect of noise in the image is addressed. We show some failed cases in Figure 13 below. The second row in the figure is the segmentation result of the baseline method. The third row is the segmentation result of our method.

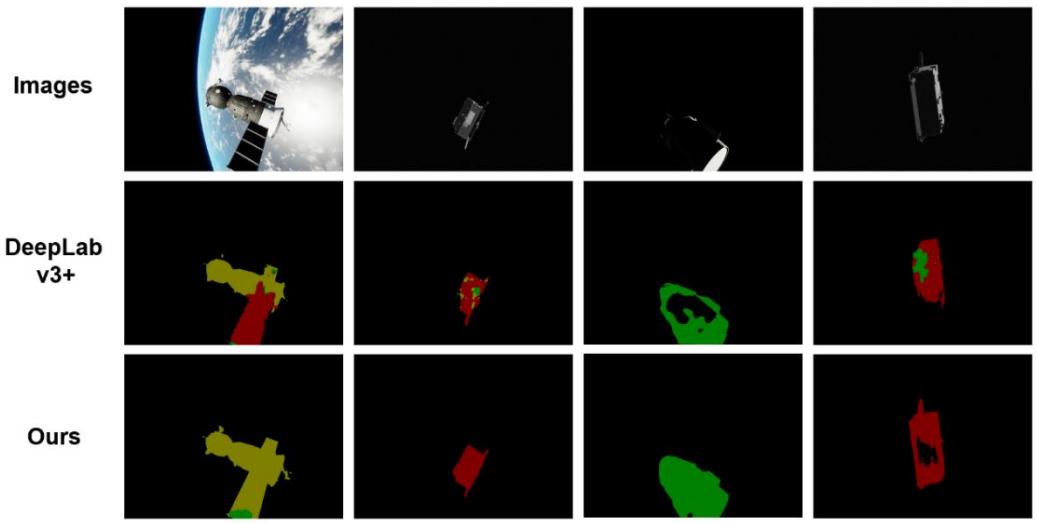

Figure 13. Failure cases.

The segmentation errors in the first column are caused by the similar structures of different types of spacecraft. As shown in Figure 14 below, the Spacecraft A and Spacecraft C objects have similar solar panel structures, and, in some perspectives of Spacecraft A, only the solar panels are more conspicuous. Similar structures are prone to the misclassification of pixels. There is also some degree of pixel clutter in the second column. It is related to the texture features similar to the spacecraft. Our network enhances the contextual information to alleviate such phenomena. The third and fourth columns of images are poor lighting conditions that make some structures invisible and make it easier for pixels to be misclassified as other spacecraft or space backgrounds. The segmentation results of DeepLabv3+ will have large cavities or misclassification. Our method can better segment a simply connected domain, but there are still cavity cases, and the segmentation effect can continue to improve in the future.

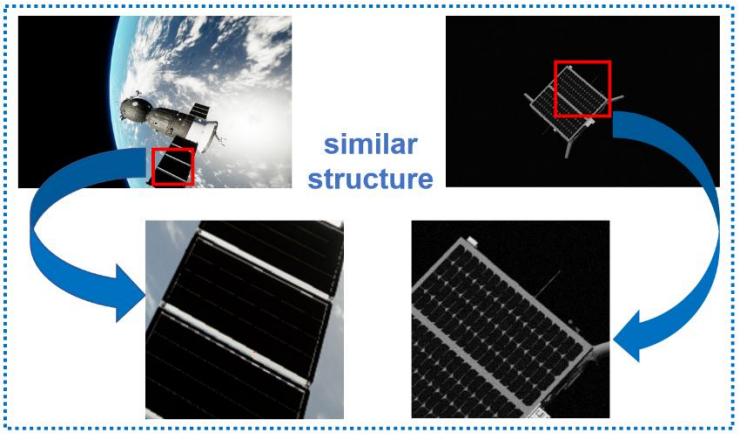

Figure 14. Comparison diagram of similar structures of different spacecraft.

6. Conclusions

In this paper, an end-to-end spacecraft image segmentation network was proposed. Different types of spacecraft objects were segmented with monocular images as inputs to obtain their corresponding masks, and good segmentation results were achieved. We designed a neural network based on a dilated convolution and encoder+ attention+ decoder structure using the self-supervised learning method. Through the dilated convolutional and parallel-ASPP structure, our network enhanced the contextual information and mitigated the effects of gridding artifacts. In addition, the channel attention mechanism was introduced to recalibrate the features and improve the learning ability of the network. Our method is more robust to noise in the image and can segment complete and smooth object masks. We finely labelled public spacecraft datasets to establish a spacecraft segmentation dataset. We conducted various comparison experiments and an ablation study on the dataset. The experimental results show that our method outperforms other methods and has a better segmentation performance for spacecraft objects. In the future, we hope to improve the effect of spacecraft segmentation by the feature matching method in small-sample segmentation technology. At the same time, we would like to experiment with the deep unsupervised active learning strategy [58]. This allows the network to continuously acquire new knowledge for learning during the test phase. We also hope to build more complex space object segmentation datasets and evaluate our network more comprehensively.

