remote sensing

MDPI

Article

Adaptive Granularity-Fused Keypoint Detection for 6D Pose Estimation of Space Targets

Xu Gu [1], Xi Yang [1,*], Hong Liu [2] and Dong Yang [3]

[1] State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; ryangu@stu.xidian.edu.cn
[2] Space Engineering University, Beijing 101416, China; hliu@nwpu.edu.cn
[3] Xi'an Institute of Space Radio Technology, Xi'an 710100, China; yangd@cast504.com
* Correspondence: yangx@xidian.edu.cn

Abstract: Estimating the 6D pose of a space target is an intricate task due to factors such as occlusions, changes in visual appearance, and background clutter. Accurate pose determination requires robust algorithms capable of handling these complexities while maintaining reliability under various environmental conditions. Conventional pose estimation for space targets unfolds in two stages: establishing 2D–3D correspondences using keypoint detection networks and 3D models, followed by pose estimation via the perspective-n-point algorithm. The accuracy of this process hinges critically on the initial keypoint detection, which is currently limited by predominantly singular-scale detection techniques and fails to exploit sufficient information. To tackle the aforementioned challenges, we propose an adaptive dual-stream aggregation network (ADSAN), which enables the learning of finer local representations and the acquisition of abundant spatial and semantic information by merging features from both inter-layer and intra-layer perspectives through a multi-grained approach, consolidating features within individual layers and amplifying the interaction of distinct resolution features between layers. Furthermore, our ADSAN implements the selective keypoint focus module (SKFM) algorithm to alleviate problems caused by partial occlusions and viewpoint alterations. This mechanism places greater emphasis on the most challenging keypoints, ensuring the network prioritizes and optimizes its learning around these critical points. Benefiting from the finer and more robust information of space objects extracted by the ADSAN and SKFM, our method surpasses the SOTA method PoET (5.8°, 8.1° / 0.0351%, 0.0744%) by 0.5°, 0.9°, and 0.0084%, 0.0354%, achieving 5.3°, 7.2° in rotation angle errors and 0.0267%, 0.0390% in normalized translation errors on the Speed and SwissCube datasets, respectively.

Keywords: 6D pose estimation; space target; keypoint detection

Citation: Gu, X.; Yang, X.; Liu, H.; Yang, D. Adaptive Granularity-Fused Keypoint Detection for 6D Pose Estimation of Space Targets. *Remote Sens.* **2024**, *16*, 4138. https://doi.org/10.3390/rs16224138

Academic Editors: Hossein M. Rizeei, Qi Zhao, Guangliang Cheng and Shuchang Lyu

Received: 23 September 2024
Revised: 28 October 2024
Accepted: 5 November 2024
Published: 6 November 2024

1. Introduction

With increasing activity in space operations, the task of 6D pose estimation for space targets has become a major area of interest. By obtaining the three-dimensional translation and rotation of space targets relative to the camera coordinate system, their physical state can be better monitored [1], and subsequent trajectories can be predicted. Therefore, accurate and efficient 6D pose estimation of space targets is of strategic significance for the increasingly complex demands of space missions, such as space defense, space observation, satellite docking, satellite capture, and other tasks.

Benefiting from the rapid development of deep learning, the most effective method for 6D pose estimation of space targets currently uses a two-step strategy: first, using a keypoint detection network, followed by pose recovery. However, most methods rely on single-grained feature extraction techniques, which extract feature information of space targets at a single scale. Single-scale techniques, while computationally efficient, lack the ability to capture both fine and large-scale object details, reducing detection accuracy in complex environments. As illustrated in Figure 1, the challenges of occlusions, scale variation, and viewpoint changes present major obstacles to accurate pose estimation. In such environments, single-grained features struggle to effectively capture all necessary details, directly leading to imprecise keypoint localization. For example, for distant or truncated targets, single-grained features may fail to provide enough distinctiveness, making keypoint detection unreliable. Additionally, when faced with targets of different sizes and shapes, single-grained features also struggle to adapt, thereby reducing the overall accuracy and robustness of pose estimation. Chen et al. [2] improved the performance of the keypoint detection network by enhancing the high-resolution network (HRNet) [3] to predict more accurate keypoint positions, thus recovering more precise pose information. Although this layer-wise fusion-based method repeatedly enhances inter-layer features to obtain high-resolution features, these features are still coarse-grained and contain limited information. Alternatively, a few works address only features within layers, which also have significant limitations. Moreover, due to issues of truncation and viewpoint changes, the challenging keypoint detection strategy initially used by the keypoint detection network is no longer applicable, and these challenging keypoints require a more robust algorithm for effective handling.

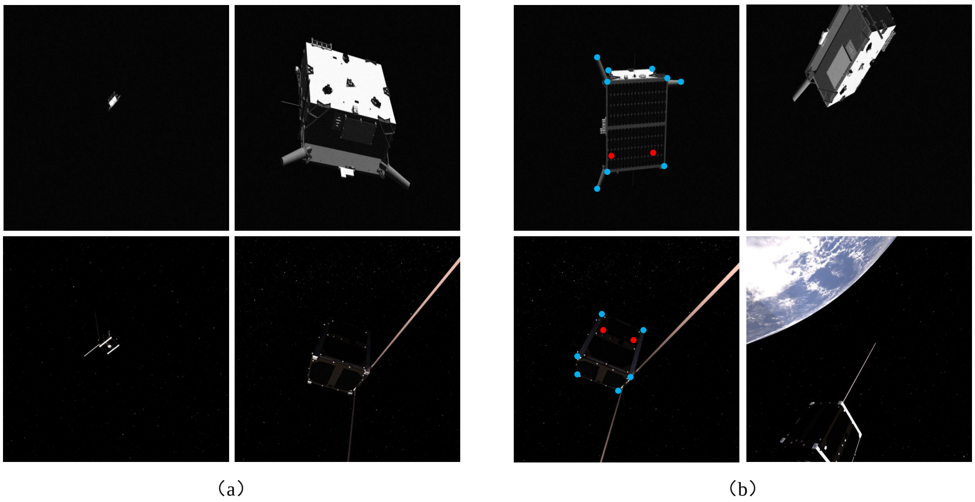


Figure 1. The challenges faced in pose estimation of space targets. (**a**) Scale variation: large-scale space targets can occupy thousands of pixels in an image, while small-scale space targets may cover only a few tens of pixels. (**b**) Viewpoint changes and truncation: viewpoint changes often lead to some points becoming invisible, such as the red points in the image, while truncation results in missing points.

To improve the detection accuracy of keypoints in space target data and thus achieve more accurate pose estimation, we propose an adaptive dual-stream aggregation network (ADSAN). This is a multi-grained keypoint detection framework that combines inter-layer fusion and intra-layer aggregation, breaking the limitations of traditional single-grained keypoint detection networks. The network employs an inter-layer fusion strategy to sample and fuse feature maps of different resolutions, ensuring that each layer's feature map shares information from other layers so that the final output features contain global information from different dimensions. At the same time, our ADSAN divides features within each layer into finer and smaller features, continuously aggregating these fine-grained features within the layer to ensure that each layer's output features contain rich local information. Finally, the output features containing rich global and local information are used to detect keypoints. By combining inter-layer fusion and intra-layer aggregation, this network extracts more informative features to avoid the limitations of single-grained detection, thereby improving the accuracy and robustness of keypoint detection. In addition, we propose a selective keypoint focus module, which first calculates the mean square error loss of each keypoint, and then sets a threshold manually to retain keypoints with a loss greater than the threshold and discard the rest, iteratively calculating the overall loss. By dynamically selecting the number of keypoints in this manner and retaining occluded or invisible points, the network's ability to detect challenging keypoints can be further improved.

The contributions of this work can be summarized as follows:

(1) Unlike previous single-grained keypoint detection networks, we design a multi-grained dual-stream aggregation network that fully integrates inter-layer features and intra-layer features to obtain richer keypoint detection features.

(2) We propose a selective keypoint focus module, which addresses the problem of challenging keypoint detection in space target data by dynamically selecting keypoints that are hard to detect and improves the robustness of detection.

(3) We demonstrate the significant performance improvements of our approach compared to the state-of-the-art approaches on the Speed and SwissCube datasets.

