

Article

On-Orbit 3D Reconstruction and Pose Measurement of Space Non-Cooperative Satellites Using Monocular Vision

Xingguang Qu ¹, Zhen Liu ¹, Xiao Pan ^{1,*}, Qiming Liu ¹, Jiuzheng Song ¹ and Bo Feng ²

¹ Ministry of Education Key Laboratory of Precision Opto-Mechatronics Technology, Beihang University, Beijing 100191, China; qu287165037@buaa.edu.cn (X.Q.); liuzhen008@buaa.edu.cn (Z.L.); ywxkgsgysm_leap@outlook.com (Q.L.); songjiuzheng@buaa.edu.cn (J.S.)

² Equipment Management and UAV Engineering College, Air Force Engineering University, Xi'an 710051, China; fengbo876@163.com

* Correspondence: panxiao0620541@buaa.edu.cn

Highlights

- End-to-end MVS network MambaMVS integrating structured state space models into the multi-view stereo pipeline for dense 3D reconstruction.
- MCV Module leveraging Mamba's linear complexity for efficient cost volume regularization during coarse-to-fine depth map generation.
- Improved focused linear attention-based feature matching block enhancing cross-view feature representation on complex satellite surfaces.
- Few-shot pose measurement method integrating 3D Gaussian Splatting augmentation and reprojection-based automatic annotation for on-orbit applications.

Abstract

Dense 3D reconstruction and accurate pose measurement of non-cooperative satellites are essential for on-orbit servicing missions. Multi-view stereo (MVS) is the preferred approach for high-precision dense reconstruction from monocular imagery, yet current methods struggle with the speed–accuracy trade-off and reconstruction degradation under complex on-orbit conditions, while keypoint-based pose measurement is further constrained by limited training data and heavy reliance on manual annotations. To address these challenges, we present a monocular vision-based method coupling dense 3D reconstruction with pose estimation. We develop MambaMVS, an end-to-end MVS network featuring a Mamba Cost Volume (MCV) Module for efficient cost volume regularization via the linear complexity of structured state space models, and an improved focused linear attention module for robust cross-view feature matching. The reconstructed point cloud provides 3D keypoint coordinates for pose measurement, where a few-shot scheme employing 3D Gaussian Splatting-based view augmentation and reprojection-driven self-labeling enables training with minimal manual annotations. During online inference, detected 2D keypoints are combined with offline-reconstructed 3D coordinates via the Perspective-n-Point algorithm for fast solving of the 6-DOF pose. Experiments on the Customized Space Target dataset demonstrate state-of-the-art reconstruction quality with 1.5× faster inference than current leading methods. Physical satellite model experiments achieve rotation ME within 0.2° and translation ME below 0.01 m, while robustness is validated across varying illumination, reflective surfaces, motion blur, and occlusion. The proposed reconstruction-guided measurement paradigm offers a practical pathway toward autonomous on-orbit perception and in situ structural assessment.

Academic Editor: Stefano Silvestrini

Received: 14 August 2026

Revised: 24 September 2026

Accepted: 29 September 2026

Published: 30 September 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: dense reconstruction; monocular vision; non-cooperative satellites; pose measurement

1. Introduction

On-orbit servicing missions, including satellite life extension, component replacement, active debris removal, and on-orbit assembly, require precise perception of target spacecraft to ensure safe and reliable operations. These missions also involve mission-level design problems, such as the planning of finite-thrust transfer trajectories [1] and the scheduling of relay satellites for multi-type tasks [2]; the present work addresses the perception layer that such operations rely on. Non-cooperative targets, such as malfunctioning satellites and space debris that lack dedicated vision markers or communication interfaces, pose particular difficulty: their geometry, pose, and dynamic state must be inferred entirely from sensor observations. Among available sensors, the monocular camera stands out for its compact form factor, low mass, and power efficiency, which are critical for resource-constrained spacecraft platforms. Fast and accurate 3D reconstruction from monocular imagery enables structural evaluation and defect assessment, while precise pose measurement provides the dynamic reference essential for autonomous approach and manipulation [3,4].

Dense 3D reconstruction methods for space targets fall into laser-based and image-based categories. Lidar directly acquires depth but cannot capture surface texture, is material-dependent, and imposes mass and power penalties incompatible with resource-constrained spacecraft [5]. Image-based methods instead reconstruct geometry from monocular multi-view images [6–8]. Neural radiance fields [9–11] achieve compelling visual fidelity yet optimize for novel view synthesis rather than geometric accuracy. Their implicit representations lack explicit depth, and extracting metric-accurate point clouds introduces artifacts through postprocessing. These methods further require per-scene optimization and substantial computation, precluding immediate on-orbit deployment. Alternative approaches each carry specific limitations for onboard use: SLAM-based methods [12,13] are complex and drift-prone and operate at short range; stereo [14] needs a fixed, well-calibrated baseline and loses depth accuracy with range; and structured light [15] requires an active projector and works only at short range. They therefore trade geometric precision for real-time performance or hardware simplicity. Multi-view stereo (MVS) addresses these limitations by explicitly estimating per-pixel depth through cost volume matching computed on deep features rather than raw pixel intensities, directly yielding dense point clouds with metric accuracy. Its feed-forward architecture generalizes across scenes without per-scene training, delivering the high-fidelity geometry that downstream pose measurement demands. When geometric precision is the primary requirement, MVS provides the most suitable solution.

In the target application, MVS-based dense reconstruction is performed on multi-view images captured by a monocular camera aboard a chaser spacecraft circumnavigating non-cooperative satellites at close proximity. The reconstructed point clouds provide 3D keypoint coordinates and enable training data generation for subsequent pose measurement. This workflow adopts an offline–online paradigm. During the offline phase, the reconstruction and the training of the keypoint model are completed for each target. During online operation, the pre-trained models enable real-time pose estimation without reprocessing. However, harsh and variable lighting, complex satellite surface materials, and limited images captured during high-speed motion pose significant challenges to reconstruction quality [16]. The downstream pipeline further requires highly reliable point clouds for accurate 3D keypoint extraction, sufficient training samples for robust

detection, and precise 2D–3D correspondences for reliable pose solving. These requirements motivate an enhanced MVS framework that improves reconstruction accuracy under adverse conditions, enables efficient training data generation from limited observations, and supports end-to-end pose measurement with minimal manual intervention.

To address these challenges, we propose a unified framework for on-orbit 3D reconstruction and pose measurement of space non-cooperative satellites using monocular vision. The framework is built upon MambaMVS, an end-to-end MVS network that integrates structured state space models (SSMs) into the multi-view stereo pipeline for efficient dense reconstruction. The network features a Mamba Cost Volume (MCV) Module that exploits the linear complexity of SSMs [17] for cost volume regularization and an improved focused linear attention module [18] for enhanced cross-view feature matching. The offline–online paradigm links reconstruction with pose measurement. During the offline phase, MambaMVS reconstructs a dense point cloud from which 3D semantic keypoint coordinates are extracted. To overcome the limited training data [19], 3D Gaussian Splatting (3DGS) [10] augments the training set from the reconstruction, and a reprojection-based self-labeling approach automatically propagates keypoint annotations with minimal manual effort, enabling robust training of a keypoint detection network. During the online operation, the trained network localizes 2D keypoints in new images, and the Perspective-n-Point solver combines them with the pre-computed 3D coordinates for real-time 6-DOF pose estimation. Therefore, our contributions are as follows.

- (1) We propose MambaMVS, an end-to-end MVS network for dense 3D reconstruction from monocular images, in which structured state space models are employed for cost volume regularization. Experiments on satellite targets show that this design achieves a more favorable speed–accuracy trade-off than both 3D CNN-based and attention-based regularizers.
- (2) We propose the MCV Module, which leverages the linear complexity of SSMs for efficient cost volume regularization, together with an improved focused linear attention-based feature matching block that mitigates performance degradation in cross-view feature extraction.
- (3) We propose a satellite pose measurement method tailored for data-scarce scenarios, which leverages the dense point cloud reconstructed by MambaMVS to provide 3D keypoint coordinates. By integrating 3D Gaussian Splatting-based novel-view augmentation with reprojection-based automatic annotation, this method reduces the number of manually annotated images while preserving the accuracy of fully manual annotation and improves it further under adverse imaging conditions.
- (4) Extensive experiments demonstrate that MambaMVS achieves state-of-the-art reconstruction quality in terms of point cloud accuracy, dimensional measurement, and plane fitting error while operating 1.5× faster than current leading methods. Physical satellite experiments confirm accurate 6-DOF pose estimation, and robustness is validated across diverse challenging conditions, including varying illumination, reflective surfaces and motion blur.

