remote sensing

MDPI

Technical Note

A Category–Pose Jointly Guided ISAR Image Key Part Recognition Network for Space Targets

Qi Yang *, Hongqiang Wang, Lei Fan and Shuangxun Li

College of Electronic Science, National University of Defense Technology, Changsha 410073, China; wanghongqiang@nudt.edu.cn (H.W.); fanlei15@nudt.edu.cn (L.F.); jiangwy_wally@foxmail.com (S.L.)
* Correspondence: yangqi08@nudt.edu.cn; Tel.: +86-173-7573-5001

Abstract

It is a crucial interpretation task in space target perception to identify key parts of space targets through the inverse synthetic aperture radar (ISAR) imaging. Due to the significant variations in the categories and poses of space targets, conventional methods that directly predict identification results exhibit limited accuracy. Hence, we make the first attempt to propose a key part recognition network based on ISAR images, which incorporates the knowledge of space target categories and poses. Specifically, we propose a fine-grained category training paradigm that defines the same functional parts of different space targets as distinct categories. Correspondingly, additional classification heads are employed to predict category and pose, and the predictions are then integrated with ISAR image semantic features through a designed category–pose guidance module to achieve high-precision recognition guided by category and pose knowledge. Qualitative and quantitative evaluations on two types of simulated targets and one type of measured target demonstrate that the proposed method reduces the complexity of the key part recognition task and significantly improves recognition accuracy compared to the existing mainstream methods.

Keywords: inverse synthetic aperture radar imaging; space target key part recognition; category and pose guidance; deep learning

check for updates

Academic Editor: John Trinder

Received: 27 April 2025
Revised: 19 June 2025
Accepted: 25 June 2025
Published: 27 June 2025

Citation: Yang, Q.; Wang, H.; Fan, L.; Li, S. A Category–Pose Jointly Guided ISAR Image Key Part Recognition Network for Space Targets. *Remote Sens.* **2025**, *17*, 2218. https://doi.org/10.3390/rs17132218

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The inverse synthetic aperture radar (ISAR) plays a crucial role in the identification tasks of space targets, enabling high-resolution longitudinal and lateral imaging of moving non-cooperative targets through a stationary radar [1]. With the improvement of radar performance, ISAR imaging can now obtain high-precision structural information of targets, which can support research in identification tasks like classification, attitude estimation, 3D reconstruction, and key part recognition [2–5]. Key part recognition, or ISAR image semantic segmentation, divides the space targets in ISAR images into different key parts, which is useful for scene interpretation and automating recognition.

Semantic segmentation is a hot topic in natural image processing, and in recent years, a large number of deep-learning-based algorithms have been successfully applied to it [6,7]. FCN [8] is a pioneer in deep learning for image segmentation, using convolutional and up-sampling operations to output predicted heatmaps, achieving remarkable segmentation results. U-Net [9] innovatively proposed a multi-scale architecture with skip connections, which has proven effective in various fields, including image segmentation. The introduction of the Transformer further advanced the field. Apple combined CNN with Transformer

to propose a lightweight semantic segmentation network, namely, MobileViT [10]. Microsoft introduced the concept of CNN into the Transformer [11], using shifted windows to compute attention weights, overcoming the size limitation of Vanilla ViT. In recent years, semantic segmentation methods based on large datasets have achieved impressive results. Meta released the Segment Anything Model (SAM) [6], trained on a dataset of over 11 million images, demonstrating outstanding zero-shot performance and achieving results comparable to supervised paradigms. S-Lab proposed a universal model to address all tasks in the segmentation domain [12]. By leveraging task-specific queries and outputs, it supports more than ten different segmentation tasks while significantly reducing computational and parameter overhead.

Researchers have drawn on mature experiences from the field of semantic segmentation and conducted a multitude of ISAR key part recognition studies. A straightforward idea is to apply mainstream semantic segmentation networks to ISAR key part recognition, such as FCN [8], DeepLab [13], U-Net [9], etc. Some studies consider the particularities of ISAR key part recognition when designing methods. Wang et al. used a region-growing algorithm to separate the satellite body and proposed a novel clustering algorithm to isolate the remaining components [14]. Coe et al. achieved fast segmentation of ISAR images based on statistical region merging, while using support vector machines (SVMs) for component classification [15]. Wang et al. used principal component analysis (PCA) to analyze the structure of space targets and employed the concept of style transfer to achieve more precise segmentation details [16]. Li et al. proposed a framework to alleviate the scarcity of radar data while enhancing edges by leveraging the scattering characteristics of targets [17]. Zhu et al. proposed a method that combines binary segmentation with mask matching to reduce the complexity of the task, using mask matching to restore binary segmentation results into segmentation results [18]. Kou et al. utilized the non-local self-attention mechanism to explore the structural symmetry of ISAR targets and proposed a contrast learning approach to address the issue of poor precision in small part recognition [4]. Chen et al. developed a multi-task segmentation network capable of simultaneously performing three tasks, with the network improving segmentation performance through a novel symmetric regularization loss [19]. Zhong et al. incorporated the scattering characteristics of targets into the network design, developing a cross-scale self-attention mechanism to suppress sidelobe interference and noise in features. The scattering characteristics were integrated into semantic features through an auxiliary segmentation head, enabling more precise segmentation [20].

However, mainstream methods still have the following drawbacks: (1) It is unreasonable to simply categorize the same functional parts of different space targets as the same class, for example, directly classifying the antennas of all satellites as “antennas”. The antennas, bodies, and solar panels of different satellites, even if functionally similar, often have geometric differences, which poses a great challenge to the generalization of the model. (2) ISAR images still suffer from low contrast and numerous artifacts, and it is difficult for mainstream algorithms to ensure high accuracy solely through semantic label-based learning.

In this paper, a novel ISAR image key part recognition network is proposed to address the existing challenges. Specifically, we improve the training paradigm for key part recognition by categorizing the same component into different classes according to categories of targets, making the segmentation results more specific. Moreover, the proposed method introduces category and pose knowledge of the space targets during the training process and designs two classification heads to output category and pose predictions, respectively, achieving joint estimation in category and key parts. Furthermore, we utilize a global self-attention-based (i.e., Transformer [21,22]) encoder to obtain multi-scale semantic fea-

tures. To better aggregate the learned category and pose knowledge, a category–pose guidance module (CPGM) is designed to refine the semantic features, resulting in more accurate key part recognition outcomes. We designate the proposed approach as CPGNet. Both qualitative and quantitative experiments on two types of simulated targets and one type of measured target demonstrate the effectiveness of our algorithm compared to mainstream methods.

This paper is organized as follows. In Section 2, the ISAR imaging mechanism is presented briefly, followed by the detailed network architecture, like multi-scale feature extraction blocks, semantic decoding layer, etc. Section 4 concludes this paper.