2. Related Works

2.1. Six-Dimensional Pose Estimation

Methods for estimating 6D poses often require pre-processing, such as detection [4–6] or segmentation [7,8] for space targets, to eliminate background interference. For instance, Huo et al. [9] applied a Tiny-YOLOv3-based [10] architecture with a detection subnetwork. Huan et al. [11] utilized a cascade Mask R-CNN with HRNet as the backbone to extract the masks of specific satellite objects. Lotti et al. [12], on the other hand, used a single-stage detector with a Swin Transformer [13] as the backbone and an additional discriminator head present during training, achieving a further boost in performance. After the detection stage, methods can be further divided into template-based, voting-based, correspondence-based, regression-based, and reconstruction-based approaches. Template-based methods select the most similar template from pre-labeled templates and use its pose as the estimation for the current object. For example, Hinterstoisser et al. [14] compared gradient information between observed RGB images and template RGB images to find the most similar template and used its corresponding pose as the estimation. Voting-based methods [15] generate multiple pose predictions, which are then refined and selected to obtain the final pose. Correspondence-based methods establish correspondences between feature points in the observed data and those in a reference model and then use these correspondences to compute the transformation that aligns the model with the observations. Lin et al. [16] proposed AG-Pose to establish robust keypoint-level correspondences for unseen instances. Regression-based methods [17] learn a direct mapping from raw sensor input, such as images or depth maps, to the pose parameters using machine learning techniques, often requiring large annotated datasets for training. Reconstruction-based [18,19] methods first reconstruct the scene or the object from sensor data and then infer the pose by aligning this reconstruction with a known model or by analyzing the reconstructed geometry directly. Antoine et al. [20] introduced an innovative domain generalization method for 6D pose estimation by utilizing Neural Radiance Field (NeRF)-based [21] image synthesis to enrich the diversity of datasets, leading to significant improvements in spacecraft pose estimation models. Recently, works utilizing foundation models like diffusion [22] have shown promise. Wang et al. [23] proposed aggregating diffusion features with different granularities, greatly improving the generalizability of object pose estimation. However, space objects often lack distinct textures and are subject to harsher conditions, such as extreme lighting changes, background interference (e.g., Earth or stars), and severe occlusions, which are less common in ground-based pose estimation tasks, making it difficult for the above methods to be applied directly.

The most commonly used method for pose estimation of space targets is based on corresponding points. The standard framework involves establishing a correspondence between 3D and 2D data, followed by using a perspective-n-point (PnP) solver to recover the object's pose [24–27]. Previous methods mainly relied on manually crafting features [28–30], but they often produced low-quality outputs under challenging conditions. Therefore, recent methods use neural networks to form 2D–3D matching relationships. These networks are typically trained to predict the image location of the corners of the 3D object bounding box, either in a single global manner [31–34] or by aggregating multiple local predictions to improve robustness to occlusions [35–40]. Xiang et al. [34] presented the PoseCNN model for object pose estimation, which is divided into two branches for pose estimation. A Convolutional Neural Network (CNN) architecture based on the AlexNet network [41] was used in non-cooperative spacecraft to solve classification problems and return the relative pose of the space target associated with each image [42]. Sun et al. [43] combined deep learning and geometric optimization to propose a milestone regression model [2] based on HRNet. Harvard et al. [44] used CNN-based keypoints and visibility maps to determine the pose of the target spacecraft. Real-Time Structure from Motion (RTSfM) [45] involves a robust tracking approach, where the relative pose and its scale are estimated separately and then jointly optimized by considering PnP and epipolar constraints. However, the above methods, such as PoseCNN, still struggle when keypoints are occluded by the spacecraft's own structure, leading to significant errors in pose recovery. In contrast, the ADSAN's selective focus on occluded keypoints allows for more accurate pose estimation under these conditions.

2.2. Keypoint Detection

The keypoint detection network serves as a prerequisite in the pose estimation task, and its performance determines whether the final pose is accurate. Especially when the coordinates of 3D keypoints have been determined, detecting accurate 2D keypoints becomes the core of the whole pose estimation network. Traditional keypoint detection methods involved the use of hand-crafted features [28,46,47] designed to be invariant to changes in image scale, rotation, and lighting. While these approaches were successful in many applications, they were limited due to their reliance on manually engineered features.

To address these limitations, deep learning-based keypoint detection methods have emerged in recent years. These methods usually require a large amount of labeled data to train the network, but they have higher scalability and better performance. They use CNNs or Transformers to learn features directly from images, enabling end-to-end training for keypoint detection tasks [48–56]. There are two mainstream methods: regressing the locations of keypoints [57–60] and estimating keypoint heatmaps [61–63], and then selecting the location with the highest heat value as the keypoint. The former method directly takes the coordinates of keypoints as the target that the network needs to return. In this case, the direct position information of each coordinate point can be obtained directly. In the latter method, each type of coordinate is represented by a probability map, and a probability is assigned to each pixel position in the picture, indicating the probability that the point belongs to the keypoints of the corresponding category. The probability of the nearest pixel is closer to 1, and the probability of the pixel farther away from the keypoints is closer to 0. Specifically, it can be simulated using corresponding functions, such as two-dimensional Gaussian. Despite the progress in keypoint detection using deep learning, there are still some challenges that need to be addressed. One of the main challenges is robustness to changes in image scale, rotation, and illumination. Another challenge is the ability to detect keypoints in cluttered or occluded scenes, where keypoints might be partially or fully occluded. Addressing these challenges requires the development of more powerful and discriminative feature representations, as well as the use of more sophisticated training and inference strategies that can handle variations in the input data.

Our focus is on the keypoint detection task for space target images, which involves numerous interfering factors, such as background noise, truncation, and viewpoint changes. In such complex scenarios, it is essential to extract more informative features to provide the keypoint detection network with richer information. Therefore, we propose a dual-stream aggregation network architecture. This network effectively integrates features of different scales between layers and finer features within layers, acquiring both global and local information, and it resolves interference issues in space target images through rich dual information.

3. Materials and Methods

3.1. Overview

In this section, we illustrate the general pipeline of our method for pose estimation of space targets. As shown in Figure 2, our pose estimator first uses an object detection network to locate the space target. The keypoint detection network then predicts and positions the keypoints on the target. Finally, the detected 2D keypoints are matched with 3D model keypoints, and the pose information is recovered using the PnP solver.

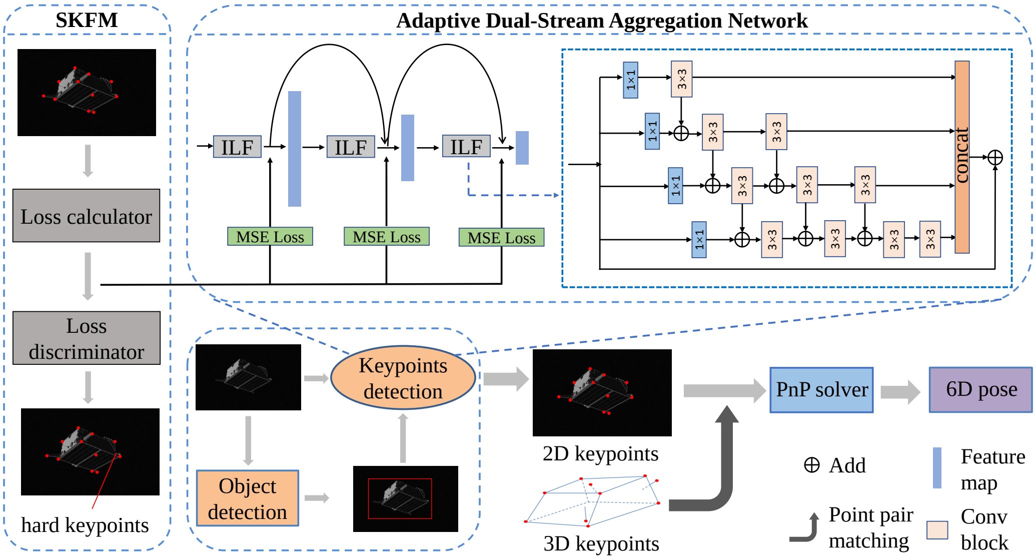


Figure 2. The structure of the proposed adaptive dual-stream aggregation network. The ADSAN first performs object detection to locate the position of the space target and detects 2D keypoints. Then, the 3D keypoints pre-matched with the 2D keypoints are input together into the PnP solver to recover the true pose of the space object. To obtain rich feature information, the keypoint detection network adopts a dual-stream aggregation method, repeatedly fusing features between layers and within layers to extract more comprehensive global features and finer local features for keypoint prediction. ILF stands for intra-layer fusion. Additionally, an SKFM is employed to further improve keypoint detection performance.

Object Detection. Due to the complex space environment, space target images often suffer from Earth background interference and lighting effects, as shown in Figure 3. These conditions can disturb keypoint detection networks, leading to a decrease in performance. Therefore, it is necessary to perform object detection on space target images, locate the specific positions of the space targets, and eliminate irrelevant background and lighting interference. We employ advanced single-stage object detection methods to detect and locate space targets.