The nomenclature of the key variables is provided in Table 1. The remainder of this article is organized as follows. Section 2 reviews related work. Section 3 details the proposed method. Section 4 presents experimental results and analysis. Section 5 concludes with future work.

Table 1. Nomenclature.

Symbol	Meaning
I_0	Reference image
$\{I_i\}_{i=1}^N$	Source images
$\{F_i\}_{i=0}^N$	Feature maps

$\{\mathbf{V}_i\}_{i=0}^N$	4D feature volumes
$\mathbf{C} \in \mathbb{R}^{C \times D \times H \times W}$	Cost volume
$\mathbf{P} \in \mathbb{R}^{D \times H \times W}$	Probability volume
FPN	Feature Pyramid Network
DINO	Self-DIstillation with NO labels
FLA	Focused linear attention-based
FSA	Focused Self-Attention
FCA	Focused Cross-Attention
SSMs	State space models

2. Related Work

Dense 3D reconstruction and pose estimation for on-orbit non-cooperative satellites require advances in multi-view stereo, efficient sequence modeling, and data-efficient pose measurement.

2.1. Learning-Based Multi-View Stereo

MVSNet [20] pioneered deep learning-based MVS by leveraging deep features and differentiable homography warping to construct cost volumes for depth regression. Subsequent works [14,21,22] adopt coarse-to-fine strategies to balance accuracy and efficiency by refining depth hypotheses iteratively. Recent architectures enhance feature matching through transformer-based mechanisms [23]. TransMVSNet [18] introduces the Feature Matching Transformer (FMT), MVSTER [24] leverages the Epipolar Transformer, and PA-MVSNet [25] employs multi-scale pyramid attention [26], while MVSFormer [27] and MVSFormer++ [28] augment the FPN with a pre-trained ViT and employ efficient attention mechanisms such as FlashAttention [29] to leverage large-scale dataset priors. While linear attention reduces computational complexity [30], it compromises performance on complex texture surfaces commonly found in space targets. For cost volume regularization, traditional 3D U-Nets [20,26,31] suffer from limited receptive fields, while transformer-based methods such as Costformer [32] still face quadratic complexity with 4D cost volumes. Therefore, designing a cost volume regularization module that achieves fast inference while maintaining performance remains a valuable research direction.

2.2. State Space Models in Computer Vision

State space models (SSMs) exhibit linear scaling with sequence length, defined as the number of tokens obtained by flattening a feature map or volume, and benefit from parallel training, showing great potential for long-sequence modeling. Mamba [17] integrates a selection mechanism and hardware-aware designs, outperforming transformers in modeling long-range dependencies while maintaining linear complexity. Vision Mamba [33] adapts SSMs to computer vision via bidirectional models and cross-scanning modules, inspiring applications in tracking [34], segmentation [35], point cloud processing [36,37], and 3D reconstruction [38]. State space models have also been applied to multi-view stereo, including Mamba-based aggregation of intra- and inter-view features [39], cost aggregation for space target reconstruction [40], and Mamba-based feature enhancement combined with 3D cost volume regularization [41]. While these methods focus on the reconstruction network, the present work applies a state space model to cost volume regularization, employs focused linear attention for cross-view feature matching, and further couples the reconstruction with a keypoint-based pose measurement pipeline, as detailed in Section 3.

2.3. Pose Estimation for Space Targets

Accurate relative pose estimation between spacecraft is a fundamental requirement for on-orbit servicing missions. Monocular passive-camera sensing for non-cooperative targets predates the deep learning-based methods reviewed below. Volpe et al. [42] addressed pose and shape determination from passive imagery, and the same group later extended this line to a guidance, navigation and control architecture for rendezvous with a tumbling target [43]. Recent works have explored diverse approaches: PVSPE [19] employs a pyramid vision multitask transformer for simultaneous spacecraft segmentation and pose estimation from monocular images; Atallah et al. [44] propose a twistor-based unscented Kalman filter for robust filtering under noisy observations; and SPTN [45] develops a transformer-based tracking network capturing long-range temporal dependencies. These methods perform end-to-end pose estimation directly from image observations.

In contrast, our approach first obtains high-precision dense 3D reconstruction via MambaMVS, providing accurate 3D coordinates for semantic keypoints on the satellite structure. Combined with a deep learning-based keypoint detection network and the Perspective-n-Point (PnP) solver, this reconstruction-then-measurement pipeline decouples geometric reconstruction from pose solving, enabling more interpretable and verifiable measurement. The 3D keypoint coordinates derived from dense reconstruction serve as a reliable geometric foundation independent of single-frame inference, improving measurement accuracy under limited training data conditions.

3. Proposed Method

3.1. Overall Architecture

As illustrated in Figure 1, the proposed method comprises three main stages: (1) MambaMVS for dense 3D reconstruction, (2) offline training for keypoint detection, and (3) online pose measurement. In the reconstruction stage, MambaMVS takes a reference image and multiple source images with known camera parameters as input. Through a coarse-to-fine multi-stage pipeline of feature extraction enhanced with the FLA Module, cost volume construction, and Mamba-based cost volume regularization (MCV Module), the network estimates per-pixel depth maps at multiple scales. Dense point clouds are obtained by fusing depth maps across all input views. The detailed architecture of MambaMVS is described in Sections 3.2 and 3.3. For pose measurement, the dense point cloud and camera poses from the MVS pipeline provide 3D coordinates of predefined semantic keypoints. To address limited training data, 3D Gaussian Splatting (3DGS) renders augmented views from the reconstructed model, and a reprojection-based self-labeling approach automatically propagates keypoint annotations across the image set using the known camera poses, requiring manual labeling of only a small subset. A modified RTMpose network is then trained to detect 2D keypoint coordinates in satellite images. During online inference, the detected 2D coordinates are combined with the known 3D keypoint positions via the Perspective-n-Point (PnP) algorithm to solve for the six-degrees-of-freedom (6-DoF) pose. The pose measurement pipeline is detailed in Section 3.5.

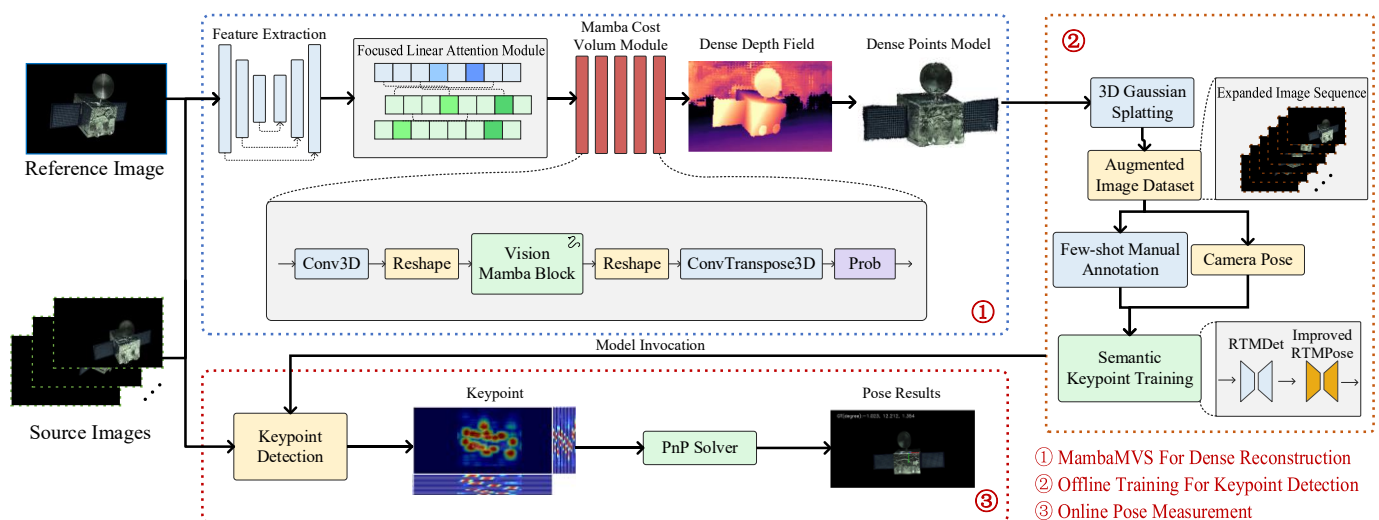


Figure 1. Overall structure of the proposed method.