Our work will contribute to the research on on-orbit assembly in space. It is helpful for the space manipulator to move quickly to the vicinity of the spacecraft through the image, which provides a basis for further target detection and docking missions.

Author Contributions: Y.L. and M.Z. designed this study; Y.L. and J.W. (Jing Wang) contributed to the theory research; Y.L., X.G. and Y.Y. performed the experiments; Y.L. and J.W. (Jiarong Wang) analyzed the data; Y.L. and X.G. wrote the paper and created the diagrams; Y.L. and M.Z. contributed to scientific advising and proofreading. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Department of Jilin Province, China: 20200401123GX.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SPEED dataset and the URSO dataset are made publicly available for research purposes. For more information, please refer to the websites https://kelvins.esa.int/satellite-pose-estimation-challenge/data/ (accessed on 4 May 2021) and https://zenodo.org/record/3279632 (accessed on 21 June 2021). For the data license, please refer to the websites https://creativecommons.org/licenses/by-nc-sa/3.0/legalcode (accessed on 20 May 2022) and https://creativecommons.org/licenses/by/4.0/legalcode (accessed on 20 May 2022).

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments.

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

References

1. Uriot, T.; Izzo, D.; Simes, L.F.; Abay, R.; Einecke, N.; Rebhan, S.; Martinez-Heras, J.; Letizia, F.; Siminski, J.; Merz, K. Spacecraft collision avoidance challenge: Design and results of a machine learning competition. *Astrodynamics* **2020**, *6*, 121–140. [CrossRef]
2. Carruba, V.; Aljbaae, S.; Domingos, R.C.; Lucchini, A.; Furlaneto, P. Machine learning classification of new asteroid families members. *Mon. Not. R. Astron. Soc.* **2020**, *496*, 540–549. [CrossRef]
3. Reed, B.B.; Smith, R.C.; Bo, J.N.; Pellegrino, J.F.; Bacon, C. The Restore-L Servicing Mission. In Proceedings of the AIAA SPACE 2016, Long Beach, CA, USA, 13–16 September 2016.
4. Proenca, P.F.; Gao, Y. Deep Learning for Spacecraft Pose Estimation from Photorealistic Rendering. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–15 June 2020; pp. 6007–6013.
5. Phisannupawong, T.; Kamsing, P.; Torteeka, P.; Yooyen, S. Vision-based attitude estimation for spacecraft docking operation through deep learning algorithm. In Proceedings of the International Conference on Advanced Communication Technology, Chuncheon City, Korea, 16–19 February 2020.
6. Forshaw, J.L.; Aglietti, G.S.; Navarathinam, N.; Kadhem, H.; Salmon, T.; Pisseloup, A.; Joffre, E.; Chabot, T.; Retat, I.; Axthelm, R.; et al. RemoveDEBRIS: An in-orbit active debris removal demonstration mission. *Acta Astronaut.* **2016**, *127*, 448–463. [CrossRef]
7. Dung, H.A.; Chen, B.; Chin, T.J.; Soc, I.C. A Spacecraft Dataset for Detection, Segmentation and Parts Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 2012–2019.
8. Opromolla, R.; Fasano, G.; Rufino, G.; Grassi, M. A review of cooperative and uncooperative spacecraft pose determination techniques for close-proximity operations. *Prog. Aerosp. Sci.* **2017**, *93*, 53–72. [CrossRef]
9. Kisantal, M.; Sharma, S.; Park, T.H.; Izzo, D.; Martens, M.; Damico, S. Satellite Pose Estimation Challenge: Dataset, Competition Design, and Results. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *56*, 4083–4098. [CrossRef]
10. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
11. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
12. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 3213–3223.
13. Chen, L.C.E.; Zhu, Y.K.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
14. Yang, M.K.; Yu, K.; Zhang, C.; Li, Z.W.; Yang, K.Y. DenseASPP for Semantic Segmentation in Street Scenes. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.