Keypoint Detection. We then proceed with keypoint detection on noise-free data. We primarily utilize the ADSAN to predict accurate keypoints, and this network is detailed in Section 3.2. Our ADSAN performs sample fusion between feature maps of different resolutions across layers, enabling each layer's feature map to share information from other layers. Within each layer, features are further divided into finer, smaller features, which continuously aggregate within the layer. This process ensures that each layer's output features contain rich local and detailed information. Additionally, the selective keypoint focus module (SKFM) module dynamically selects and retains keypoints with the highest losses while discarding the rest, allowing the network to focus more on these high-loss truncation and occlusion points, leading to more reliable keypoint detection. We further explain the specific algorithm of the SKFM in Section 3.3.

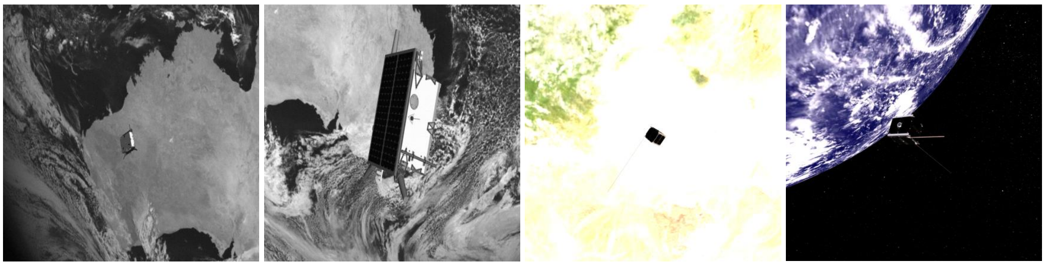


Figure 3. The issues of background interference and lighting effects in space target images.

Pose Estimation. The keypoint detection network outputs the 2D keypoint positions on space target images, inherently correlating with the predefined coordinates on the 3D models of space targets. By utilizing these 2D–3D correspondences as input to the non-differentiable PnP algorithm, the pose information of the space targets, encompassing both translation and rotation, can be accurately determined.

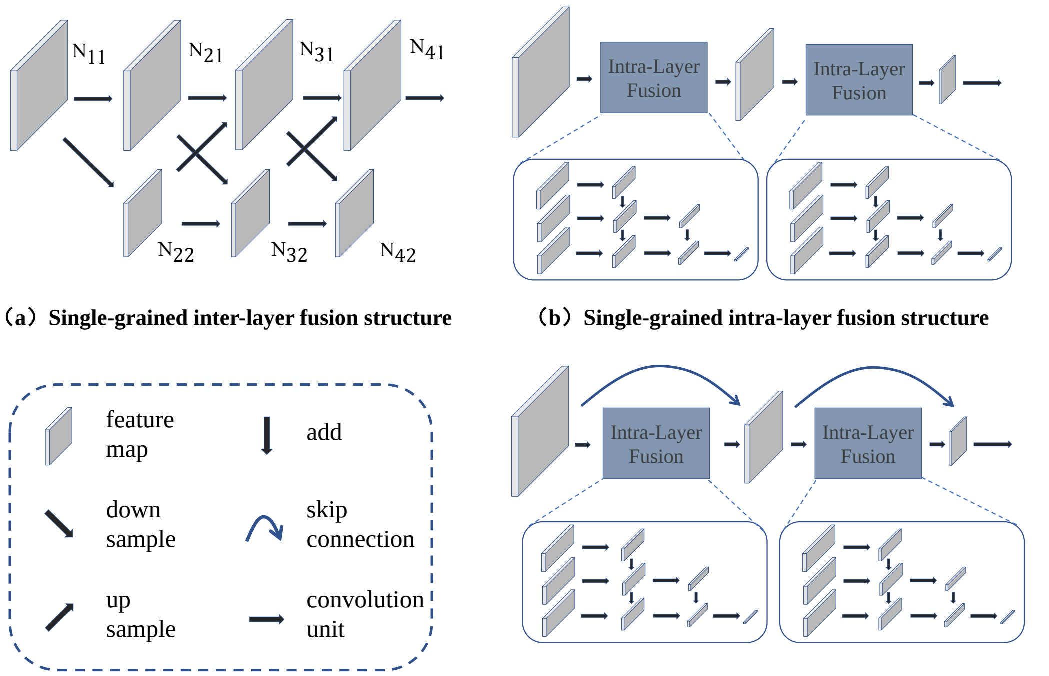
3.2. Adaptive Dual-Stream Aggregation Network

As depicted in Figure 4, single-grained feature fusion has limitations in both inter-layer and intra-layer contexts. Our proposed ADSAN enhances feature extraction by performing both inter-layer and intra-layer fusion to generate more robust information features. Inter-layer fusion combines feature maps from different layers, integrating information across multiple scales to capture a broader context. This fusion helps in gathering coarse, global features necessary for understanding the overall structure. Intra-layer fusion, on the other hand, aggregates finer details within each layer, focusing on spatially informed features that capture local information. By combining these two fusion types, the ADSAN incorporates both rich global and local information, enabling more precise keypoint detection. This dual approach allows the network to leverage complementary information from different scales, addressing the limitations of single-grained fusion methods that often miss critical details in complex environments.

Inter-layer Fusion Module. The inter-layer fusion module in the ADSAN is inspired by the skip connection mechanism in the ResNet architecture [64], which facilitates information fusion by adding the input directly to the output of each feature layer. Given input features x, the network learns a mapping $H(x)$ through multiple layers. By setting the residual function as $F(x) = H(x) - x$, the network reformulates the original mapping as $F(x) + x$. This structure allows the network to learn residuals $F(x)$ more effectively while maintaining the input x across layers. Through this design, the ADSAN effectively combines features across layers without altering the overall objective.

Intra-layer Aggregation Module. The intra-layer aggregation module divides the features of each layer more finely and then aggregates these finer features to obtain finer local features. Specifically, the intra-layer aggregation module divides the features into multiple parts, represented as $fm(m = 1, 2, 3, 4, \ldots)$, which then pass through the convolution layer with a convolution kernel size of 1. The features output from the 1 × 1 convolutional layer are also sent to the 3 × 3 convolutional layer and then added to the output of the latter feature 1 × 1 convolutional layer. Finally, the final output of each finer feature fm is connected and passed through another 1 × 1 convolutional layer to obtain the output of the intra-layer aggregation module.

On each finer feature branch, the intra-layer aggregation module receives the features output by the convolutional layer from the previous finer feature branch (except for the first branch) and then continuously refines these features through 3 × 3 convolution; the refined features are sent to the next finer feature branch. Through this dense fusion method, the small receptive field in finer features is completely fused, thereby expanding the receptive field, generating more accurate and finer local features, and retaining richer spatial features. At the same time, the densely connected structure provides sufficient gradients during training, which better supervises low-level features and ensures the normal progress of training.



（a）**Single-grained inter-layer fusion structure**

（b）**Single-grained intra-layer fusion structure**

（c）**Granularity-fused dual stream aggregation structure**

Figure 4. Comparison of single-grained fusion structure and multi-grained fusion structure. The single-grained fusion structure includes the inter-layer fusion structure and the intra-layer fusion structure. These two single-grained structures obtain single features and contain limited information. The granularity-fused dual-stream aggregation structure fuses inter-layer features and intra-layer features, and the acquired feature information is more abundant.

3.3. Selective Keypoint Focus Module

When performing keypoint detection on space targets, the following problems exist: (1) Some keypoints are invisible due to different viewing angles; and (2) due to the presence of truncation phenomena, information for some keypoints is blocked, resulting in poor feature learning. Keypoints that are difficult to detect are called "challenging keypoints", as indicated by the red dots in Figure 5. These challenging keypoints are often important factors that are difficult to learn in keypoint detection networks and are also the main reason for poor network performance, so it is necessary to focus on challenging keypoint detection. In real-world scenarios, like a satellite partially occluded by its own solar panels, focusing on hard-to-detect keypoints allows the model to maintain pose accuracy even when large parts of the object are not visible. In order to help the network learn better features and pay more attention to the challenging keypoints, we introduce a selective keypoint focus module in the dual-stream aggregation network.