3.2. Enhanced Feature Extraction

Unlike the closely related regularizer of Li et al. [41], which unfolds the cost volume into three 2D cost volumes along the height, width and depth axes, our MCV Module flattens the volume into a single sequence and applies a bidirectional selective scan to it, so that depth and spatial dependencies are modeled jointly.

MVS is inherently a one-to-many matching task, in which each pixel in the reference image searches along epipolar lines in the source images across plane-sweep depth hypotheses to determine the optimal depth corresponding to the minimum matching cost. Thus, extracting more discriminative multi-scale hierarchical features for matching is crucial. Given a reference image and N source images, the network first performs multi-scale feature extraction using a shared-weight Feature Pyramid Network (FPN); a bottom-up pathway downsamples the feature maps to $1/2$, $1/4$ and $1/8$ of the original resolution. This is followed by a top-down pathway that upsamples features to higher resolutions, integrated with lateral connections that fuse these upsampled features with corresponding low-level features from the bottom-up pathway. Following the approach in MVSFormer++ [28], we employ a pre-trained frozen DINOv2-Base model [46] to complement the FPN and enrich visual features. The above process can be formulated as follows

$$\{\mathbf{F}_i\}_{i=0}^N = \{\text{FPN}(\mathbf{I}_i) + \text{DINO}(\mathbf{I}_i)\}_{i=0}^N \quad (1)$$

To enhance the model's ability to represent image-related features and perceive global features, while enabling more efficient computation when processing long sequences, we introduce a focused linear attention-based module, termed the FLA Module. Specifically, to enhance feature representations both within and across views in the input image sequence, we design two components, a Focused Self-Attention Block (FSA Block) and a Cross-Attention Block (FCA Block), which respectively handle intra-attention and inter-attention. The network architecture of the FSA Block and the FCA Block is shown in Figure 2.

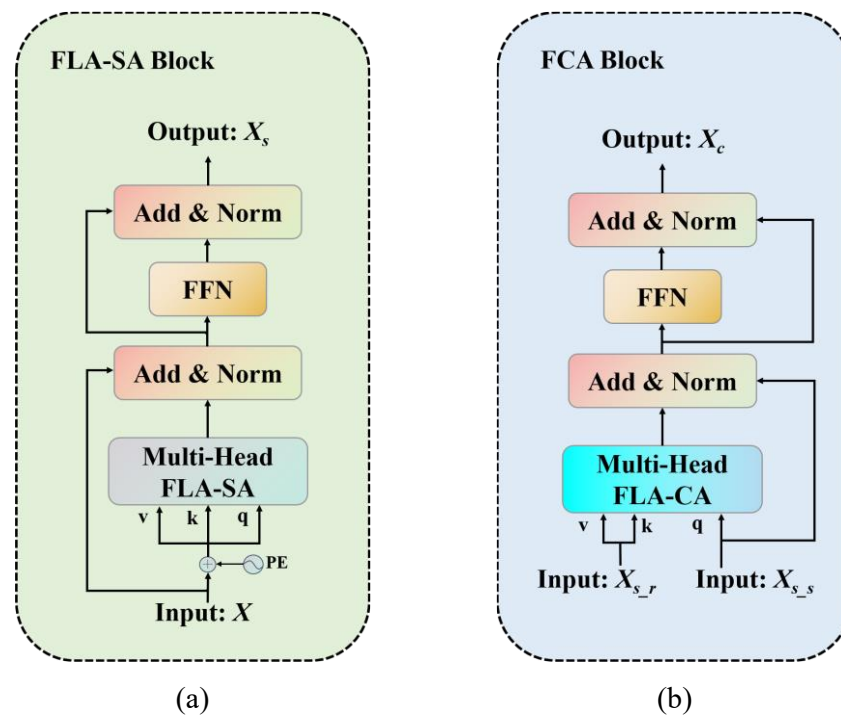


Figure 2. Components of the FLA Module: (a) FLA Block. (b) FCA Block.

The computation of the focused attention weights is expressed as follows

$$\text{FLA - Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})_i = \frac{\sum_{j=1}^n \Phi(q_i) \Phi(k_j)^\top v_j}{\sum_{j=1}^n \Phi(q_i) \Phi(k_j)^\top} \quad (2)$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ denote the query, key and value matrices obtained by linear projections of the input feature map and $\Phi(\cdot)$ denotes the focusing function defined in Equation (3).

In standard linear attention, the kernel function is typically $\phi(x) = \text{elu}(x) + 1$, which produces smooth attention weights that approximate the average of all features, limiting the model's ability to concentrate on the most informative regions. To overcome this, the FLA Module replaces the standard kernel with a focusing function that sharpens the attention distribution. Specifically, after applying ReLU to ensure non-negativity, the focusing function is defined as

$$\Phi(x) = f_p(\text{ReLU}(x)), f_p(x) = \frac{\|x\|}{\|x^{**p}\|} x^{**p} \quad (3)$$

where x^{**p} denotes the element-wise p -th power of x . A key property of this focusing function is its ability to discriminate between similar and dissimilar query–key pairs. This property ensures that the focusing function pushes similar query–key pairs closer while separating dissimilar ones, concentrating attention on the most relevant features and suppressing irrelevant ones [47]. Following empirical best practices, p is set to 3 in our implementation.

The FSA Block aggregates long-range global context within a single image, while the FCA Block facilitates cross-image feature interactions. For the FSA Block, the attention vectors, are derived from the same image, that is, from either the source or the reference feature map. To enhance its global context awareness and ability to capture long-range dependencies, position encoding (PE) is added to the attention outputs, which then pass

through residual connections, layer normalization and a feed-forward network (FFN) to yield the feature representation. For the FCA Block, the attention vector originates from the source image feature map, while $\mathbf{K}$ and $\mathbf{V}$ are derived from the reference image feature map. This facilitates information fusion between the source and reference images. Specifically, attention weights are computed between each position in the source image and all positions in the reference image. This enables the source image to focus on and integrate relevant information from the reference image, thereby enhancing the understanding and expression of inter-image relationships.

Beyond the focusing mechanism, the FLA Module addresses another fundamental limitation of linear attention: the rank constraint on the attention matrix. In linear attention, the output is computed as

$$\text{Output} = \Phi(\mathbf{Q})\Phi(\mathbf{K})^\top \mathbf{V} = \mathbf{M}\mathbf{V} \quad (4)$$

where $\mathbf{M} = \Phi(\mathbf{Q})\Phi(\mathbf{K})^\top$ is the attention matrix. The rank of $\mathbf{M}$ satisfies

$$\begin{aligned} \text{rank}(\mathbf{M}) &= \text{rank}(\Phi(\mathbf{Q})\Phi(\mathbf{K})^\top) \\ &\leq \min\{\text{rank}(\Phi(\mathbf{Q})), \text{rank}(\Phi(\mathbf{K})^\top)\} \\ &\leq \{N, d\} \end{aligned} \quad (5)$$

To alleviate this rank bottleneck, a depthwise convolution (DWC) branch is introduced. The DWC branch operates independently on the value matrix, and the output becomes

$$\begin{aligned} \text{Output} &= \Phi(\mathbf{Q})\Phi(\mathbf{K})^\top \mathbf{V} + \text{DWC}(\mathbf{V}) \\ &= (\Phi(\mathbf{Q})\Phi(\mathbf{K})^\top + \mathbf{M}_{\text{DWC}})\mathbf{V} \\ &= \mathbf{M}_{\text{eq}}\mathbf{V} \end{aligned} \quad (6)$$

where $\mathbf{M}_{\text{DWC}}$ is the sparse matrix corresponding to the depthwise convolution function and $\mathbf{M}_{\text{eq}}$ is the equivalent attention matrix. In our implementation, the DWC branch employs a 5×5 depthwise convolution with grouped channels, which effectively enriches the aggregated features and mitigates the homogenization issue inherent in standard linear attention.

Furthermore, inspired by the Feature Matching Transformer designed in TransMVSNet [18], we introduce the FLA Module based on the FSA Block and FCA Block to perform intra-view and inter-view attention on features. As illustrated in Figure 3, the feature map of each view is processed sequentially by n FLA Modules, the number of which is set to 3 in our experiments. Within each FLA Module, the reference feature and each source feature are first processed by a shared-weight FSA Block to compute intra-attention and then processed by an FCA Block for inter-attention. The FLA Module provides more prominent image features for cost volume construction and regularization.