15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
16. Yu, F.; Koltun, V.; Funkhouser, T. Dilated Residual Networks. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 636–644.
17. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
18. Zhang, Z.P.; Zhang, K.P. FarSee-Net: Real-Time Semantic Segmentation by Efficient Multi-scale Context Aggregation and Feature Space Super-resolution. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–15 June 2020; pp. 8411–8417.
19. Bai, R.F.; Jiang, S.; Sun, H.J.; Yang, Y.F.; Li, G.J. Deep Neural Network-Based Semantic Segmentation of Microvascular Decompression Images. *Sensors* **2021**, *21*, 1167. [CrossRef]
20. Hu, J.; Shen, L.; Sun, G.; IEEE. Squeeze-and-Excitation Networks. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
21. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
22. Sun, Q.; Niu, Z.; Wang, W.; Li, H.; Lin, X. An Adaptive Real-Time Detection Algorithm for Dim and Small Photoelectric GSO Debris. *Sensors* **2019**, *19*, 4026. [CrossRef]
23. Schildknecht, T.; Ploner, M.; Hugentobler, U. The search for debris in GEO. *Adv. Space Res.* **2001**, *28*, 1291–1299. [CrossRef]
24. Castellani, L.T.; Llorente, J.S.; Ibarz, J.M.F.; Ruiz, M.; Mestreau-Garreau, A.; Cropp, A.; Santovincenzo, A. PROBA-3 mission. *Int. J. Space Sci. Eng.* **2013**, *1*, 349–366. [CrossRef]
25. Khan, R.; Eichmann, T.; Buttsworth, D.; Upcroft, B. Image-based visual servoing for the super-orbital re-entry of Hayabusa spacecraft. In Proceedings of the 2011 Australasian Conference on Robotics and Automation (ACRA 2011), Melbourne, Australia, 7–9 December 2011; pp. 1–10.
26. Sharma, S.; D'Amico, S. Comparative assessment of techniques for initial pose estimation using monocular vision. *Acta Astronaut.* **2016**, *123*, 435–445. [CrossRef]
27. D'Errico, M. *Distributed Space Missions for Earth System Monitoring*; Springer: New York, NY, USA, 2013. [CrossRef]
28. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-Driven 6D Object Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
29. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G.R. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, 6–13 November 2011.
30. Lowe, D. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **2003**, *20*, 91–110.
31. Herbert, B.; Andreas, E.; Tinne, T.; van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359.
32. Harris, C.G.; Stephens, M.J. A combined corner and edge detector. In Proceedings of the Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988.
33. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698. [CrossRef]
34. Ballard, D.H. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognit.* **1981**, *13*, 111–122. [CrossRef]
35. Sumant, S.; Jacopo, V.; Simone, D. Robust Model-Based Monocular Pose Initialization for Noncooperative Spacecraft Rendezvous. *J. Spacecr. Rocket.* **2018**, *55*, 1–16.
36. Sharma, S.; Beierle, C.; D'Amico, S.; IEEE. Pose Estimation for Non-Cooperative Spacecraft Rendezvous Using Convolutional Neural Networks. In Proceedings of the IEEE Aerospace Conference, Big Sky, MT, USA, 3–10 March 2018.
37. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV, USA, 3–8 December 2012.
38. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
39. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
40. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
41. Zhou, B.; Hang, Z.; Fernandez, F.X.P.; Fidler, S.; Torralba, A. Scene parsing through ADE20K dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
42. Liu, Z.W.; Li, X.X.; Luo, P.; Loy, C.C.; Tang, X.O.; IEEE. Semantic Image Segmentation via Deep Parsing Network. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1377–1385.
43. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

44. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
45. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
46. Visin, F.; Ciccone, M.; Romero, A.; Kastner, K.; Cho, K.; Bengio, Y.; Matteucci, M.; Courville, A. ReSeg: A Recurrent Neural Network-Based Model for Semantic Segmentation. In Proceedings of the Computer Vision & Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016.
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
48. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
49. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. *CBAM: Convolutional Block Attention Module*; Springer: Cham, Switzerland, 2018.
50. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. PSANet: Point-wise Spatial Attention Network for Scene Parsing. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
51. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2020.
52. Li, X.; Wang, W.H.; Hu, X.L.; Yang, J.; Soc, I.C. Selective Kernel Networks. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
53. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Liu, H. Expectation-Maximization Attention Networks for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
54. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
55. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv* **2019**, arXiv:1902.09212.
56. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *arXiv* **2018**, arXiv:1801.04381.
57. Huang, G.; Liu, Z.; Laurens, V.D.M.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Computer Society, Pittsburgh, PA, USA, 11–13 July 2016.
58. Khaldi, Y.; Benzaoui, A.; Ouahabi, A.; Jacques, S.; Taleb-Ahmed, A. Ear Recognition Based on Deep Unsupervised Active Learning. *IEEE Sens. J.* **2021**, *21*, 20704–20713. [CrossRef]