In the training phase of the keypoint detection network, the model classifies high-confidence keypoints as positive-sample keypoints and low-confidence keypoints as negative-sample keypoints. These positive-sample and negative-sample keypoints can be represented by a loss function. "Challenging keypoints" can be classified as negative-sample keypoints because they are difficult to detect and have a large loss. The selective keypoint focus module makes the network pay more attention to these negative-sample keypoints by strengthening their training. In the experiments, we select 11 keypoints in the Speed dataset and 8 keypoints in the SwissCube dataset, including all vertices of the space target itself. The network typically computes an overall loss for all keypoints. However, after introducing the SKFM, the network is forced to pay more attention to the keypoints with the largest loss during the training process, thereby improving detection accuracy. The specific process of the algorithm involves generating the most likely keypoints based on the initial MSE loss and then performing negative-sample keypoint mining on the generated keypoints. The original method calculated the loss of all keypoints and incorporated them into the overall loss. However, the SKFM sets a loss threshold and includes only those keypoints whose loss is greater than the threshold in the overall loss while discarding the other keypoints, as shown in Figure 5. Through this operation, different numbers of keypoints are dynamically retained in each image, and these keypoints are the most challenging negative-sample keypoints for the network to detect. Therefore, the overall performance of the network is improved. The loss function of the space target keypoint detection network is as follows:

$$L_{overall} = \sum_{n=1}^{N} |y_{gt} - y_{pre}|, N = \sum_{k=1}^{11} \left(L_{\mathrm{MSE}}^{(k)} > \delta \right), \tag{1}$$

where y_{gt}, y_{pre}, n, N, and δ represent the real coordinate values of the keypoints of the space target, the coordinate values predicted by the keypoint detection network, the number of keypoints, the number of negative-sample keypoints, and the MSE loss threshold, respectively.

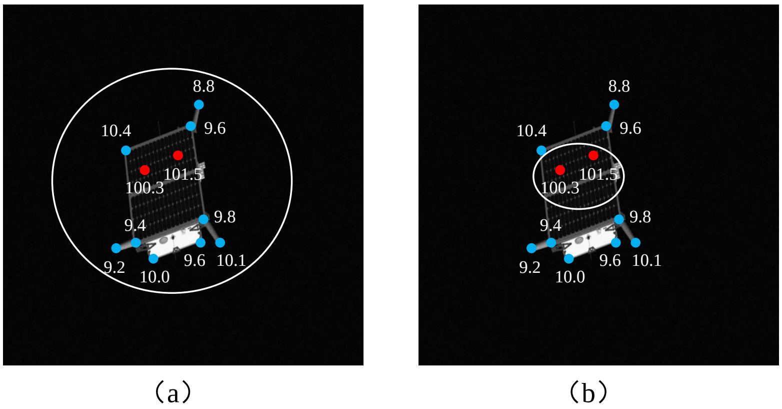


Figure 5. Comparison between traditional mining strategy and the SKFM. (**a**) The traditional mining strategy focuses on all selected fixed-number points (the circle contains both blue and red points). (**b**) The SKFM focuses on invisible or truncation points (the circle contains only red points) and reduces attention to other points (the circle does not contain blue points).

4. Results

4.1. Experimental Settings

Dataset. Our dual-stream aggregation pose estimator was evaluated on the Speed [65], SwissCube [66], and Speed-noise datasets.

The Speed dataset contains a total of 15,303 images, including 305 real images and 14,998 synthetic images. The synthetic ones were created using camera emulator software, with half of them incorporating random Earth images as backgrounds to generate photo-realistic images. The synthetic images are divided into 12,000 training images and 2998 test images, of which the test images and real images have no ground truth; only the training images have the true values of the keypoints and the true values of the poses. So, in the subsequent experiments, we randomly divided the training images into a training set, a validation set, and a test set in a ratio of 9:1:2, with the validation set used to tune the hyperparameters and prevent overfitting during training. The size of the images in the Speed dataset is 1920 × 1200.

The SwissCube dataset is used to evaluate 6D pose estimation algorithms for space targets in a wide depth range. The images were created by rendering a highly realistic 3D model of a satellite in a space environment, using accurate material properties and realistic illumination from the Sun, Earth, and surrounding star fields. This dataset comprises 500 scenes, each containing a sequence of 100 frames, totaling 50,000 images. From these images, 40,000 were extracted from 400 scenes for training, while the remaining 10,000 images were taken from 100 scenes for testing purposes.

To further test the model's performance and robustness under real noisy scenarios, we modified the Speed dataset to create its noisy variant, named the Speed-noise dataset, by adding Gaussian noise with a mean of 0 and a variance of 0.00001, along with motion blur parameterized with a degree of blur of 10 and an angle of blur of 5. The ground-truth keypoints remained consistent with the original positions, ensuring that the comparison between the predicted and actual keypoints remained valid, as the physical keypoint locations on the objects do not change; only the image quality does. The noisy version of the Speed dataset provided a significant challenge for the inference resistance tests.

Implementation Details. In our experiments, unless otherwise specified, we always used ResNet-50 as the backbone network with a four-stage encoder. These experiments were all carried out on a single RTX TITAN, using the Adam optimizer with a weight decay of 0.00001, and the batch size for training on the GPU was 16. The learning rate started at 5×10^{-4} and was multiplied by 0.1 at each decay step. During our training and inference processes, we re-scaled the Speed images to 160×240 and the SwissCube images to 256×256. The loss threshold of the SKFM was empirically set to 50 due to the observation that both the Speed dataset and the SwissCube dataset exhibited "challenging keypoint" loss values consistently above 100, while the loss for the other keypoints remained around 10, as shown in Figure 5. Such a loss value gap allowed us to choose a median value of 50 for a rather good separation between the normal keypoints and "challenging keypoints". Across both datasets, we conducted a total of 76,800 iterations, adjusting the learning rate at the 19,200th iteration. During the testing phase, we adopted a batch size of 32.

Evaluation Metrics. The accuracy of pose estimation was evaluated by calculating the magnitude of the object rotation angle E_R and translation error E_T in the prediction space. In the fields of aerospace and aviation, quaternions are widely used to represent these two metrics. They typically consist of four components:

$$q = q_0 + iq_1 + jq_2 + kq_3, \tag{2}$$

where *i*, *j*, and *k* represent three imaginary units, respectively. In the formula, q_0, q_1, q_2, and q_3 satisfy the constraint condition $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$. The basic mathematical equation for a quaternion is given as follows:

$$q = \cos(\theta/2) + i(x \cdot \sin(\theta/2)) + j(y \cdot \ \sin(\theta/2)) + k(z \cdot \ \sin(\theta/2)), \tag{3}$$

where *(x, y, z)* represents the rotation axis and θ represents the angle of rotation about the rotation axis. Specifically, q^*, q, t^*, t, $|\cdot|$, $<\cdot>$, $|\cdot|_2$ and m represent the real label of the quaternion of the target image, the predicted label of the quaternion, the real label of the translation vector of the target image, the predicted value of the translation vector, the absolute value, the dot product, the 2-norm, and the m-th space target image, respectively. The rotation angle error $E_R^{(m)}$ and translation vector error $E_T^{(m)}$ of the estimated rotation angle of the m-th space target image are formulated as

$$E_R^{(m)} = 2 \cdot \arccos(| < q_{(m)}, q^*_{(m)} > |), E_T^{(m)} = \frac{|t^*_{(m)} - t(m)|_2}{|t^*_{(m)}|_2}. \tag{4}$$

The overall rotation angle error and translation vector error are computed separately as the average of the rotation angle errors and the average of the translation errors across all images, defined as follows:

$$E_R = \frac{1}{M} \sum_{m=1}^{M} E_R^{(m)}, E_T = \frac{1}{M} \sum_{m=1}^{M} E_T^{(m)}, \tag{5}$$

where M represents the number of space target images.

In addition, we used $ADI-0.1d$ as an evaluation metric, which encodes the percentage of samples with 3D pose errors less than 10% of the object's diameter. Its definition is as follows:

$$ADI-0.1d = \frac{1}{N}\sum_{n=1}^{N}\psi(err_n < 0.1d) \tag{6}$$

where N is the number of samples, err_n is the 3D pose error of the n-th sample, d represents the diameter of the target object, and ψ is the indicator function that returns 1 if the condition inside the parentheses is true and 0 otherwise.

4.2. Comparison with State-of-the-Art Methods

We compared our method with other state-of-the-art pose estimators on the Speed and SwissCube datasets. Additionally, since the main network of our pose estimator is a keypoint detection network, which can also be a standalone task in computer vision, we compared our keypoint detection method with other state-of-the-art keypoint detection methods and integrated them into our PnP algorithm for comparison in pose estimation tasks.