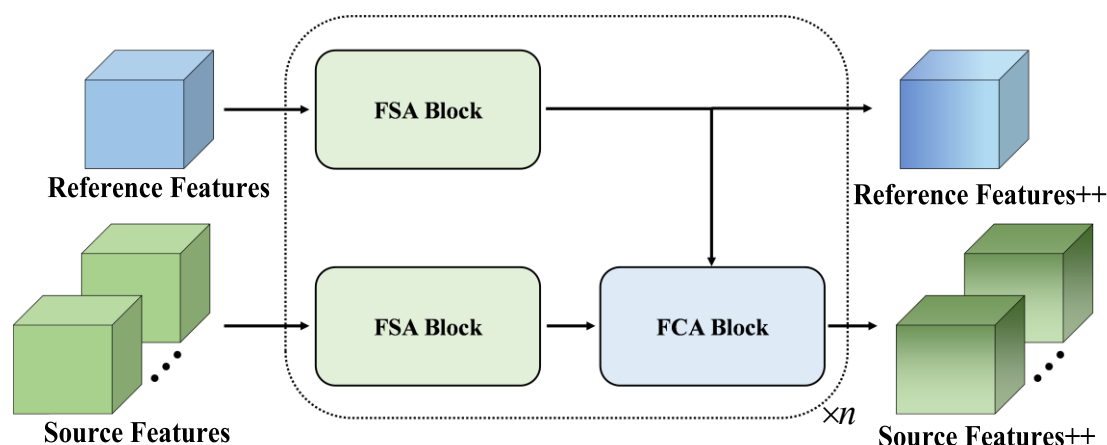


Figure 3. Structure of the FLA Module.

The process above can be formally denoted as follows

$$\{\mathbf{F}_i^j\}_{i=0}^N = \{\text{FSA}(\mathbf{F}_i^{j-1})\}_{i=0}^N \quad (7)$$

$$\{\mathbf{F}_i^j\}_{i=1}^N = \{\text{FCA}(\text{FSA}(\mathbf{F}_0^{j-1})), \text{FSA}(\mathbf{F}_i^{j-1})\}_{i=1}^N \quad (8)$$

where $\{\mathbf{F}_i^j\}_{i=0}^N$ and $\{\mathbf{F}_i^j\}_{i=1}^N$ mean the output of feature maps of the j -th FLA Module of the reference image and the source images, respectively.

The FLA Module builds upon two established frameworks and combines them within the MVS feature extraction pipeline. The alternating self-attention and cross-attention architecture follows the Feature Matching Transformer (FMT) design of TransMVSNet [18], while the focusing function, the depthwise convolution branch for attention matrix rank enhancement, and learnable scale normalization are adopted from the focused linear attention framework by Han et al. [47], which was originally designed for self-attention in general visual recognition. To our knowledge, the proposed FLA Module is among the first applications of focused linear attention to cross-view feature matching in MVS. Unlike the original FMT, which relies on standard softmax attention, and the original focused linear attention, which operates solely within single-view self-attention, the proposed FLA Module applies the focusing mechanism to both intra-view and inter-view attention, enabling sharpened cross-view feature correspondence under linear computational complexity. This integration is particularly well-suited for satellite surface reconstruction, where repetitive textures on solar panels and weakly textured thermal insulation layers cause standard linear attention to produce overly diffuse matching distributions. The element-wise power function with $p=3$ concentrates attention on the most discriminative correspondences, while the DWC branch preserves local geometric structures essential for accurate depth estimation. The experimental results in Section 4.2 validate that the FLA Module yields measurable improvements in reconstruction accuracy and completeness on satellite datasets.

3.3. Mamba Cost Volume Module

The cost volume encodes multi-view feature similarity across depth hypotheses and serves as the input to the regularization module. Its construction proceeds as follows. Based on differentiable homography warping, pixel-wise correspondences between the reference and source views are established using camera parameters and depth hypotheses. Let $\mathbf{K}_1$, $\mathbf{R}_1$, and $\mathbf{t}_1$ denote the intrinsic matrix, rotation, and translation of the reference camera and $\mathbf{K}_i$, $\mathbf{R}_i$, and $\mathbf{t}_i$ denote those of the i -th source camera. Given a depth hypothesis d and the plane normal $\mathbf{n}$, the homography matrix warping the i -th source feature map to the reference frame is

$$\mathbf{H}_i(d) = \mathbf{K}_1 \mathbf{R}_1 \left(\mathbf{I} - \frac{(\mathbf{R}_1^{-1} \mathbf{t}_1 - \mathbf{R}_i^{-1} \mathbf{t}_i) \mathbf{n}^T \mathbf{R}_1}{d} \right) \mathbf{R}_i^{-1} \mathbf{K}_i^{-1} \quad (9)$$

Through this transformation, each feature vector in the reference view, that is, the vector of channel values at a given pixel location of the feature map, is matched with its corresponding position in each source view, yielding N warped 4D feature volumes.

To measure feature similarity while reducing channel dimensionality, group-wise correlation is employed. The feature volume with Ch channels is evenly divided into G groups along the channel dimension, and the per-group similarity is computed as

$$s_i^g = \frac{1}{Ch/G} \langle \mathbf{F}_0^g, \mathbf{F}_i^g \rangle \quad (10)$$

where $g = 0, 1, 2, \dots, G-1$; the angle brackets denote the inner product. $\mathbf{F}_0^g$ and $\mathbf{F}_i^g$ denote the g -th set of features in feature map $\mathbf{F}_0$ and $\mathbf{F}_i$, respectively. The per-group similarities are concatenated along the channel dimension to form the cost volume s_i for each source view. Finally, a lightweight 3D CNN produces per-view pixel-wise weights $\{\mathbf{w}_i\}_{i=0}^{N-1}$, and the individual cost volumes are aggregated through weighted fusion,

$$\mathbf{C} = \frac{\sum_{i=0}^{N-1} \mathbf{w}_i s_i}{\sum_{i=0}^{N-1} \mathbf{w}_i} \quad (11)$$

where $\mathbf{C}$ serves as the initial cost volume, which is subsequently refined by the MCV Module to produce the probability volume for depth estimation.

Cost volume regularization is critically important, as it filters noise introduced by environmental disturbances such as illumination variations and occlusions while enhancing discriminative features for more accurate depth estimation. For an input image of size $3 \times 1152 \times 1536$ and an initial depth hypothesis count of $D=16$, the cost volume input to the regularization module possesses dimensions of $64 \times 16 \times 72 \times 96$ after feature extraction and downsampling. Flattening the volume along its depth and spatial dimensions yields a sequence of $16 \times 72 \times 96 \approx 10^5$ tokens, which imposes a significant computational burden on Transformer-based architectures. Leveraging the inherent advantages of state space models (SSMs), Mamba demonstrates superior capability in handling extremely long sequences with near-linear complexity. Given that cost volume regularization does not require cross-attention operations akin to those in feature extraction, Mamba is particularly well-suited for this module.

A Mamba-based cost aggregation network named MCV takes Vision Mamba as its core computing unit, with the objective of capturing global spectral-spatial semantic information and filtering cost volume noise. The detailed structure is illustrated in Figure 4. Specifically, a cost volume with the size of $C \times D \times W \times H$ is first processed by a 3D convolutional module to $\times 2$ downsampling after the addition of 3D positional encoding, denoted as $\mathbf{E}_{pos}$, resulting in an output of size $C' \times D/2 \times W/2 \times H/2$, noted as $\mathbf{C}_{down}$. This can be formally expressed as

$$\mathbf{C}_{down} = \text{conv}_{3d}(\mathbf{C} + \mathbf{E}_{pos}) \quad (12)$$

The 3D positional encoding employs sinusoidal functions applied independently along each spatial axis (x, y, z) , where the coordinates are normalized via min-max scaling using the global spatial extent for x and y and the depth range $[d_{min}, d_{max}]$ for z . The resulting encoding produces 3C channels (C per axis), which are then projected back to C channels through a learnable $1 \times 1 \times 1$ 3D convolution before being added to the cost volume [27].