2. Methods

2.1. ISAR Imaging Mechanism

ISAR imaging is essentially a two-dimensional projection in range and azimuth directions of the three-dimensional scattering information of the space target. The positional relationship between the ISAR and the space target is shown in Figure 1, where ONME and O'XYZ represent the northeast celestial coordinate system and the orbital coordinate system, respectively, and $\vec{LOS}$ represents the line-of-sight vector defined in the O'XYZ.

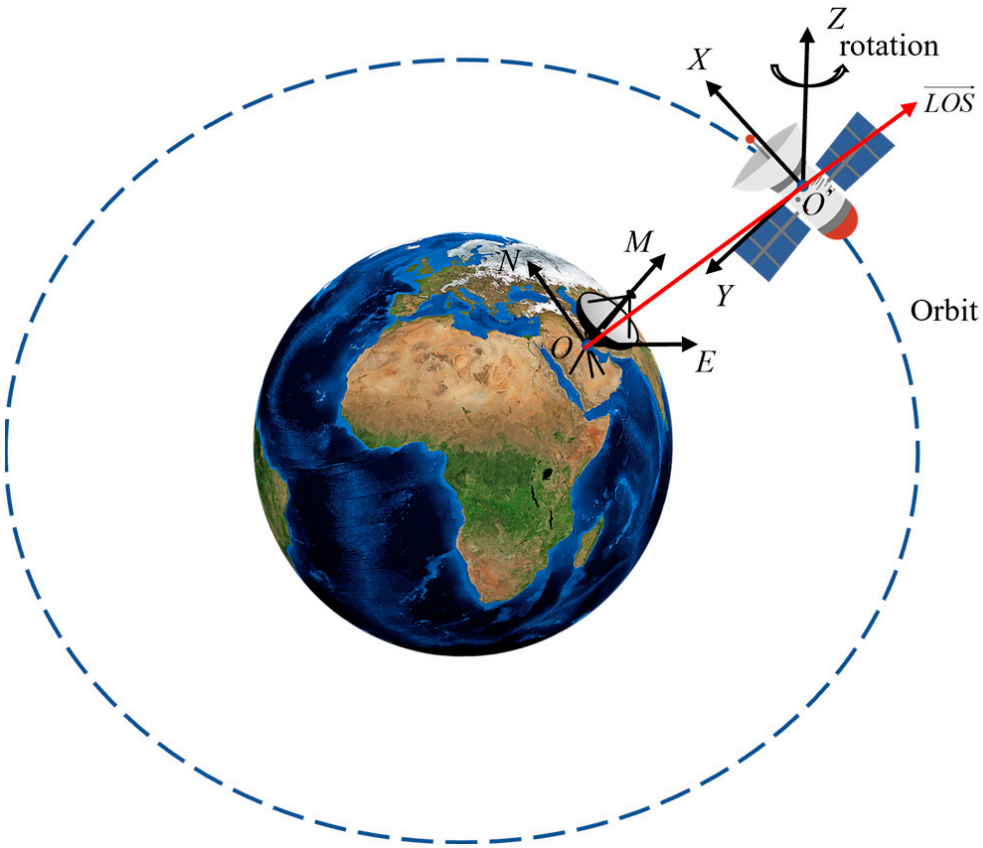


Figure 1. The relative position between the observation radar and the space target.

ISAR imaging is determined by the range dimension projection vector $\vec{k_r}$ and the azimuth dimension projection vector $\vec{k_a}$: $\vec{k_r}$ is equal to the line-of-sight vector at the imaging moment $\vec{LOS}_0$, i.e.,

$$\vec{k_r} = \vec{LOS}_0. \tag{1}$$

Thus, $\vec{k_a}$ is obtained by the cross-product of $\vec{LOS_0}$ and the effective rotation vector $\vec{\omega_e}$, i.e.,

$$\vec{k_a} = \frac{2}{\lambda}(\vec{LOS_0} \times \vec{\omega_e}), \tag{2}$$

where λ means the wavelength, and $\vec{\omega_e}$ is determined by the LOS and target rotation. Following [4], we define $\vec{l_{st}}$, $\vec{l_0}$, and $\vec{l_{end}}$ as the LOS vector at the beginning, middle, and end moments of the coherent processing interval (CPI), respectively. Then, the target rotation vector $\vec{\omega_{LOS}}$ can be written as:

$$\vec{\omega_{LOS}} = \frac{\arccos(\vec{l_{st}}.\vec{l_{end}})}{\text{CPI}} \cdot \frac{\vec{l_{st}} \times \vec{l_{end}}}{\left|\vec{l_{st}} \times \vec{l_{end}}\right|}. \tag{3}$$

For targets with relative rotation, the total rotation vector $\vec{\omega_{syn}}$ can be formed with $\vec{\omega_{LOS}}$ and the target's rotation vector. During the observation, $\vec{\omega_e}$ is only related to the component of $\vec{\omega_{syn}}$ that is orthogonal to $\vec{l_0}$:

$$\vec{\omega_e} = \vec{\omega_{syn}} - \vec{l_0} \cdot \vec{\omega_{syn}} \cdot \vec{l_0^T}. \tag{4}$$

When we get $\vec{k_r}$ and $\vec{k_a}$, each pixel (r, a) in the ISAR image is given by the following expression:

$$\begin{bmatrix} r \\ a \end{bmatrix} = \begin{bmatrix} \vec{k_r} \\ \vec{k_a} \end{bmatrix} \cdot S^T, \tag{5}$$

where S^T denotes the scatter position (x, y, z) of the target.

2.2. Fine-Grained Category Training Paradigm

When sufficient ISAR satellite data are acquired, we can manually designate key parts as the ground truth (GT) and train a key part recognition network under supervised learning. The mainstream methods group functionally similar parts from different satellites into the same class. It is reasonable in natural image semantic segmentation, as different instances share similar morphological features; however, for the recognition of key parts of space targets, geometric variations among satellites force the network to classify dissimilar components as the same class, hindering training and resulting in suboptimal recognition accuracy.

To enhance the network's ability to recognize key parts with high precision, we refine the labels of the dataset and expand the number of prediction classes of the network's output, enabling the network to learn both the structure of the key part and its associated satellite category. The specific configuration of the dataset will be explained in the Experimental Section. This simple yet effective approach reduces training complexity, facilitating high-accuracy recognition, and we show its effectiveness in the Ablation Study Section.

2.3. Network Architecture

We designed a customized network for key part recognition, named CPGNet, and the overall architecture is shown in Figure 2. The network was divided into four modules, which are discussed in detail below.