We divided the comparison methods into three categories: keypoint detection methods based on heatmaps, combined with our PnP algorithm for space target pose estimation; keypoint detection methods based on regression, combined with our PnP algorithm for space target pose estimation; and other 6D pose estimation methods for space target pose estimation. We present the results on the Speed and SwissCube test sets in Table 1. It can be observed that our method incorporated more parameters, primarily due to the dual-stream aggregation and SKFM, with enhance feature representations in challenging scenarios where keypoints may be occluded or truncated. Despite having more parameters, the GFLOPs (Giga Floating-Point Operations per Second) of our method remained competitive with those of other approaches and achieved comparable performance due to the efficient operations and overall architecture of our proposed approach. For example, compared to the PoET network on the Speed dataset, our pose estimator was significantly better, with a reduction of 0.5° (5.8° vs. 5.3°) in the rotation error and 0.0084% (0.0351% vs. 0.0267%) in the translation error. Moreover, our keypoint detection network also outperformed all other methods. Compared to keypoint detection methods based on heatmaps, our method achieved at least 3.1% higher AP than other methods (92.1% vs. 95.8%). Since the keypoint detection network is the front-end network of pose estimation, the performance of keypoint detection directly affects the performance of final pose estimation. Therefore, our method reduced the rotation error by 5.2° (10.5° vs. 5.3°) and the translation error by 0.2799% (0.3066% vs. 0.0267%) compared to other heatmap-based keypoint detection methods. Compared to regression-based methods, our method was also significantly better. Compared to the best PRTR, our keypoint detection network improved by 8.6% in AP (87.2% vs. 95.8%), and even when the backbone network was replaced with HRNet-W32, our method was still 5.4% higher in AP. Our method also has a significant advantage compared to other recent methods. AG-Pose [16] aims to establish robust keypoint-level correspondences for better performance on unseen objects; however, it relies heavily on discriminative features and correspondences across varied textures or shapes, in which spacecraft objects lack such cues. Wang et al. [23] leveraged additional diffusion features, but their method cannot be sufficiently generalized to space objects, leading to reduced accuracy.

Table 1. Comparison results with related methods on the Speed and SwissCube datasets. Values in bold refer to the best, and those in italic refer to the second-best.

	Method	Source	Backbone	Params	GFLOPs	$mAP(\%)\uparrow$	$E_q(deg)\downarrow$	$E_T(\%)\downarrow$	$ADI-0.1d(\%)\uparrow$
Speed	HRNet + PnP	CVPR 2019 [2]	HRNet-W32	27.2M	17.1	91.2	11.8	0.3487	69.45
	HRNet + PnP	CVPR 2019 [2]	HRNet-W48	62.4M	37.2	*92.1*	10.5	0.3066	71.10
	LitePose + PnP	CVPR 2022 [67]	ResNet-50	-	-	82.5	20.9	1.0218	63.85
	CPN + PnP	CVPR 2018 [68]	ResNet-50	57.6M	33.1	89.3	13.6	0.3297	68.27
	PointSetNet + PnP	ECCV 2020 [69]	HRNet-W48	-	-	85.0	17.9	0.7483	67.50
	PRTR + PnP	CVPR 2021 [70]	ResNet-50	40.2M	15.3	87.2	16.1	0.6299	68.05
	PRTR + PnP	CVPR 2021 [70]	HRNet-W32	56.1M	28.9	90.4	13.4	0.2306	68.94
	PVNet	CVPR 2019 [38]	-	71.2M	57.8	-	8.4	0.1069	72.06
	CDPN	ICCV 2019 [40]	-	57.9M	35.2	-	6.1	0.0425	75.23
	PoET	PMLR 2023 [71]	-	44.6M	49.5	-	*5.8*	*0.0351*	*75.49*
	AG-Pose	CVPR 2024 [16]	-	42.5M	35.1	-	6.5	0.0441	75.30
	diff-pose	CVPR 2024 [23]	-	1.02B	65.3	-	7.2	0.0712	73.46
	ours	-	ResNet-50	104.3M	39.9	**95.8 (+3.7)**	**5.3 (−0.5)**	**0.0267 (−0.0084)**	**76.15 (+0.66)**
SwissCube	HRNet + PnP	CVPR 2019 [2]	HRNet-W32	23.7M	15.9	90.4	12.7	0.4435	66.79
	HRNet + PnP	CVPR 2019 [2]	HRNet-W48	57.3M	31.6	*90.9*	11.3	0.4186	68.23
	LitePose + PnP	CVPR 2022 [67]	ResNet-50	-	-	81.6	21.6	1.1149	60.53
	CPN + PnP	CVPR 2018 [68]	ResNet-50	50.4M	28.4	88.5	14.1	0.3944	65.24
	PointSetNet + PnP	ECCV 2020 [69]	HRNet-W48	-	-	84.4	18.5	0.7824	64.29
	PRTR + PnP	CVPR 2021 [70]	ResNet-50	38.5M	10.4	86.5	17.3	0.6625	64.83
	PRTR + PnP	CVPR 2021 [70]	HRNet-W32	53.6M	20.7	90.2	13.9	0.2873	65.51
	PVNet	CVPR 2019 [38]	-	66.8M	57.8	-	9.9	0.1229	68.74
	CDPN	ICCV 2019 [40]	-	49.2M	33.1	-	8.5	0.0993	71.88
	PoET	PMLR 2023 [71]	-	40.3M	47.8	-	*8.1*	*0.0744*	*72.13*
	AG-Pose	CVPR 2024 [16]	-	41.1M	34.1	-	8.7	0.0764	70.75
	diff-pose	CVPR 2024 [23]	-	1.00B	63.1	-	8.9	0.0801	69.22
	ours	-	ResNet-50	97.5M	35.2	**95.2 (+4.3)**	**7.2 (−0.9)**	**0.0390 (−0.0354)**	**72.71 (+0.58)**

4.3. Visual Results

To demonstrate the superiority of our method over previous methods, we provide visual comparison results. In Figure 6, the keypoints detected and the estimated pose information are compared across several methods on the Speed dataset. The first column shows the ground-truth keypoints, the second column presents the keypoints detected by our ADSAN, while the third and fourth columns display the keypoints detected by

LitePose and PRTR (using a ResNet-50 backbone). Our method achieved results nearly indistinguishable from the ground truth, demonstrating the ADSAN's robustness to motion blur and noisy space targets. In contrast, LitePose and PRTR, which rely on single-grained feature extraction, faced challenges in capturing both global context and fine local details, particularly in complex space environments. These methods struggled under occlusions, missing critical information when certain keypoints were obscured. Moreover, they did not prioritize challenging keypoints with higher loss, which are often essential for accurate pose recovery in space imagery. As a result, LitePose and PRTR exhibited unavoidable missed detections (red points in Figure 6) and false detections (yellow points in Figure 6). While our ADSAN occasionally had minor deviations, it consistently achieved higher accuracy with fewer missed or false detections. Figure 7 further illustrates the keypoint detection and pose estimation results on the SwissCube dataset, where the ADSAN's ability to adaptively focus on challenging keypoints enabled more robust pose recovery. This emphasizes our method's capacity to address the limitations associated with occlusions and complex spatial contexts, ensuring precise keypoint prediction across datasets.

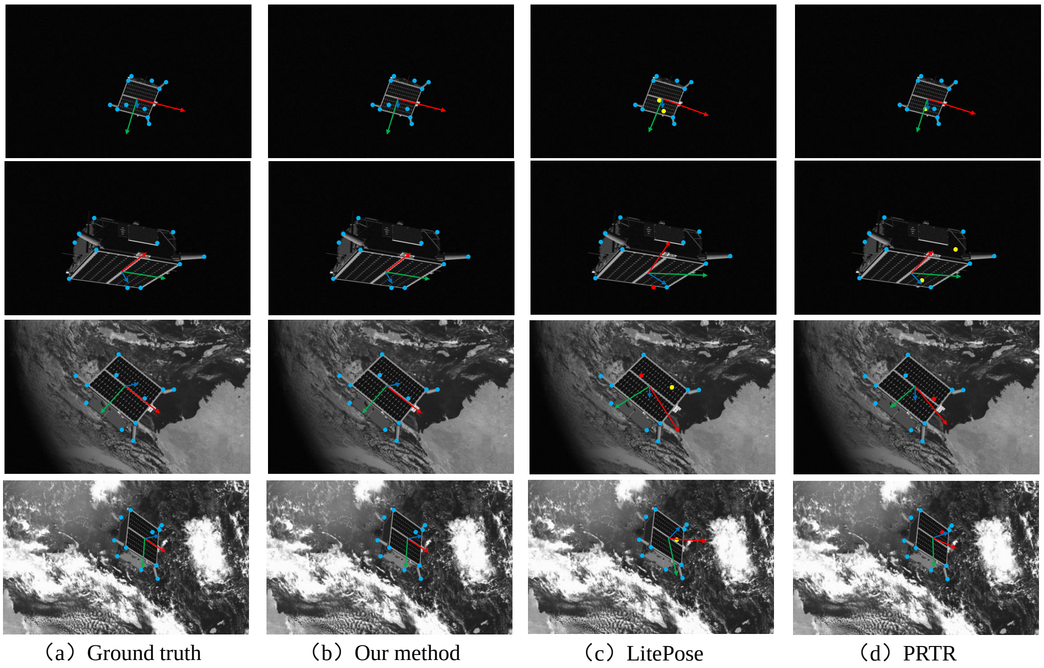


Figure 6. Visual comparison on the Speed dataset. In terms of keypoint detection, our method is nearly identical to the ground truth, while LitePosite and PRTR both suffer from missed detections (red points) and false detections (yellow points).