The convolutional layer captures local features and produces a representation that matches the input requirements of Mamba. Subsequently, to match the input size of the Vision Mamba Block, the output undergoes a flattening operation, which converts it into a sequence of tokens $\mathbf{S}$ with dimensions of $C \times L$, where $L = D \times W \times H$ is the sequence length. This process can be formally represented as

$$\mathbf{S} = \text{Flatten}(\mathbf{C}_{down}) \quad (13)$$

Subsequently, the output passes sequentially through multiple Vision Mamba (Vim) blocks, which perform global feature extraction and sequence modeling and thereby suppress noise and integrate features. In our implementation, the number of Vim Blocks is set

to 2. Within each Vim Block, the input first passes through a normalization layer and is then split into a scan branch and a gating branch. In the scan branch, following a linear projection, the features are further split into forward and backward scan paths, each comprising a 1D convolutional layer and an SSM layer. Their outputs are denoted as X_f and X_b , respectively. This process can be formulated as follows

$$X_f = \text{SSM}_f(\text{Conv}(\text{Norm}(\mathcal{S}))) \quad (14)$$

$$X_b = \text{SSM}_b(\text{Conv}(\text{Norm}(\mathcal{S}))) \quad (15)$$

where Norm is the normalization layer and SSM_f and SSM_b represent the forward SSM and the backward SSM, respectively. Another branch of Mamba introduces a gated mechanism, which comprises a linear layer and an activation layer, formulated as

$$\mathbf{Z} = \sigma(\text{Linear}(\text{Norm}(\mathcal{S}))) \quad (16)$$

where $\sigma(\cdot)$ is SiLU activation. The output $\mathbf{z}$ is then multiplied with the X_f and X_b , with the results denoted as $\mathcal{S}_f$ and $\mathcal{S}_b$, respectively, as follows

$$\mathcal{S}_f = X_f \otimes \mathbf{Z} \quad (17)$$

$$\mathcal{S}_b = X_b \otimes \mathbf{Z} \quad (18)$$

where $\otimes$ is element-wise multiplication. The outputs of the two SSM branches are summed and then passed through a linear layer for channel restoration. Then, a skip connection adds the input to obtain the final output. This process can be denoted as

$$\text{Vim}(\mathcal{S}) = \text{Linear}(\text{add}(\mathcal{S}_f, \mathcal{S}_b)) \quad (19)$$

$$\mathcal{S}^i = \text{Vim}(\mathcal{S}^{i-1}) + \mathcal{S}^{i-1} \quad (20)$$

Subsequently, the output of the final Vim layer is reshaped and processed by a 3D deconvolution for dimension and resolution recovery. A 1×1 convolutional layer is then adopted to output a probability volume $\mathbf{P}$ of dimensions $D \times W \times H$, where $\mathbf{P}(k, i, j)$ represents the probability of the k -th hypothesis d_k at (i, j) in the image, which can be formally expressed as

$$\mathbf{C}_m = \text{Deconv}_{3d}(\text{reshape}(\mathcal{S}^n)) \quad (21)$$

$$\mathbf{P} = \text{prop}(\mathbf{C}_m) \quad (22)$$

where $\text{prop}(\cdot)$ means a 1×1 convolutional layer.

Cost volume regularization dominates the computational cost of the pipeline. At the sequence length reported in Section 3.3 (of the order of 10^5 tokens), a Transformer regularizer must relate every pair of tokens, i.e., about 10^{10} interactions with quadratic growth in sequence length, whereas the MCV Module propagates the same sequence through a linear recurrence. Section 4.2 quantifies the resulting trade-off: at comparable numbers of trainable parameters (39.484 M for the Transformer baseline and 39.475 M for MCV) and comparable memory consumption (5964.6 MB and 5947.2 MB), MCV reduces the inference time from 1.675 s/img to 1.085 s/img and the overall error from 0.856 to 0.737.

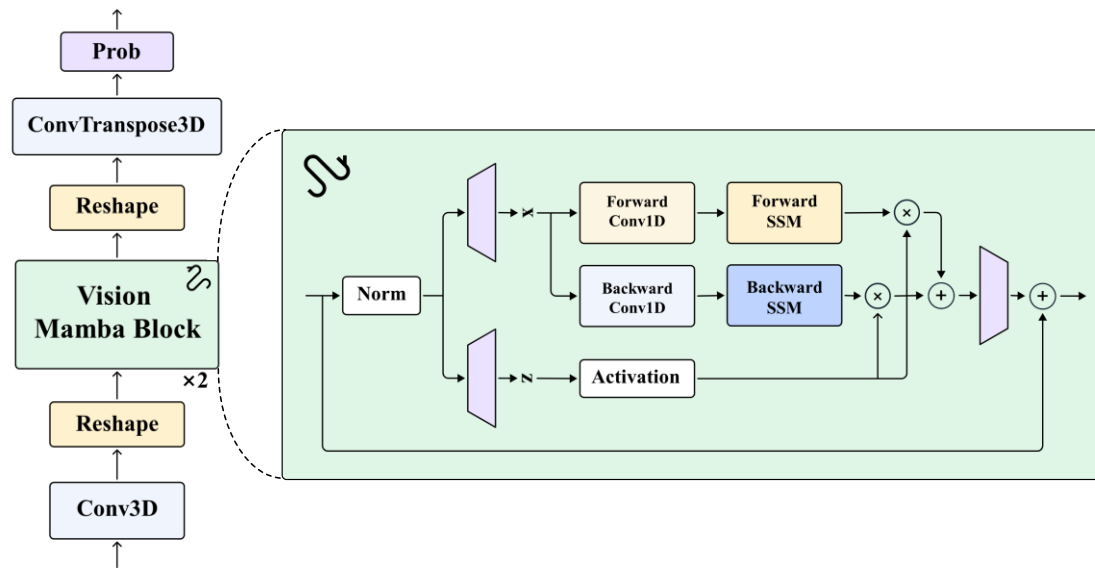


Figure 4. Structure of the Mamba Cost Volume Regularization Module.

3.4. Loss Function

Following the classification-based formulation of MVSNet [20], the cross-entropy loss is employed for depth estimation. In multi-view stereo (MVS) methods based on plane-sweep depth hypotheses, depth estimation is formulated as a classification task, where the model selects the optimal depth hypothesis from discrete depth candidates. Furthermore, when hierarchical depth hypotheses are adopted, the cross-entropy loss facilitates multi-stage probability refinement. By measuring the divergence between the predicted depth probability distribution and the ground truth distribution, the cross-entropy loss is particularly well-suited for this categorical modeling task.

$$loss = \sum_{p \in \{p_v\}} -(1 - P^{(\hat{d})}(p)) \log(P^{(\hat{d})}(p)) \quad (23)$$

where $P^{(\hat{d})}(p)$ represents the predicted probability of the depth hypothesis $\hat{d}$ at the pixel p , $\{p_v\}$ represents the subset of pixels with valid truth values, and $\hat{d}$ represents the depth value that is closest to the true value among all depth assumptions. Since the algorithm adopts a multi-stage coarse-to-fine deep prediction framework, the total loss of the algorithm is obtained by summing the depth losses of each stage. Then, the total loss is expressed as

$$Loss = \sum_{i=1}^N loss_i \quad (24)$$

where N represents the number of stages and is set to 4.

3.5. Pose Measurement Method

Beyond dense 3D reconstruction, accurate relative pose measurement is equally essential for on-orbit servicing missions. The reconstructed 3D point clouds and estimated camera poses from the preceding MVS pipeline serve as the geometric foundation for the pose measurement framework described below, which addresses the challenge of training reliable keypoint detectors under limited data conditions.

Figure 5 illustrates the workflow of reconstruction and pose measurement, in which the pose measurement process comprises two stages: offline training and online measure-

ment. During the offline training phase, we first augment the dataset using 3DGS rendering, leveraging camera poses and reconstructed point clouds obtained from the MVS pipeline. This augmentation effectively expands our limited real image dataset, which is critical for training models under few-shot scenarios. Subsequently, we adopt a geometry-based reprojection method to automatically annotate semantic keypoints across the entire training set. This method projects the 3D world coordinates of keypoints onto each image using the known camera pose information, requiring manual labeling of only a small subset of images.

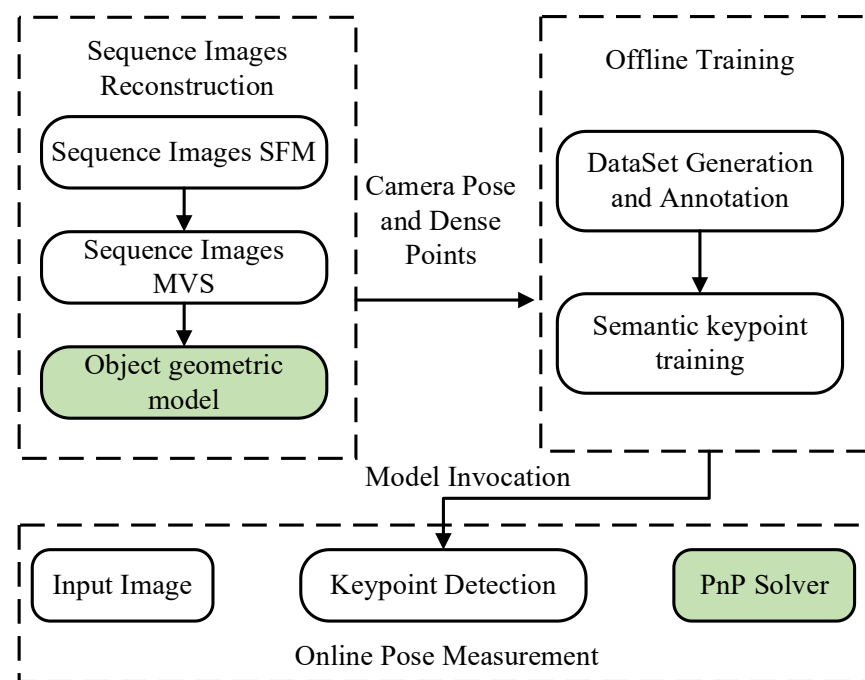


Figure 5. Flowchart of reconstruction and pose measurement. The green box indicates the obtained output.