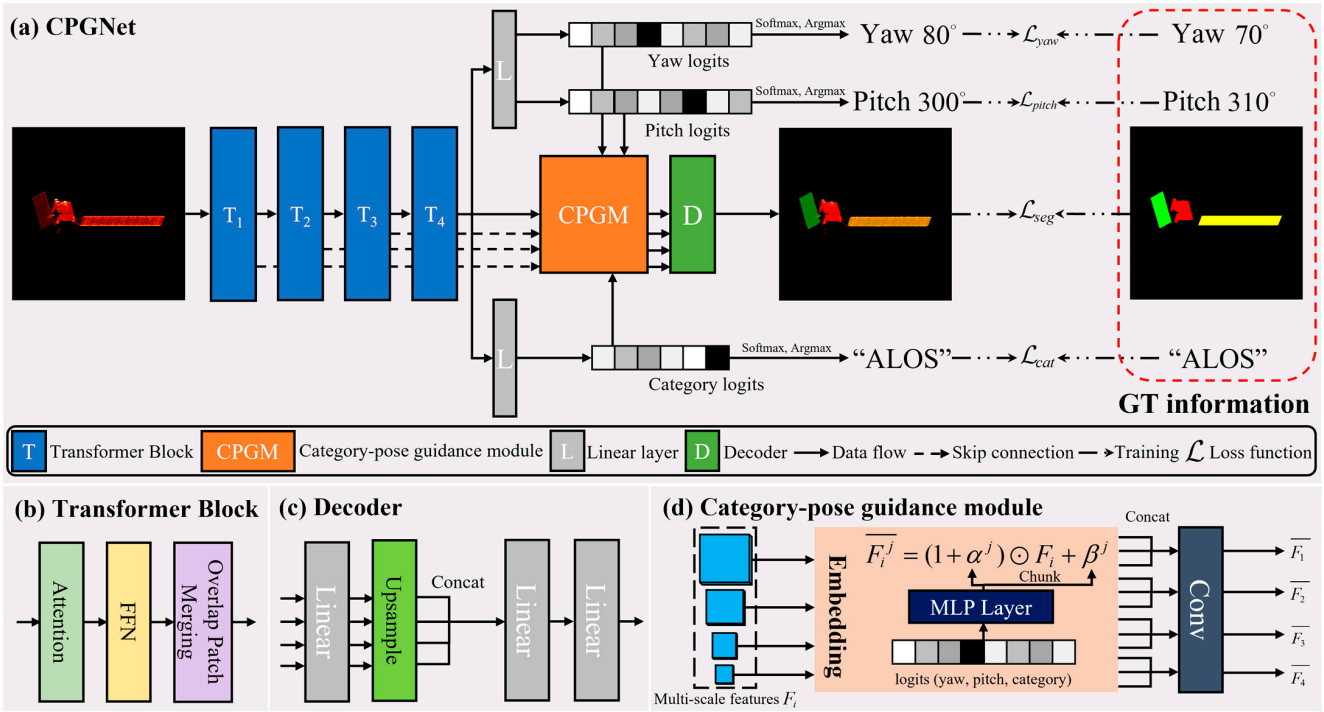


Figure 2. The overall framework of the proposed CPGNet. (**a**) The data flow of the network. (**b**) The structure of the Transformer block, where "FFN" means feed-forward network and "Overlap Patch Merging" is a customized convolutional layer to down-sample the feature. (**c**) The structure of the decoder, where "Up-sample" is implemented based on bilinear interpolation. (**d**) The structure of the category–pose guidance module. Features at each scale are fed into the CPGM for category–pose information embedding, where "MLP Layer" refers to a multi-layer perceptron.

2.3.1. Multi-Scale Feature Extraction Blocks

Traditional CNN-based methods are limited by their local receptive field, which prevents them from obtaining fine-grained segmentation results. Therefore, we utilize a Transformer-based semantic feature extraction backbone to address this issue. The backbone, through its hierarchical design, obtains multi-scale features, F_i, $(i = 1, 2, 3, 4)$, ranging from coarse to fine, which is beneficial for key part recognition tasks. Each Transformer block consists of an attention module and a feed-forward network (FFN), and features are down-sampled via overlap patch merging (a customized convolutional layer [21]) and passed to the next scale. The core attention module, by computing the global relevance of the semantic features, can better capture the overall contours of the key components, while also suppressing irrelevant artifacts and background interference. The attention mechanism can be expressed as:

$$Attention\ (Q,\ K,\ V) = Softmax(\frac{Q \cdot K^T}{\sqrt{d}}) \cdot V, \tag{6}$$

where "$\cdot$" is matrix multiplication, matrices Q, K, V represent the query, key, and value projections of the semantic feature F_i, respectively, d means the dimension of the feature, and $Softmax(\cdot)$ denotes the SoftMax normalization function.

The data flow of the Transformer block can be expressed as:

$$\hat{F} = LayerNorm(F + Attention(F)), \tag{7}$$

$$\widehat{F} = LayerNorm(\widehat{F} + FFN(\hat{F})), \tag{8}$$

where F, $\hat{F}$, $\widehat{F}$ represent the sequentially processed features, $LayerNorm(\cdot)$ denotes layer normalization, $Attention(\cdot)$ refers to the attention computation, and $FFN(\cdot)$ stands for the feed-forward neural network. The Transformer layer ensures network stability through residual connections.

2.3.2. Satellite Category and Pose Classification Heads

To explicitly guide the network in learning satellite category and pose features, we introduced three simple linear layers to transform F_4 into category logits $\vec{c} \in \mathbb{R}^{1\times C}$ and pose logits $\vec{m}$, $\vec{n} \in \mathbb{R}^{1\times P}$, where C and P represent the predefined number of space target categories and poses, respectively. In this work, $C = 3$, and the poses are discretized into 10-degree intervals, such that the index "0" position in the pose logits represents the possibility of 0 degrees, and the index "35" represents the possibility of 350 degrees (i.e., $P = 36$). It is worth noting that the predicted pose refers to the yaw and pitch angles of the target in the orbital coordinate system, which is different from the task of attitude estimation. The predicted pose only participates in the refinement of semantic features together with the class logits.