4.4. Ablation Study

To verify the effectiveness of our method, we conducted ablation experiments on the Speed and SwissCube datasets, respectively. The integration of inter-layer features and intra-layer features provided by the ADSAN and dynamic keypoint focusing aided by the SKFM each contributed to the enhanced performance. Specifically, we used a residual step network as our baseline network. The positions of the keypoints of the space target were predicted by the dual-stream aggregation network, and the selective keypoint focus module was used to assist the dual-stream aggregation network in improving detection accuracy. Table 2 summarizes the quantitative results on the Speed and SwissCube datasets, with the first row displaying the results of the baseline network.

Effectiveness of ADSAN: As shown in Table 2, the baseline network RSN [72] achieved 94.7% AP on the Speed dataset and 93.4% AP on the SwissCube dataset. Compared to RSN, the performance of ADSAN+RSN improved by 0.8% on the Speed dataset and by 1.2% on the SwissCube dataset, indicating that the ADSAN is beneficial for enhancing keypoint detection performance. This is because the ADSAN not only aggregates several finer features within each layer to obtain finer local representations, thereby providing more accurate spatial information to help the network locate keypoints, but also continuously fuses features from different layers to obtain richer global representations, thus gaining deep semantic information to assist the network in predicting keypoints. In addition, during the final pose estimation stage, on the Speed dataset, E_q decreased by 2.4°, E_T decreased by 0.0135%, and $ADI - 0.1d$ increased by 1.65%. The pose estimation performance on the SwissCube dataset also significantly improved.

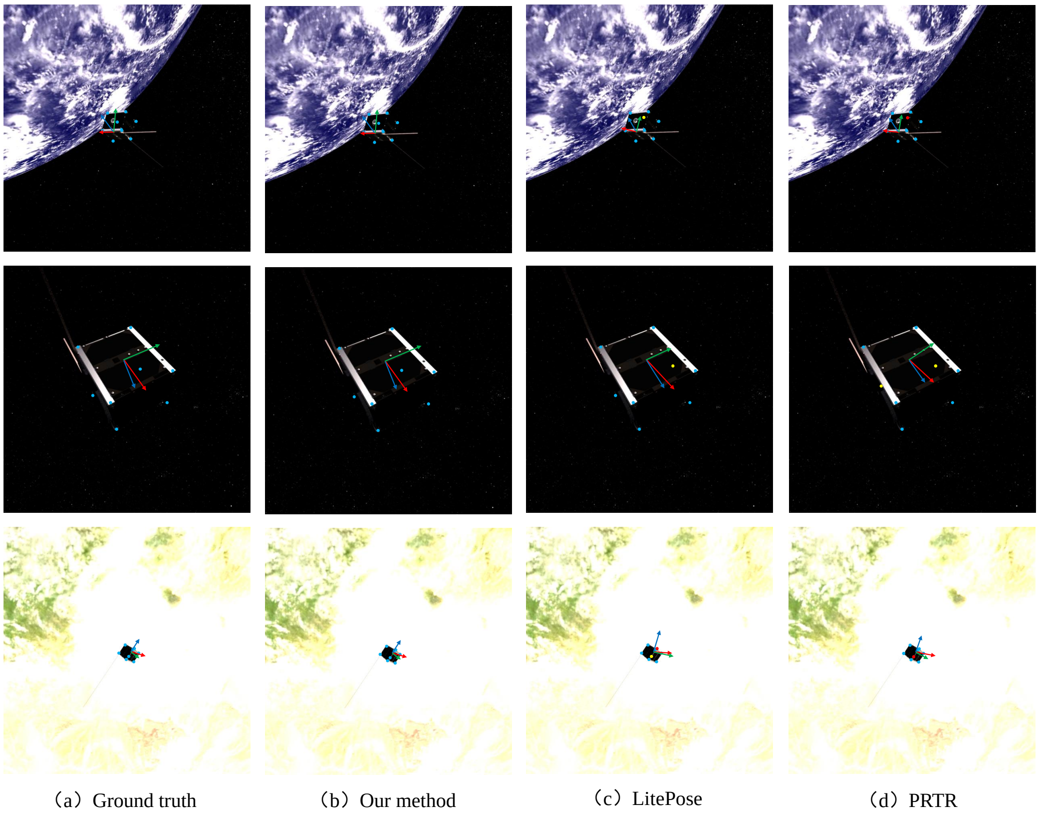


Figure 7. Visual comparison on the SwissCube dataset. Our method shows very small differences compared to the ground truth, while LitePose and PRTR exhibit varying degrees of false detections (yellow points) and missed detections (red points).

Table 2. Experimental results for each component on the Speed and SwissCube datasets.

	ADSAN	SKFM	$mAP(\%)\uparrow$	$E_q(deg)\downarrow$	$E_T(\%)\downarrow$	$ADI-\mathbf{0.1}d(\%)\uparrow$
Speed			94.7	8.7	0.0460	73.89
	√		95.5	6.3	0.0325	75.54
		√	95.1	7.0	0.0397	74.72
	√	√	95.8	5.3	0.0267	76.15
SwissCube			93.4	9.3	0.0533	71.79
	√		94.6	8.1	0.0429	72.83
		√	94.2	8.6	0.0484	72.35
	√	√	95.2	7.2	0.0390	73.22

Effectiveness of SKFM: When the SKFM was included, the performance of RSN+SKFM improved to 95.1% on the Speed dataset and 94.2% on the SwissCube dataset compared to the baseline network RSN. This indicates that the SKFM can identify challenging keypoints of space targets, thereby addressing interference caused by truncation and viewpoint changes. Through this strategy, which focuses on challenging keypoints, our keypoint detection network achieved high-precision detection of keypoints on space targets.

4.5. Interference Resistance Experiment

To validate the robustness of our method against motion blur and noise interference, we conducted interference resistance experiments on the Speed dataset by processing noise and adding motion blur, thus creating the more realistic Speed-noise dataset. As shown in Figure 8a, the original simulated image had minimal noise interference, with ideal image quality even in the presence of an Earth background. By simulating noise and motion blur on the original simulated images, as shown in Figure 8b, their texture information was significantly reduced, making them closer to real-world data.

Comparison with State-of-the-Art Methods: To demonstrate the superiority of our method over other methods on the Speed-noise dataset, we conducted comparative experiments, as shown in Table 3. Consistent with the Speed and SwissCube datasets, the comparative experiments on the Speed-noise dataset were divided into the same three categories: heatmap-based keypoint detection methods with PnP geometric algorithms, regression-based keypoint detection methods with PnP geometric algorithms, and 6D pose estimation algorithms without independent keypoint detection networks. The numbers in parentheses in the table represent the changes compared to the corresponding results in Table 1. It can be observed that the change in the mAP was 0, indicating that the ADSAN achieved the same keypoint detection performance on both the Speed-noise and Speed datasets. This can be attributed to the dual fusion of inter-layer and intra-layer features, which obtained more robust information and overcame interference from noise and motion blur. Additionally, the ADSAN significantly outperformed other methods, with the smallest change value, indicating its superior anti-interference performance.

Table 3. Comparison results with related methods on the Speed-noise dataset. Values in bold refer to the best, and those in italic refer to the second-best.