The resulting labeled dataset is then employed to train a top-down keypoint detection model, producing a model optimized for online inference. In the online measurement phase, the pre-trained keypoint detection network processes real-time satellite images to predict the 2D image coordinates of predefined semantic keypoints. By combining these 2D coordinates with their known 3D world positions, camera poses can be efficiently computed using the Perspective-n-Point (PnP) algorithm. This approach achieves accurate semantic keypoint detection and pose measurement under few-shot conditions while significantly reducing the burden of manual annotation.

The architecture of the keypoint detection network is shown in Figure 6. The RTMpose [48] network architecture employed in this method consists of a backbone network, 7×7 convolutional layers, fully connected layers (FC), and gated attention units (GAUs). For pose measurement, the model employs a Simple Coordinate Classification (SimCC) strategy. This decouples the prediction of keypoint coordinates into two independent classification tasks: horizontal (X-axis) and vertical (Y-axis) directions. Dual-channel classifiers separately output keypoint positions within the image, enabling efficient and precise location inference. During feature pyramid fusion, standard convolution operations involving subsampling can cause misalignment between feature maps and original spatial positions, leading to cumulative errors and reduced keypoint regression accuracy. Furthermore, in applications like satellite pose measurement, target components (e.g., solar panels, antennas) exhibit significant perspective distortion despite their rigid structures. This distortion arises from shooting angles, occlusions, or attitude

changes, which traditional fixed-geometric convolutional kernels struggle to model effectively. To address these challenges, we modify the CSPNeXt backbone network in the original RTMpose algorithm by replacing some standard convolutions with Deformable Convolution v4 (DCNv4), constructing a novel CSP_D Layer. This design enhances the network's adaptability to spatial geometric deformations and effectively mitigates spatial misalignment issues caused by downsampling, thereby achieving higher-precision spatial keypoint extraction under complex observation conditions.

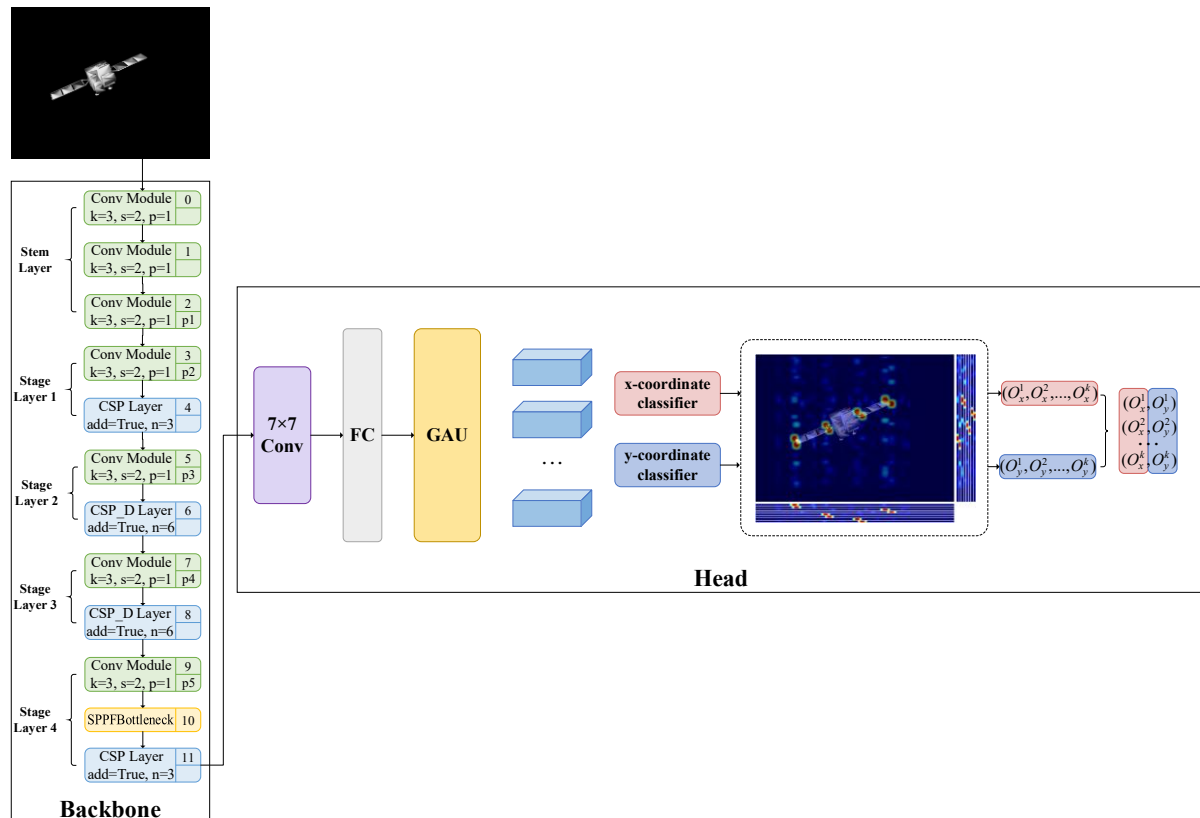


Figure 6. Architecture of the keypoint detection network.

The selection of semantic keypoints follows three principles to ensure reliable pose measurement: (1) structural stability: keypoints are placed at geometrically well-defined corner locations that are invariant to illumination changes; (2) semantic unambiguity: each keypoint corresponds to a physically meaningful and identifiable location on the satellite structure; and (3) multi-view visibility: keypoints are selected to be visible from a wide range of viewing angles during circumnavigation.

Pose accuracy depends on how the reconstruction uncertainty propagates into the estimated pose. Let the keypoints carry a three-dimensional error whose component perpendicular to the viewing ray is $\delta X_{\perp}$, and let σ_u denote the keypoint localization error in the image. A lateral displacement produces an equivalent image displacement of $(f/z)\delta X_{\perp}$ ($f \approx 2311.6px$, Section 4.6; $z = 0.425-0.935m$, Section 4.1.2), and both terms enter the reprojection residual additively, so that

$$\sigma_{u,\text{total}}^2 = \sigma_u^2 + \left(\frac{f}{z}\right)^2 \delta X_{\perp}^2 \quad (25)$$

and the pose covariance follows from first-order propagation through the PnP refinement,

$$\Sigma_{\xi} = \sigma_{u,\text{total}}^2 (\mathbf{J}^T \mathbf{J})^{-1}, \mathbf{J} = \frac{\partial \mathbf{u}}{\partial \xi} \in \mathbb{R}^{2n \times 6} \quad (26)$$

with the keypoint errors treated as mutually independent and a global scale error of the reconstruction excluded as systematic. Since $\text{diag}[(J^T J)^{-1}]$ depends only on the keypoint configuration and the observing geometry, the reconstruction term grows as the range decreases and competes in quadrature with the localization term, so that refining the reconstruction alone yields diminishing returns once $(f/z), \delta X_{\perp} < \sigma_u$. Conversely, an unfavorable configuration, such as under partial occlusion, amplifies both terms at once, which is consistent with the increase in the worst-case, but not mean, error reported in Section 4.5.2.

The pose measurement stage presupposes a target-specific offline preparation, namely, a set of semantic keypoints defined on the target structure and a detector trained on the resulting annotated views, whereas the reconstruction stage is target-agnostic, as the network is trained on the DTU benchmark alone and applied without fine-tuning. This requirement is met in missions in which the target is identified in advance and observed over an extended period, but not for a target that has not been observed in advance and can only be imaged over a short window, in which case the limiting stage is pose measurement rather than dense reconstruction; the sensitivity of the pose solution to keypoint localization and reconstruction error is analyzed above, and relaxing the requirement of predefined keypoints is left as future work.