2.3.3. Category–Pose Guidance Module

To fully leverage the class and pose knowledge of the space targets, achieving refined key part recognition, a category–pose guidance module (CPGM) was designed to fuse the semantic features and predicted logits shown in Figure 2d. Specifically, the predicted logits $\vec{c}$, $\vec{m}$, $\vec{n}$ are first passed through three MLP (multi-layer perceptron) layers to get the affine weights α^j and biases β^j ($j = \vec{c}$, $\vec{m}$, $\vec{n}$). The MLP follows the structure shown in Figure 3, where the predicted logits are mapped to the same dimension as the semantic feature F_i for alignment, which can be expressed as:

$$\hat{j} = \text{Linear}(\text{Relu}(\text{Linear}(j))),\ j = \vec{c},\ \vec{m},\ \vec{n}, \tag{9}$$

$$\alpha^j,\ \beta^j = \text{Chunk}(\hat{j}). \tag{10}$$

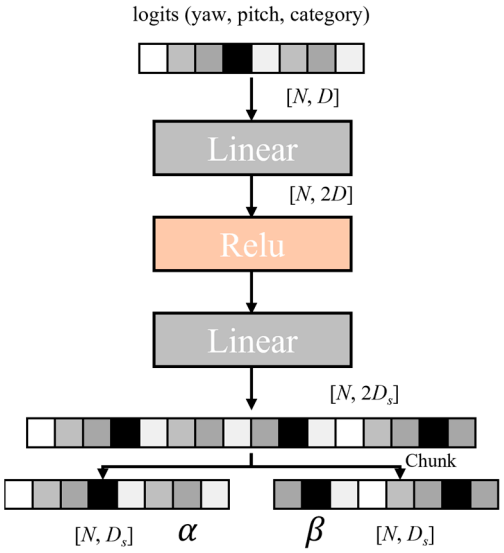


Figure 3. The structure of the MLP. "Relu" means the activation function, while "N", "D", and "D_s " mean the batch size and the dimensions of the logits and semantic features, respectively.

Subsequently, the category and pose information are embedded into the semantic features F_i in the form of weights and biases. To preserve effective information while

maintaining network stability, we added residual connections besides the embedding. The entire process can be expressed as:

$$\overline{F_i^j} = (1 + \alpha^j) \odot F_i + \beta^j, \tag{11}$$

where $\odot$ means the Hadamard product. To fully utilize the different types of information, the fused features $\overline{F_i^j}$ are concatenated along the channel dimension and then passed through a 1×1 convolutional layer to reduce the dimension, resulting in the final refined features $\overline{F_i}$, as can be represented by:

$$\overline{F_i} = conv(concat(\overrightarrow{\overline{F_i^c}}, \overrightarrow{\overline{F_i^m}}, \overrightarrow{\overline{F_i^n}})). \tag{12}$$

We first addressed the simpler image-level tasks, such as category and pose prediction, which in turn assisted in solving the more complex pixel-level task, i.e., key part recognition. Encoding category and pose information is equivalent to imposing additional prior knowledge on semantic features. The proposed CPGM can strengthen semantic features that are relevant to the current satellite state, guiding the network to achieve directed semantic detail recovery.

2.3.4. Semantic Decoding Layer

The refined semantic feature $\overline{F_i}$ is passed through an MLP layer to obtain the predicted key part annotation logits $Y \in \mathbb{R}^{T \times H \times W}$, where T represents the total number of classes of the key parts, and H and W denote the height and width of the annotation logits, respectively.

After the logits $Y \in \mathbb{R}^{T \times H \times W}$ were output, we added a relatively large value $e = 10$ to the dimensions of the predicted class. For example, if the network predicts the class as "ALOS", then the dimensions corresponding to "ALOS Antenna", "ALOS Solar Panel", and "ALOS Body" in the annotated logits Y will all be added with the value. Under the guidance of class-level knowledge, this can prevent the confusion of components from different satellites and reduce the difficulty of key part recognition for multiple target classes.

Finally, the key part recognition results were obtained through an argmax operation along the classes dimension (dimension 0), i.e.,

$$\hat{Y} = argmax(Y, dim = 0). \tag{13}$$

2.4. Loss Function

To train CPGNet, we need to accomplish two types of tasks: predicting the category and pose of the ISAR target and identifying its key parts. These two tasks can be regarded as image-level and pixel-level classification, respectively. Therefore, we uniformly constrained the predictions against the labels by the cross-entropy loss, which is widely used in classification tasks [21]. The total loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{cat} + \mathcal{L}_{yaw} + \mathcal{L}_{pitch} + \mathcal{L}_{seg}, \tag{14}$$

where,

$$\mathcal{L} = -\log \frac{\exp(x_{label})}{\sum_{l=1}^{L} \exp(x_l)},\ L = \begin{cases} 3, if\ \mathcal{L}_{cat} \\ 36,\ if\ \mathcal{L}_{yaw} \\ 36,\ if\ \mathcal{L}_{pitch} \\ 8, if\ \mathcal{L}_{seg} \end{cases}, \tag{15}$$

where $\mathcal{L}_{cat}$, $\mathcal{L}_{yaw}$, $\mathcal{L}_{pitch}$, and $\mathcal{L}_{seg}$ represent the loss functions used to optimize category prediction, yaw prediction, pitch prediction, and key part recognition, respectively. Here, x_{label} and x_l represent the values in the output logits corresponding to the GT label and label l, respectively. L represents the number of labels for each task, as defined in Section 2.3.2, the number of categories $C = 3$, the number of poses $P = 36$, and the number of key parts $T = 8$. Through explicit constraints on $\mathcal{L}_{cat}$, $\mathcal{L}_{yaw}$, and $\mathcal{L}_{pitch}$, the network can correctly learn knowledge related to satellite categories and poses, transforming them into the feature space aligned with semantic features, which helps $\mathcal{L}_{seg}$ converge more effectively.

3. Experiment and Discussion

3.1. Implementation Details

We trained on the simulated data from three satellites: KeyHole (KH), ALOS, and GSSAP (GP), and conducted validation on simulated scenes from KH and ALOS, as well as real-measured scenes from GP. The simulated parameters were set as follows: the center frequency was 216 GHz, the bandwidth was 20 GHz, and the range resolution and azimuth resolution were 7.5×10^{-3} m and 2.2×10^{-4} rad, respectively. The radar echoes were calculated by the physical optics method, and then the ISAR images were obtained through the range Doppler imaging method. The annotated key part images of the three target types are shown in Figure 4. We generated 1000 ISAR images with different poses for the three space targets (pitch and yaw angles ranging from 0 to 350 degrees, with an interval of 10 degrees) and partitioned the dataset into a training set and a testing set at an 8:2 ratio. Unlike previous key part recognition work, we refined the key part categories based on the target type, dividing them into eight classes: "KH Solar Panel", "KH Body", "ALOS Solar Panel", "ALOS Body", "ALOS Antenna", "GP Solar Panel", "GP Body", and "GP Antenna". We also annotated the target class and the pose for each ISAR image.