Method	Source	Backbone	$mAP(\%)\uparrow$	$E_q(deg)\downarrow$	$E_T(\%)\downarrow$	$ADI-0.1d(\%)\uparrow$
HRNet + PnP	CVPR 2019 [2]	HRNet-W32	90.2(−1)	13.9(+2.1)	0.3729(+0.0242)	59.31(−10.14)
HRNet + PnP	CVPR 2019 [2]	HRNet-W48	*91.5(−0.6)*	13.1(+2.6)	0.3242(+0.0176)	61.89(−9.21)
LitePose + PnP	CVPR 2022 [67]	ResNet-50	81.6(−0.9)	22.3(+1.4)	1.3118(+0.2900)	55.26(−8.59)
CPN + PnP	CVPR 2018 [68]	ResNet-50	87.9(−1.4)	16.4(+2.8)	0.3519(+0.0222)	59.05(−9.22)
PointSetNet + PnP	ECCV 2020 [69]	HRNet-W48	84.2(−0.8)	19.5(+1.6)	0.7989(+0.0506)	58.77(−8.73)
PRTR + PnP	CVPR 2021 [70]	ResNet-50	86.0(−1.2)	18.8(+2.7)	0.6962(+0.0633)	59.03(−9.02)
PRTR + PnP	CVPR 2021 [70]	HRNet-W32	89.7(−0.7)	16.0(+2.6)	0.3157(+0.0851)	59.28(−9.66)
PVNet	CVPR 2019 [38]	-	-	10.3(+1.9)	0.1663(+0.0594)	62.75(-9.85)
CDPN	ICCV 2019 [40]	-	-	8.7(+2.6)	0.0605(+0.0180)	65.90(-9.33)
PoET	PMLR 2023 [71]	-	-	*8.6(+2.8)*	*0.0496(+0.0145)*	*66.04(-9.45)*
AG-Pose	CVPR 2024 [16]	-	-	9.5(+3.0)	0.0607(+0.0166)	65.8(-9.50)
diff-pose	CVPR 2024 [23]	-	-	9.8(+2.6)	0.0881(+0.0169)	63.55(-9.91)
ours	-	ResNet-50	**95.8(−0)**	**7.5(+2.2)**	**0.0376(+0.0109)**	**68.80(−7.35)**

Visual Results: In Figure 8, we present the visual comparison results on the Speed-noise dataset. It can be seen that our predicted keypoints and pose results are consistent with the ground truth because the ADSAN extracted rich multi-grained features, enabling more accurate keypoint detection. Additionally, the SKFM focused on challenging keypoints, further enhancing keypoint detection performance. Due to keypoint misdetection, LitePose's recovered pose deviated somewhat from the ground truth. Similarly, PRTR exhibited deviations from the ground truth due to missed detections.

Ablation Study: To validate the effective improvement of our method in addressing motion blur and noise interference issues, we conducted ablation experiments on the Speed-noise dataset. As shown in Table 4, under the detection of the dual-stream aggregation network, the average precision of keypoint detection reached 95.1%, an increase of 1.9% compared to the baseline network. Additionally, E_q decreased from 10.6° to 8.5°, E_T decreased from 0.0677% to 0.0483%, and the $ADI - 0.1d$ index also showed a significant improvement. Furthermore, by applying the selective keypoint focus module to handle truncation points and occluded points, the final performance of the keypoint detection network reached 95.8%, with E_q at 7.5° and E_T at 0.0376%, resulting in a significant improvement in accuracy compared to the baseline network. Thus, each component of our method has a mitigating effect on motion blur and noise interference issues.

Table 4. Experimental results for each component on the Speed-noise dataset.

ADSAN	SKFM	$mAP(\%) \uparrow$	$E_q(deg) \downarrow$	$E_T(\%) \downarrow$	$ADI - 0.1d(\%) \uparrow$
		93.2	10.6	0.0677	65.32
√		95.1	8.5	0.0483	67.97
	√	94.6	9.1	0.0574	67.05
√	√	95.8	7.5	0.0376	68.80

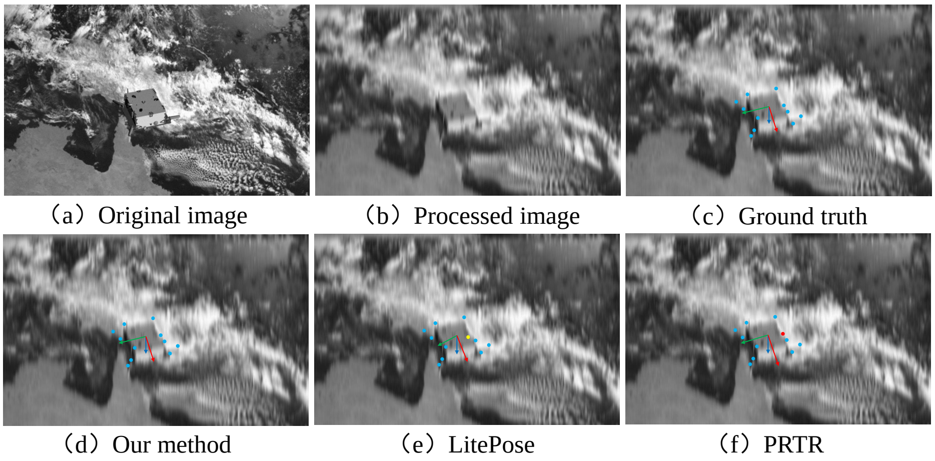


Figure 8. Comparison of Speed-noise images and their interference resistance visualization results. (**a**) Original image, with clear image quality and visible target texture. (**b**) Processed image, which is blurry with noise present and significantly reduced image quality. (**c**) Ground truth of the processed image. (**d**) Keypoints and pose information predicted by our method. (**e**) Keypoints and pose information predicted by LitePose. (**f**) Keypoints and pose information predicted by PRTR.

5. Conclusions

In this paper, we propose an adaptive dual-stream aggregation network (ADSAN) for pose estimation of space targets in real images. The dual-stream aggregation keypoint detection structure fuses feature maps of different sizes between layers and finer features within layers to overcome the limitations of single-grained features. We further introduce a selective keypoint focus module to improve detection accuracy by addressing challenges like truncation and viewpoint changes. Additionally, we validate the ADSAN's robustness to noise and motion blur in real-world scenarios. Extensive experiments on challenging datasets, including Speed, SwissCube, and Speed-noise, demonstrate that our method achieves state-of-the-art pose estimation performance for space target images.

In future work, we plan to extend our approach to other datasets, including those featuring diverse satellite types and imaging conditions, to assess the ADSAN's adaptability.

Furthermore, we aim to optimize the network for real-time applications and investigate methods to enhance its robustness under extreme environmental factors encountered in space. These advancements will help broaden the applicability of the ADSAN to a wider range of space exploration tasks.

Author Contributions: Methodology, X.G., X.Y. and D.Y.; Software, X.G.; Formal analysis, X.G., X.Y. and D.Y.; Writing—original draft, X.G. and X.Y.; Writing—review and editing, H.L. and D.Y.; Funding acquisition, X.Y.; Data curation, H.L.; Visualization, H.L.; Supervision, X.Y. and D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grant 62372348; in part by the Key Research and Development Program of Shaanxi under grant 2024GXZDCYL-02-10; in part by the Shaanxi Outstanding Youth Science Fund Project under grant 2023-JC-JQ-53; and in part by the Fundamental Research Funds for the Central Universities under grant QTZX24080.

Data Availability Statement: Publicly available datasets were analyzed in this study. The Speed dataset can be found here (https://kelvins.esa.int/satellite-pose-estimation-challenge/home/ (accessed on 4 November 2024)). The Swisscube dataset can be found here (https://github.com/cvlab-epfl/wide-depth-range-pose) (accessed on 4 November 2024).

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

References

1. Lei, X.; Lao, Z.; Liu, L.; Chen, J.; Wang, L.; Jiang, S.; Li, M. Telescopic Network of Zhulong for Orbit Determination and Prediction of Space Objects. *Remote Sens.* **2024**, *16*, 2282. [CrossRef]
2. Chen, B.; Cao, J.; Parra, A.; Chin, T.J. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
3. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
4. Yang, X.; Nan, X.; Song, B. D2N4: A discriminative deep nearest neighbor neural network for few-shot space target recognition. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3667–3676. [CrossRef]
5. Tian, X.; Bai, X.; Zhou, F. Recognition of micro-motion space targets based on attention-augmented cross-modal feature fusion recognition network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5104909. [CrossRef]
6. Wang, J.; Li, G.; Zhao, Z.; Jiao, J.; Ding, S.; Wang, K.; Duan, M. Space target anomaly detection based on Gaussian mixture model and micro-Doppler features. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5118411. [CrossRef]
7. Zhao, M.; Li, S.; Wang, H.; Yang, J.; Sun, Y.; Gu, Y. MP 2 Net: Mask Propagation and Motion Prediction Network for Multi-Object Tracking in Satellite Videos. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5617515.
8. Chen, Z.; Shang, Y.; Python, A.; Cai, Y.; Yin, J. DB-BlendMask: Decomposed attention and balanced BlendMask for instance segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615915. [CrossRef]
9. Huo, Y.; Li, Z.; Zhang, F. Fast and accurate spacecraft pose estimation from single shot space imagery using box reliability and keypoints existence judgments. *IEEE Access* **2020**, *8*, 216283–216297. [CrossRef]
10. Redmon, J. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
11. Huan, W.; Liu, M.; Hu, Q. Pose estimation for non-cooperative spacecraft based on deep learning. In Proceedings of the 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 3339–3343.
12. Lotti, A.; Modenini, D.; Tortora, P. Investigating vision transformers for bridging domain gap in satellite pose estimation. In Proceedings of the International Conference on Applied Intelligence and Informatics, Reggio Calabria, Italy, 1–3 September 2022; pp. 299–314.
13. 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; pp. 10012–10022.
14. Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; Navab, N. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In Proceedings of the IEEE/CVF Asian Conference on Computer Vision, Daejeon, Republic of Korea, 5–9 November 2012; pp. 548–562.
15. Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; Savarese, S. DenseFusion: 6D object pose estimation by iterative dense fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3343–3352.