4. Experiments

4.1. Experimental Configuration

4.1.1. Dataset

For training, we utilize the dataset from the Technical University of Denmark (DTU). Due to the scarcity of MVS evaluation datasets for space satellites, we compiled image sequences from various satellite types to form a Customized Space Target (CST) dataset. Sample images are shown in Figure 7. Seven groups of image sequences were acquired from three satellite models, encompassing both simulated and real-world variants.

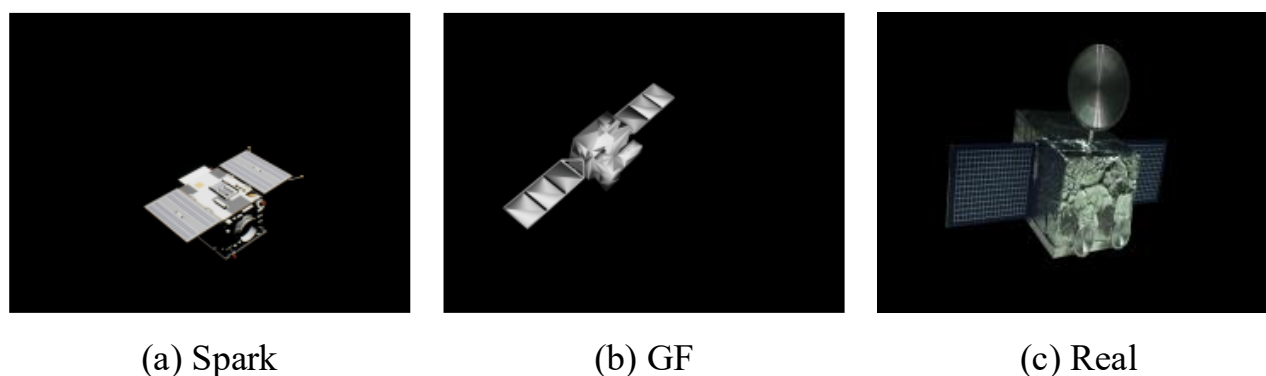


Figure 7. Different satellite models in the dataset.

The first model is derived from the SPARK dataset [49], characterized by a larger volume and abundant surface-mounted components with repetitive textures (e.g., solar panels), non-repetitive micro-structural details, and weakly textured regions (e.g., black side surfaces). Five representative 100-image sequences were selected based on varying distances, reflectivity, and illumination. The second model is a simulated satellite constructed from CAD models; the GF satellite represents a challenging textureless scenario, with 50-image sequences generated using predefined camera poses. The third model is a real-world satellite equipped with thermal insulation and reflective surfaces, accompanied by a 75-image sequence. The real model is smaller in volume and presents realistic

wrinkled thermal insulation layers, protruding occluding structures, and repetitive textures. Collectively, these targets cover rich, weak, and absent texture conditions with varying structural complexities, ensuring robust evaluation of reconstruction performance.

No CST sequence is used for training; the reconstruction network is trained on the DTU dataset alone, so the CST dataset serves exclusively as a test set. For the pose measurement stage, the keypoint detector is trained on the augmented image set described in Section 4.4.1 and evaluated on a disjoint set of real images acquired separately. Camera parameters and poses were predefined for the simulated sequences. Images of the physical model were acquired at observation distances of 425–935 mm over a viewing range covering the front and sides of the target, with the calibrated camera reported in Table 11. Keypoint annotations were obtained by manually labeling a small subset of images and propagating the labels to the rest by reprojection, as detailed in Section 3.5.

To simulate a space environment, backgrounds were removed via semantic segmentation. In practical scenarios, background elements such as the Earth can also be removed using established methods and further mitigated by controlling the satellite used to image the target.

To evaluate robustness, we established diverse lighting conditions and simulated imaging variations caused by satellite motion. As shown in Figure 8, two satellite types were included: a real satellite with rich texture and a SPARK satellite with low texture. Three lighting conditions were evaluated: standard lighting consistent with the baseline dataset; low-brightness conditions with intensity reduced by 50% and highlights suppressed; and high-brightness, high-reflectivity conditions with intensity increased by 50% and amplified specular reflections. For each scenario, the real satellite subset comprises 64 images per group, while the SPARK satellite subset contains 75 images per group.

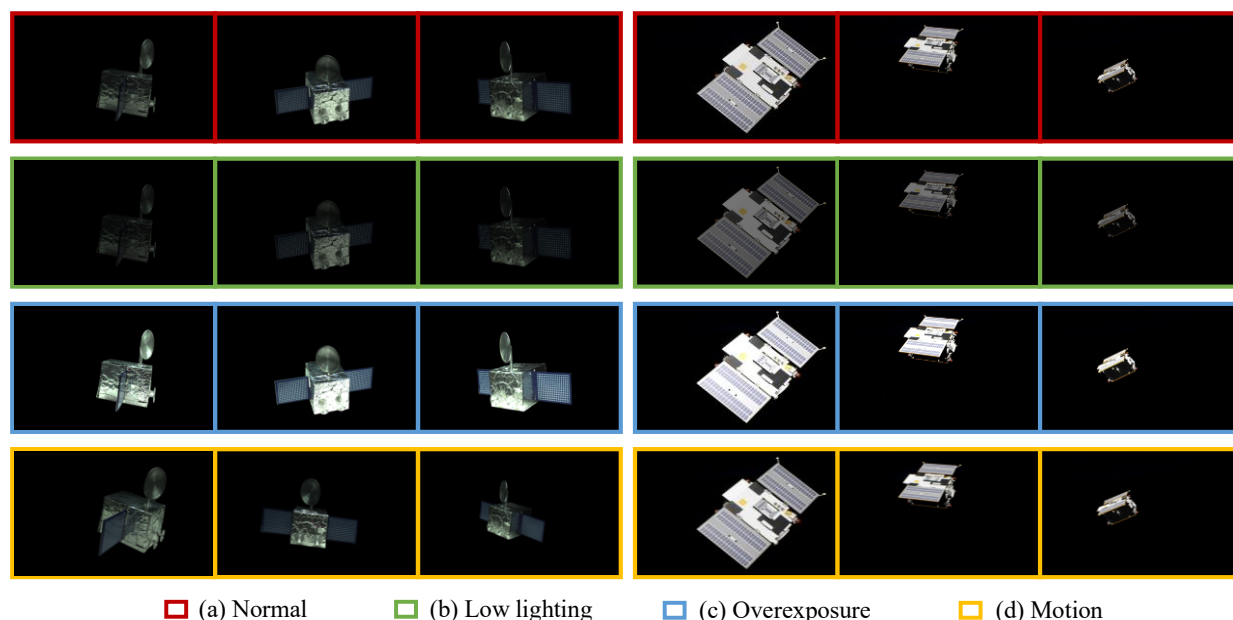


Figure 8. Datasets for simulating imaging in various on-orbit environments.

The motion dataset simulates reduced usable images and motion blur caused by relative motion during acquisition. Specifically, images exhibit an average motion blur of 12 pixels, and the number of input images is reduced by half while maintaining geometric consistency, resulting in degradation of both quality and quantity. To further evaluate pose measurement under occlusion, we assessed accuracy under scenarios where 25%, 50%, and random 0–50% of feature points were missing.

4.1.2. Method Implementation Details

In the training stage, the number of input images is $N=5$. Following previous works [21,28], we adopt a multi-scale training strategy with resolution randomly scaled between 512×640 and 576×832 . Inverse depth sampling is adopted for depth hypothesis generation, with the depth range from 425 mm to 935 mm. A four-stage coarse-to-fine cascade is employed with 32, 16, eight, and four depth hypotheses per stage. At each subsequent stage, the depth interval is adaptively refined based on the preceding prediction. The batch size is one. The model is implemented in PyTorch(2.0.1) and trained with AdamW for 15 epochs at an initial learning rate of 0.001 on four NVIDIA TITAN V GPUs, with a 500-step warmup followed by cosine annealing.

All implementation details are listed in Table 2. The pseudocode of all methods is shown in Algorithms 1–3. During evaluation on the CST dataset, input images are resized to 1152×1536 with $N=5$. Camera parameters and depth sampling range are estimated using COLMAP [50], which integrates Structure-from-Motion (SfM) into a comprehensive 3D reconstruction pipeline that combines information across images. We follow a dynamic checking strategy using geometric and photometric constraints for depth filtering, with the number of consistent views and photometric probability threshold set to four and 0.5, respectively. Finally, Gipuma [51] fuses the refined depth maps into a dense 3D point cloud.