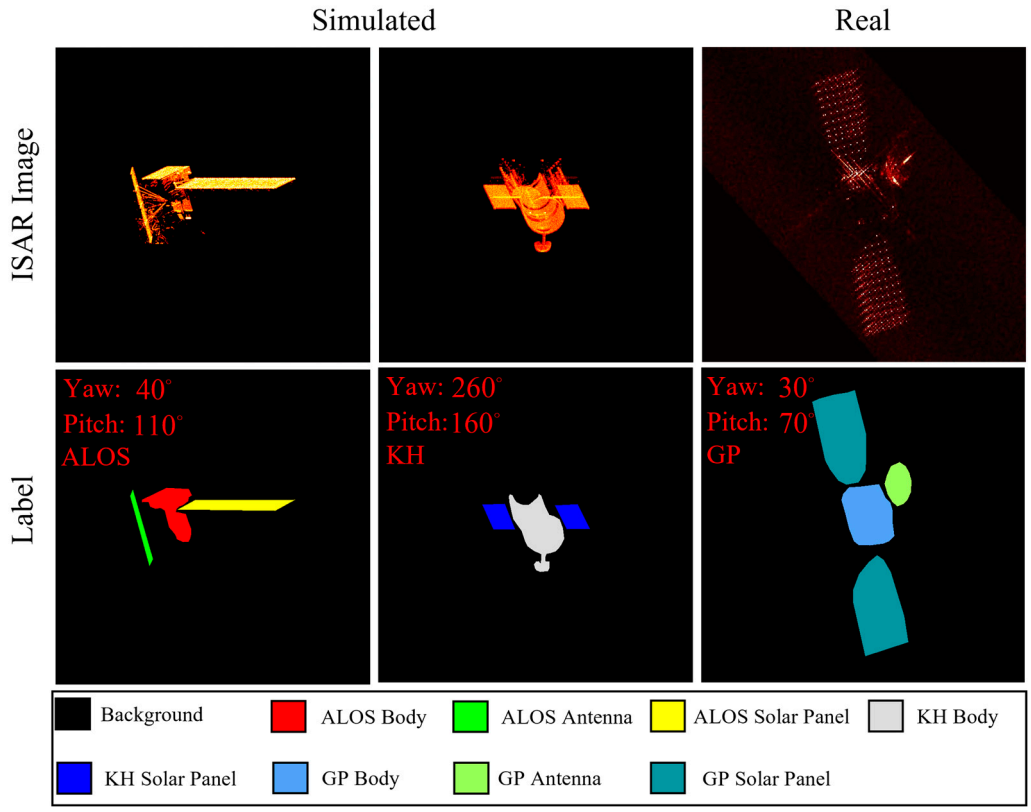


Figure 4. A demonstration of the simulated dataset, which included three space targets, "ALOS", "KeyHole", and "GP". The legend indicates the key parts represented by each color.

To ensure the stability of the training process, the network training was divided into two stages. In the first stage, the parameters of the encoder and the three classification heads were updated. In the second stage, the parameters of the entire network were updated. Both stages were trained for 1000 epochs, with a batch size of 2 and a fixed learning rate of 6×10^{-5}. The parameters were updated using the default AdamW optimizer. The configuration of the machine was NVIDIA GeForce RTX 3060 and i7-13700KF.

3.2. Experimental Configuration

The proposed CPGNet was compared with five CNN-based methods (i.e., FCN [8], Lraspp [23], Deeplab_v3 [13], U-Net [9], and ConvNeXt [24]) and two Transformer-based methods (i.e., MobileViT [10] and Segformer [21]). For evaluation metrics, MPA (mean pixel accuracy) and MIOU (mean intersection over union) [25] were employed to assess the accuracy of key part recognition, and they can be expressed as:

$$\text{MPA} = \frac{1}{L+1} \sum_{l=0}^{L} \frac{p_{ll}}{\sum_{k}^{L} p_{lk}}, \tag{16}$$

$$\text{MIOU} = \frac{1}{L+1} \sum_{l=0}^{L} \frac{p_{ll}}{p_{ll} + \sum_{k \neq l} p_{lk} + \sum_{k \neq l} p_{kl}}, \tag{17}$$

where L is the total number of key parts (L = 8 in this paper), p_{ll} means the number of pixels correctly predicted as key part l, $\sum_{k}^{L} p_{lk}$ means the total number of pixels predicted as key part l, $\sum_{k \neq l} p_{lk}$ means the number of pixels incorrectly predicted as other key parts when they belong to key part l (false negatives), and $\sum_{k \neq l} p_{kl}$ means the number of pixels incorrectly predicted as key part l when they belong to other classes (false positives). MPA reflects the classification ability of the model across different key parts, and MIOU considers the intersection and union of the predicted results and the GT. A higher value of both metrics means better effects.

3.3. Performance Analysis

The qualitative comparison results are shown in Figure 5. For simulated scenes (ALOS and KH), due to their high resolution, most methods could achieve satisfactory boundary segmentation results. However, Fcn and Lraspp could only distinguish satellites from the background, and their performance on internal key parts was poor. Both methods suffered from significant category confusion, whether within a single satellite or across different satellites. Deeplab_v3 showed noticeable improvement over the first two methods, but its recognition accuracy remained low, with blurry boundaries for key parts. U-Net and ConvNeXt achieved relatively accurate recognition of key part boundaries due to their superior architectural design; however, they still failed to address the issue of category confusion, with some regions exhibiting incorrect classifications. Although MobileViT is based on a Transformer architecture, its recognition accuracy appeared inferior to that of CNN-based ConvNeXt, suggesting that Transformer-based methods do not necessarily outperform CNNs. Among these approaches, Segformer was most closely related to our method due to its similar network structure. As shown in Figure 5, without category and pose guidance, Segformer demonstrated low recognition accuracy and severe confusion, where part of "ALOS" incorrectly appeared in the segmentation results of "KH". For the real-measured scene (GP), the low contrast and noise significantly degraded the performance of Fcn, Lraspp, and Deeplab_v3, resulting in low-confidence recognition outcomes. U-Net and ConvNeXt exhibited recognition errors in key parts, such as solar panels and antennas. MobileViT and Segformer demonstrated relatively satisfactory performance, indicating that Transformer-based methods possess robustness that surpasses CNNs.