16. Lin, X.; Yang, W.; Gao, Y.; Zhang, T. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 21040–21049.
17. Wang, G.; Manhardt, F.; Tombari, F.; Ji, X. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16611–16621.
18. Li, F.; Vutukur, S.R.; Yu, H.; Shugurov, I.; Busam, B.; Yang, S.; Ilic, S. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2123–2133.
19. Fan, Z.; Song, Z.; Xu, J.; Wang, Z.; Wu, K.; Liu, H.; He, J. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In Proceedings of the IEEE/CVF European Conference on Computer Vision, New Orleans, LA, USA, 18–24 June 2022; pp. 220–236.
20. Legrand, A.; Detry, R.; De Vleeschouwer, C. Domain Generalization for 6D Pose Estimation Through NeRF-based Image Synthesis. *arXiv* **2024**, arXiv:2407.10762.
21. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
22. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
23. Wang, T.; Hu, G.; Wang, H. Object pose estimation via the aggregation of diffusion features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 10238–10247.
24. Lu, C.P.; Hager, G.D.; Mjolsness, E. Fast and globally convergent pose estimation from video images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 610–622. [CrossRef]
25. Tulsiani, S.; Malik, J. Viewpoints and keypoints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1510–1519.
26. Pavlakos, G.; Zhou, X.; Chan, A.; Derpanis, K.G.; Daniilidis, K. 6-Dof object pose from semantic keypoints. In Proceedings of the IEEE International Conference on Robotics and Automation, Singapore, 29 May–3 June 2017; pp. 2011–2018.
27. Fan, R.; Xu, T.B.; Wei, Z. Estimating 6D Aircraft Pose from Keypoints and Structures. *Remote Sens.* **2021**, *13*, 663. [CrossRef]
28. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
29. Tola, E.; Lepetit, V.; Fua, P. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 815–830. [CrossRef]
30. Trzcinski, T.; Christoudias, M.; Lepetit, V.; Fua, P. Learning image descriptors with the boosting-trick. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25.
31. Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; Navab, N. SSD-6D: Making rgb-based 3d detection and 6d pose estimation great again. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1521–1529.
32. Rad, M.; Lepetit, V. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3828–3836.
33. Tekin, B.; Sinha, S.N.; Fua, P. Real-time seamless single shot 6d object pose prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 292–301.
34. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv* **2017**, arXiv:1711.00199.
35. Oberweger, M.; Rad, M.; Lepetit, V. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Salt Lake City, UT, USA, 18–22 June 2018; pp. 119–134.
36. Hosseini Jafari, O.; Mustikovela, S.K.; Pertsch, K.; Brachmann, E.; Rother, C. iPose: Instance-aware 6d pose estimation of partly occluded objects. In Proceedings of the IEEE/CVF Asian Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 477–492.
37. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-driven 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3385–3394.
38. Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; Bao, H. Pvnet: Pixel-wise voting network for 6dof pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4561–4570.
39. Zakharov, S.; Shugurov, I.; Ilic, S. DPOD: 6D pose object detector and refiner. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1941–1950.
40. Li, Z.; Wang, G.; Ji, X. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7678–7687.
41. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]

42. Sharma, S.; Beierle, C.; D'Amico, S. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. In Proceedings of the Aerospace Conference, Big Sky, MT, USA, 3–10 March 2018; pp. 1–12.
43. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
44. Harvard, A.; Capuano, V.; Shao, E.Y.; Chung, S.J. Spacecraft pose estimation from monocular images using neural network based keypoints and visibility maps. In Proceedings of the AIAA Scitech Forum, Orlando, FL, USA, 6–10 January 2020; p. 1874.
45. Zhao, Y.; Chen, L.; Zhang, X.; Xu, S.; Bu, S.; Jiang, H.; Han, P.; Li, K.; Wan, G. RTSFM: Real-time structure from motion for mosaicing and dsm mapping of sequential aerial images with low overlap. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607415. [CrossRef]
46. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded up robust features. *Lect. Notes Comput. Sci.* **2006**, *3951*, 404–417.
47. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
48. Gkioxari, G.; Toshev, A.; Jaitly, N. Chained predictions using convolutional neural networks. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Las Vegas, LV, USA, 26 June–1 July 2016; pp. 728–743.
49. Lifshitz, I.; Fetaya, E.; Ullman, S. Human pose estimation using deep consensus voting. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Las Vegas, LV, USA, 26 June–1 July 2016; pp. 246–260.
50. Tang, W.; Yu, P.; Wu, Y. Deeply learned compositional models for human pose estimation. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Salt Lake City, UT, USA, 18–22 June 2018; pp. 190–206.
51. Nie, X.; Feng, J.; Yan, S. Mutual learning to adapt for joint human parsing and pose estimation. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Salt Lake City, UT, USA, 18–22 June 2018; pp. 502–517.
52. Nie, X.; Feng, J.; Zuo, Y.; Yan, S. Human pose estimation with parsing induced learner. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2100–2108.
53. Peng, X.; Tang, Z.; Yang, F.; Feris, R.S.; Metaxas, D. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2226–2234.
54. Sun, K.; Lan, C.; Xing, J.; Zeng, W.; Liu, D.; Wang, J. Human pose estimation using global and local normalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5599–5607.
55. Fan, X.; Zheng, K.; Lin, Y.; Wang, S. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1347–1355.
56. Ye, R.; Ren, Y.; Zhu, X.; Wang, Y.; Liu, M.; Wang, L. An Efficient Pose Estimation Algorithm for Non-Cooperative Space Objects Based on Dual-Channel Transformer. *Remote Sens.* **2023**, *15*, 5278. [CrossRef]
57. Toshev, A.; Szegedy, C. DeepPose: Human pose estimation via deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1653–1660.
58. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human pose estimation with iterative error feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4733–4742.
59. Zhang, S.; Fu, Z.; Liu, J.; Su, X.; Luo, B.; Nie, H.; Tang, B.H. Multilevel attention Siamese network for keypoint detection in optical and SAR images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5404617. [CrossRef]
60. Cao, J.; You, Y.; Li, C.; Liu, J. TSK: A Trustworthy Semantic Keypoint Detector for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5607120. [CrossRef]
61. Chu, X.; Ouyang, W.; Li, H.; Wang, X. Structured feature learning for pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4715–4723.
62. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 1831–1840.
63. Yang, W.; Ouyang, W.; Li, H.; Wang, X. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3073–3082.
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
65. Kisantal, M.; Sharma, S.; Park, T.H.; Izzo, D.; Märtens, M.; D'Amico, S. Satellite pose estimation challenge: Dataset, competition design, and results. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *56*, 4083–4098. [CrossRef]
66. Hu, Y.; Speierer, S.; Jakob, W.; Fua, P.; Salzmann, M. Wide-depth-range 6d object pose estimation in space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15870–15879.
67. Wang, Y.; Li, M.; Cai, H.; Chen, W.M.; Han, S. Lite pose: Efficient architecture design for 2d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13126–13136.
68. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.

69. Wei, F.; Sun, X.; Li, H.; Wang, J.; Lin, S. Point-set anchors for object detection, instance segmentation and pose estimation. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Seattle, WA, USA, 14–19 June 2020; pp. 527–544.
70. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose recognition with cascade transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1944–1953.
71. Jantos, T.G.; Hamdad, M.A.; Granig, W.; Weiss, S.; Steinbrener, J. PoET: Pose estimation transformer for single-view, multi-object 6D pose estimation. In Proceedings of the Conference on Robot Learning, Atlanta, GA, USA, 6–9 November 2023; pp. 1060–1070.
72. Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; Sun, J. Learning delicate local representations for multi-person pose estimation. In Proceedings of the IEEE/CVF European Conference on Computer Vision, Seattle, WA, USA, 14–19 June 2020; pp. 455–472.

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.