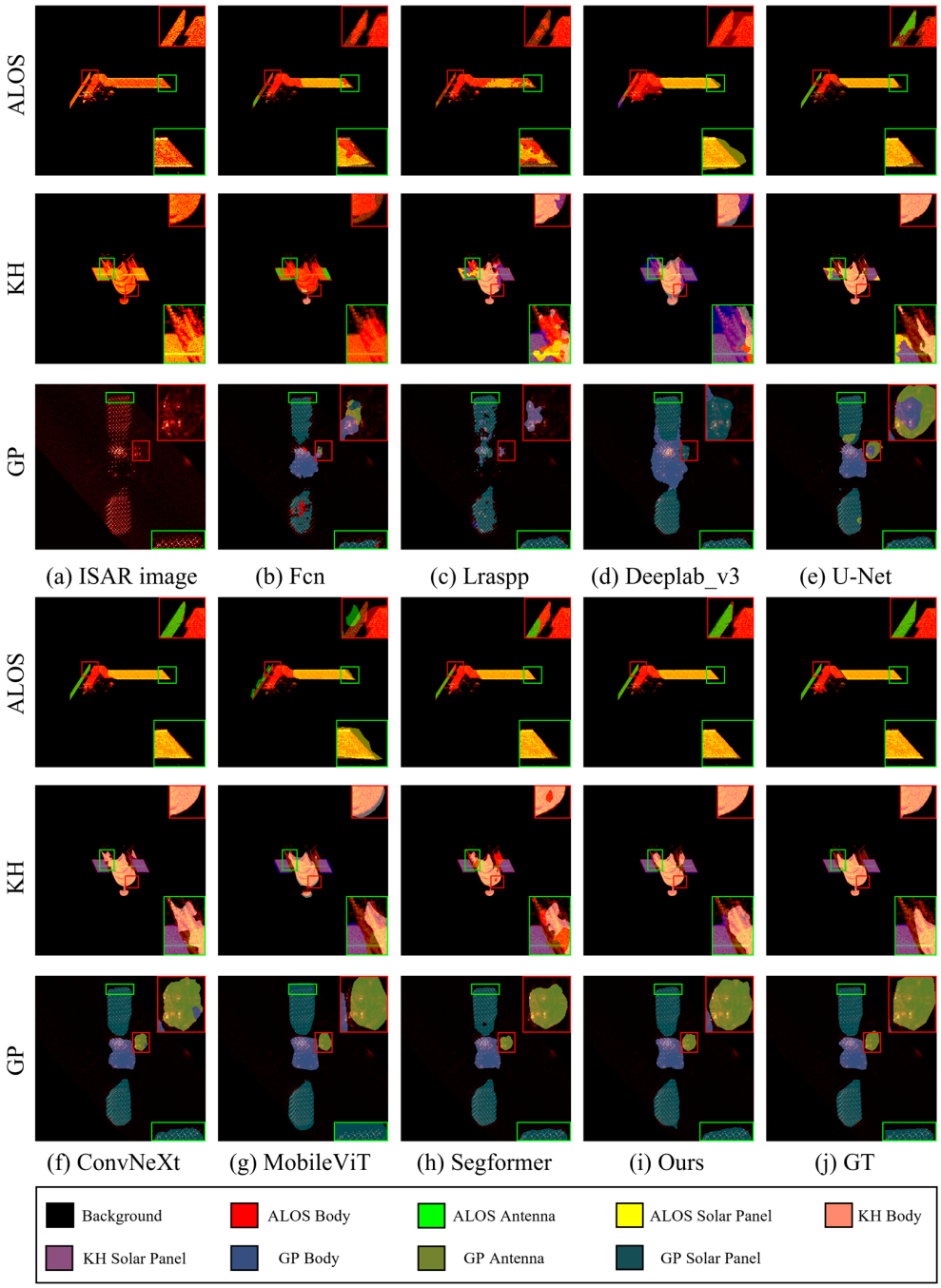


Figure 5. The key part recognition results on space targets "ALOS", "KeyHole", and "GSSAP". (**a**) The ISAR image, and (**b**–**j**) FCN, Lraspp, Deeplab_v3, U-Net, ConvNeXt, MobileViT, Segformer, ours, and the ground truth label. The legend indicates the key parts represented by each color.

In contrast, the proposed CPGNet integrates knowledge guidance with part recognition, leveraging a self-attention-based architecture and satellite structural information to guide the network in learning refined morphological features of the key part, achieving the most precise recognition results with well-defined key part contours. Moreover, under the guidance of category information, our method avoided category confusion, demonstrating the strong robustness on multi-target fine-grained category datasets. Our method also achieved the best performance on the real-measured scene. The satellite body and solar panels were segmented with results closest to the manual annotations, while the antenna's segmentation boundary, although slightly different from the GT, more accurately reflected the true morphology of the satellite (elliptical). This demonstrates that our method can cor-

rectly learn knowledge about satellite morphology, achieving results that surpassed manual annotations in terms of accuracy and realism. The optimal MPA and MIOU presented in Table 1 further demonstrate the superiority of our method in recognition accuracy.

Table 1. Quantitative results of recognition performance on three types of space targets. **Bold** indicates optimal outcomes.

Metric	Fcn	Laspp	Deeplab_v3	U-Net	ConvNeXt	MobileViT	Segformer	Ours
MPA	73.3	60.9	82.4	76.0	82.9	78.3	88.2	**94.3**
MIOU	62.4	51.5	59.7	73.1	77.3	75.6	82.0	**89.2**

3.4. Discussion on Category and Pose Prediction

In the proposed algorithm, the network additionally output predictions for category and pose. The category prediction reduced the complexity of the key part identification by providing image-level knowledge to the recognition head, thereby avoiding category confusion, as shown in Figures 5 and 6. The pose prediction head output the yaw and pitch angles of the target in the body coordinate system, which, while not directly meaningful in practical applications, represents a kind of perception of the target's pose, and we learn this knowledge through such an explicit approach. As seen in Figure 6, when predicting the "KH" target, the network's pose prediction significantly deviated from the labels. However, the ISAR image corresponding to the predicted angles shown in Figure 6c was highly similar in morphology to the input. A similar phenomenon was also observed in "ALOS", where the ISAR image corresponding to the prediction exhibited a mirror-symmetric relationship with the input. This indicates that the network developed a correct perception of the target's pose. This knowledge, integrated with semantic features through the CPGM module, assisted the network in achieving refined segmentation of ISAR images under various poses, as demonstrated in Figure 6b.

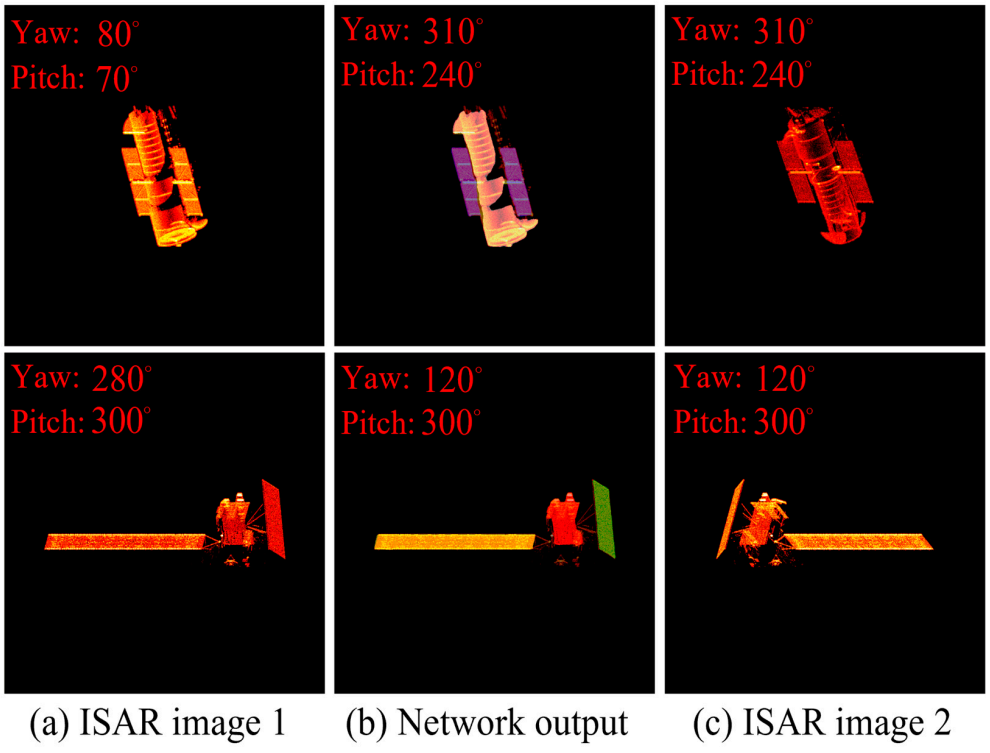


Figure 6. The discussion on pose prediction. (**a**) The input ISAR image, (**b**) the predictions of CPGNet, and (**c**) the ISAR image corresponding to the predicted poses.

Importantly, due to large morphological variations across satellite poses, our method encountered some recognition failures (Figure 7). However, the network still correctly predicted the category, which indicates that our model can degenerate from a segmentation model into a classification one, a characteristic absent in previous methods.

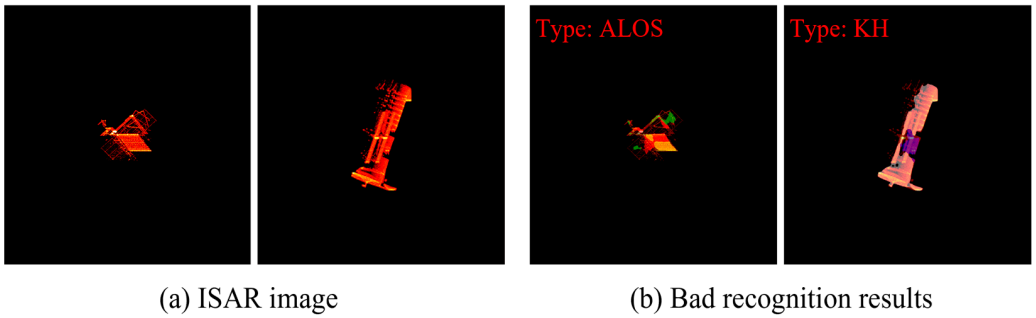


Figure 7. Some bad recognition results. The proposed method can still output the correct classification results. (**a**) Two typical ISAR images. (**b**) Two bad recognition results of these two samples.

3.5. Ablation Experiment

In this section, we first demonstrate the benefits of the fine-grained training strategy for recognition accuracy. As shown in Figure 8, the conventional strategy treated functionally similar components of different satellites as a single class. However, significant morphological differences exist between the bodies of ALOS and KH, causing the network to struggle in precisely learning spatial features during training and resulting in suboptimal recognition accuracy. In contrast, the proposed training strategy essentially refined the optimization space. The network learned structure features related to satellite categories within subspaces formed by different satellites. It no longer needs to compromise on the differences between these space targets, effectively reducing training complexity and achieving more accurate recognition results.

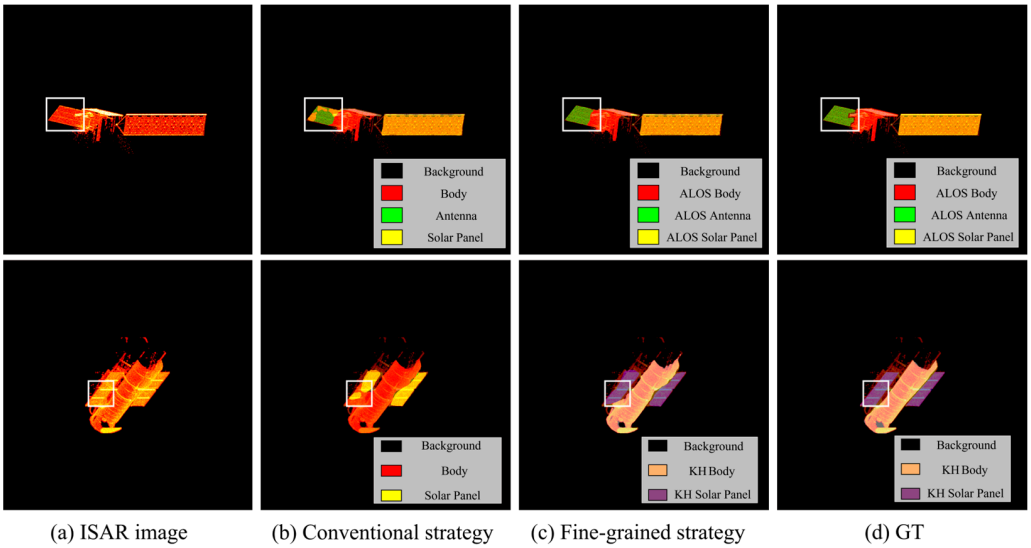


Figure 8. Discussion of the training strategy, with the corresponding colors representing the categories annotated in the legends. (**a**) ISAR images. (**b**–**d**) denote the results based on conventional strategy, fine-grained strategy and the ground-truth.

The specific impact of pose and category knowledge on the performance was also analyzed. As shown in Figure 9, "W/o pose" indicates the removal of pose knowledge, and although the network restricted the results to the "KH" through the guidance of category knowledge, the contours of the recognition were unclear, and a significant portion

of the solar panels was misidentified as the body. "W/o category" denotes the removal of category knowledge, which drastically increased the difficulty of the multi-target fine-grained category recognition task, leading to category confusion and reduced segmentation accuracy. Conversely, only by utilizing both types of knowledge could optimal performance be achieved, with the "KH" lens in the red box correctly segmented. Additionally, the quantitative evaluation metrics presented in Table 2 confirm that the current configuration was most suitable for the ISAR image key part recognition task.

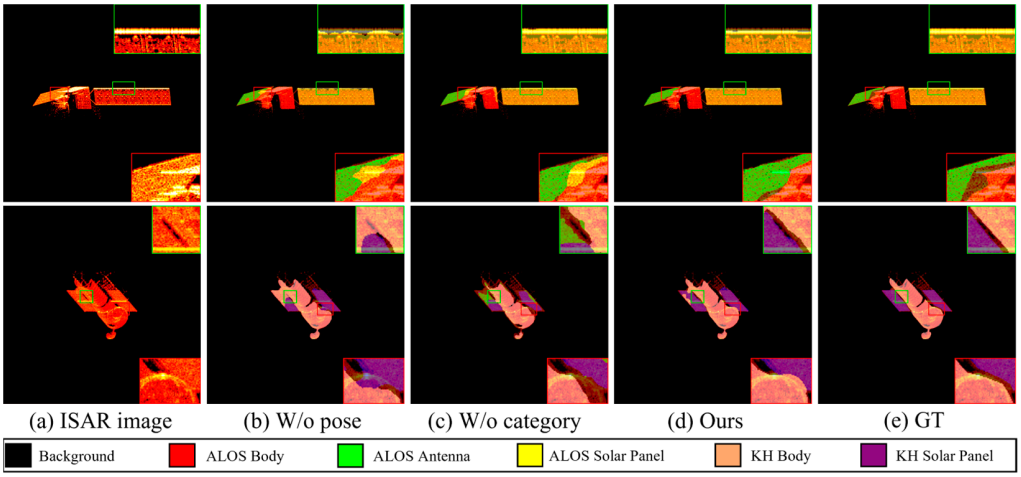


Figure 9. Ablation experiments on the proposed method. "W/o pose" and "W/o category" denote the removal of pose prediction and category prediction modules, respectively. (**a**) ISAR images. (**b**–**e**) denote the results based on W/o pose, W/o category, ours and the ground-truth.

Table 2. Quantitative results of the ablation experiment on three types of space targets. **Bold** indicates optimal outcomes.

Metric	W/o Pose	W/o Category	Ours
MPA	93.8	87.4	**94.3**
MIOU	88.9	83.5	**89.2**

3.6. Discussion on the Limitations and Future Work

The proposed CPGNet achieved high-accuracy segmentation of ISAR images through a category-refined strategy and knowledge embedding. However, the inability to obtain a large amount of real-measured data remains a major limitation of this kind of method. Additionally, the proposed training strategy made it difficult to generalize to unknown space targets, a challenge that was similarly hard to address in previous training strategies.

In the future, we will explore how to improve the quality (including quantity and diversity) of simulated data to better approximate real detection scenes, which can further enhance the generalization of the proposed method, thereby promoting the development of data-driven approaches.

4. Conclusions

In conclusion, a Transformer-based network incorporating knowledge of space target categories and poses was first proposed to address the issue of key part recognition in ISAR imaging. Compared to previous recognition methods, we introduced a key part refinement training paradigm and leveraged the full potential of category and pose knowledge during training to guide the network toward more precise key part recognition effects. Qualitative and quantitative evaluations on three space targets demonstrated the superiority of the proposed method.

Author Contributions: Methodology, Q.Y.; validation, H.W. and Q.Y., formal analysis, L.F. and H.W.; writing—original draft preparation, Q.Y.; writing—review and editing, S.L. and L.F.; visualization, Q.Y. and S.L.; supervision, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Science and Technology Innovation Program of Hunan Province under No. 2024RC3143, and in part by the National Natural Science Foundation of China under Grants 62201591 and 62035014.

Data Availability Statement: The datasets presented in this article are not readily available because the special reasons from the institute.

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

Abbreviations

The following abbreviations are used in this manuscript:

ISAR	Inverse Synthetic Aperture Radar
PCA	Principal Component Analysis
CPGM	Category–Pose Guidance Module
LOS	Line-of-Sight
GT	Ground Truth
CNN	Convolutional Neural Network
FFN	Feed-Forward Network
MLP	Multi-Layer Perceptron
ALOS	Advanced Land Observing Satellite
KH	KeyHole
GSSAP	Geosynchronous Space Situational Awareness Program

References

1. Camp, W.W.; Mayhan, J.T.; O'Donnell, R.M. Wideband radar for ballistic missile defense and range-doppler imaging of satellites. *Linc. Lab. J.* **2000**, *12*, 267–280.
2. Bai, X.; Yang, M.; Chen, B.; Zhou, F. REMI: Few-shot isar target classification via robust embedding and manifold inference. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 6000–6013. [CrossRef] [PubMed]
3. Wang, C.; Jiang, L.; Li, M.; Ren, X.; Wang, Z. Slow-spinning spacecraft cross-range scaling and attitude estimation based on sequential isar images. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 7469–7485. [CrossRef]
4. Fan, L.; Wang, H.; Yang, Q.; Deng, B. High-quality airborne terahertz video SAR imaging based on echo-driven robust motion compensation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 2001817. [CrossRef]
5. Kou, P.; Qiu, X.; Liu, Y.; Zhao, D.; Li, W.; Zhang, S. ISAR image segmentation for space target based on contrastive learning and nl-unet. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3506105. [CrossRef]
6. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
7. Fan, L.; Yang, Q.; Wang, H.; Qin, Y.; Deng, B. Sequential ground moving target imaging based on hybrid ViSAR-ISAR image formation in terahertz Band. *IEEE Trans. Circuits Syst. Video Technol.* **2025**. *early access*. [CrossRef]
8. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
9. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; proceedings, part III 18*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
10. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
11. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.

12. Li, X.; Yuan, H.; Li, W.; Ding, H.; Wu, S.; Zhang, W.; Li, Y.; Chen, K.; Loy, C.C. Omg-seg: Is one model good enough for all segmentation? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
13. Chen, L.-C. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
14. Wang, Z.H.; Wang, Z.; Jiang, L.B. Components segmentation algorithm for space target ISAR image based on clean. In Proceedings of the 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Ningbo, China, 9–12 July 2017.
15. Coe, M.; Jones, G.; Gashinova, M.; Cherniakov, M.; Martorella, M.; Alconcel, L.N.; Marchetti, E. Segmentation and Classification of Sub-THz ISAR Imagery. In Proceedings of the 2024 International Radar Symposium (IRS), Rennes, France, 21–25 October 2024.
16. Wang, J.; Du, L.; Li, Y.; Lyu, G.; Chen, B. Attitude and size estimation of satellite targets based on isar image interpretation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5109015. [CrossRef]
17. Li, C.; Zhu, W.; Qu, W.; Ma, F.; Wang, R. Component recognition of ISAR targets via multimodal feature fusion. *Chin. J. Aeronaut.* **2025**, *38*, 103122. [CrossRef]
18. Zhu, X.; Zhang, Y.; Lu, W.; Fang, Y.; He, J. An isar image component recognition method based on semantic segmentation and mask matching. *Sensors* **2023**, *23*, 7955. [CrossRef] [PubMed]
19. Chen, L.; Yang, C.; Shi, Z.; Wang, K.; Li, Y. Satellite Component Segmentation Based on Symmetry Loss and Edge Perception. In Proceedings of the 2024 IEEE 17th International Conference on Signal Processing (ICSP), Suzhou, China, 16–18 December 2024.
20. Zhong, F.; Gao, F.; Liu, T.; Wang, J.; Sun, J.; Zhou, H. Scattering Characteristics Guided Network for ISAR Space Target Component Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 4009505. [CrossRef]
21. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
23. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
24. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
25. Hao, S.; Zhou, Y.; Guo, Y. A Brief survey on semantic segmentation with deep learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]

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