Table 2. Implementation details of the proposed framework. FLA, focused linear attention; DWC, depthwise convolution; DCNv4, deformable convolution v4; SimCC, simple coordinate classification.

Category	Item	Setting
Training	Input views	5
	Training resolution	512×640 to 576×832 , randomly sampled
	Depth sampling	Inverse depth sampling, 425–935 mm
	Depth hypotheses per stage	32/16/8/4
	Optimizer	AdamW, learning rate 0.001, 500-step warmup with cosine annealing
	Epochs / batch size	15/1
	Hardware and framework	4× NVIDIA TITAN V, PyTorch
	Training data	DTU dataset
	Loss function	Per-stage cross-entropy, summed over the four stages
Evaluation	Input resolution	1152×1536
	Camera parameters	Estimated with COLMAP
	Depth filtering	Dynamic geometric and photometric checking
	Depth fusion	Gipuma
	Novel view rendering	3D Gaussian Splatting, 30,000 iterations
Network	FLA Modules per view	3
	Focusing function and DWC	$p=3$; 5×5 grouped depthwise convolution
	Vim blocks in the MCV Module	2
	Cost volume size/token length	$64 \times 16 \times 72 \times 96$ / of the order of 10^5
Pose measurement	Keypoint detector	RTMpose with SimCC head and CSP_D (DCNv4) layers
	Pose solver	OpenCV solvePnP Ransac, iterative method
	Camera	Allied Vision AVT GC2000C (Allied Vision Technologies GmbH, Stadtroda, Germany), 12 mm focal length (reprojection error 0.05–0.07 px)
	Online pose rate	≈ 30 Hz

Algorithm 1 MambaMVS for dense 3D reconstruction

Input: reference image I_0 ; source images $\{I_i\}$, $i=1,\dots,N$; camera parameters $\{K_i, R_i, t_i\}$; depth range $[d_{min}, d_{max}]$

Output: dense point cloud

```

1: for each view  $i \in \{0, 1, \dots, N\}$  do
2:    $F_i \leftarrow$  shared-weight  $FPN(I_i)$  enriched by frozen  $DINOv2(I_i)$  // Equation (1)
3: end for
4: for  $j=1$  to 2 do // FLA Modules
5:   for each view  $i \in \{0, 1, \dots, N\}$  do
6:      $F_i \leftarrow FSA\_j(F_i)$  // shared weights, intra-view attention
7:   end for
8:   for  $i=1$  to  $N$  do
9:      $F_i \leftarrow FCA\_j(F_i, F_0)$  // Equations (7) and (8);  $Q$  from source,  $K$  and  $V$  from reference
10:   end for
11: end for
12: initialize the depth hypotheses from  $[d_{min}, d_{max}]$  by inverse depth sampling
13: for  $s=1$  to 4 do //  $D_s = \{32, 16, 8, 4\}$ 
14:   if  $s > 1$  then refine the depth hypotheses from the previous stage
15:    $V_i \leftarrow$  warp  $F_i$  into the reference frame by differentiable homography // Equation (9)
16:    $C \leftarrow$  weighted fusion of the group-wise correlation volumes // Equations (10) and (11)
17:    $P \leftarrow$  probability volume obtained by the MCV Module from  $C$  // Equations (21) and (22)
18:    $d_s \leftarrow$  per-pixel depth from  $P$ 
19: end for
20: fuse the multi-view depth maps by dynamic geometric and photometric checking, then filter with Gipuma
21: return the dense point cloud

```

Algorithm 2 Offline training of the keypoint detector

Input: reconstructed point cloud; camera poses; satellite images

Output: trained keypoint detector; 3D coordinates of the semantic keypoints

```

1: extract the 3D coordinates of the predefined semantic keypoints from the point cloud
2: render novel views from the reconstruction by 3D Gaussian Splatting
3: annotate a small subset of the images manually
4: propagate the keypoint annotations to the remaining images by reprojection using the known camera poses
5: train the keypoint detector on the augmented and annotated image set
6: return the trained keypoint detector and the 3D keypoint coordinates

```

Algorithm 3 Online pose measurement

Input: satellite image; trained keypoint detector; 3D keypoint coordinates; camera intrinsics

Output: 6-DoF relative pose

```

1: detect the 2D keypoints in the input image

```

```

2: solve the 6-DoF pose by solvePnP(Ransac (2D keypoints, 3D keypoints, camera intrin-
sics) // iterative method
3: return the 6-DoF relative pose

```

4.1.3. Ground Truth

To ensure rigorous evaluation, ground truth is stratified based on data availability and target cooperativeness.

- (a) SPARK and GF models: For satellite models with accessible 3D CAD files, we generate a uniform point cloud via Monte Carlo sampling from the mesh surface as the absolute geometric ground truth for evaluating reconstruction accuracy and completeness. Camera poses are predefined during dataset generation, providing absolute pose ground truth.
- (b) Real satellite model: Dimensions were measured using a high-precision caliper (accuracy ± 0.02 mm) to evaluate reconstruction scale accuracy. For pose reference, due to the difficulty in acquiring absolute ground truth in experimental settings, we employed COLMAP to derive pose references. SfM leverages bundle adjustment over multi-view sequences to provide globally optimized relative poses with superior consistency, unlike the single-image pose estimation algorithm we propose. Consequently, SfM-derived poses serve as the reference truth for evaluating pose estimation consistency.

4.1.4. Evaluation Metrics

For the CST dataset, distinct evaluation metrics are employed based on ground truth availability. For simulated models with ground truth point clouds, we use accuracy, completeness, and overall score. Completeness is defined as the mean distance from each point in the ground truth point cloud P_G to the nearest point in the reconstructed point cloud P_R . Accuracy is defined as the mean distance from each point in the reconstructed point cloud P_R to the nearest point in the ground truth point cloud P_G . Overall is the mean of accuracy and completeness. Their formulas are given below

$$\text{Accuracy} = \frac{1}{N} \sum_{p_n \in P_R} \min_{p_m \in P_G} \|p_m - p_n\|_2^2 \quad (27)$$

$$\text{Completeness} = \frac{1}{M} \sum_{p_m \in P_G} \min_{p_n \in P_R} \|p_m - p_n\|_2^2 \quad (28)$$

$$\text{Overall} = \frac{\text{Accuracy} + \text{Completeness}}{2} \quad (29)$$

For scenes lacking ground truth, we evaluate the size measurement error (SE) and the plane fitting error (RMS), which quantify the accuracy of dimensional measurements and planar reconstruction, respectively, and are computed as follows

$$\text{SE} = |l - \hat{l}| \quad (30)$$

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{i=0}^N d(p_i, P)} \quad (31)$$

where l and $\hat{l}$ denote the measured and ground truth values of the key dimensions, n represents the number of points in the planar point cloud, P denotes the fitted plane, p_i represents the i -th point in the point cloud, and $d(\cdot)$ denotes the Euclidean distance.

However, due to scale ambiguity in monocular vision, simply averaging scores across groups is inappropriate. Therefore, we employ the normalized mean size error (NSE) and normalized mean plane fitting error (NRMS), which account for scale inconsistencies across different groups.

$$\text{NSE} = \sum_k \left| l_k^a / l_k^b - \hat{l}_k^a / \hat{l}_k^b \right| \quad (32)$$

$$\text{NRMS} = \sum_k \frac{\text{RMS}}{\left(\hat{l}_k^a \right)^2} \quad (33)$$

where k denotes the k -th image set, l_k^a and l_k^b represent the measured lengths of two key dimensions on the satellite, and $\hat{l}_k^a$ and $\hat{l}_k^b$ denote their corresponding ground truth.

4.2. Ablation Study

To evaluate the effectiveness and efficiency of the proposed modules, Table 3 presents ablation results on simulated satellite models. Compared to Group 1 (baseline), Group 2 with the FLA Module improves Overall by 0.159, with accuracy improving by 0.130 and completeness by 0.189 (lower is better). Comparing Groups 2–4, the MCV Module further enhances performance, improving Overall by 0.471 over the 3D-UNet baseline (Group 2) and by 0.119 over the Transformer baseline (Group 3). The FLA Module slightly increases parameters and inference time, while the MCV Module reduces both relative to the 3D U-Net and Transformer alternatives, with inference time reduced by $\sim 1.5\times$ relative to Transformer self-attention, demonstrating a superior accuracy–efficiency trade-off.

Table 3. Quantitative comparison of point cloud results, model memory and time cost for different module settings. The best results are denoted in bold, while the second-best results are underlined.

Module				Accuracy	Completeness	Overall	Time (s/img)	Memory (MB)	Params. (Trainable)
Feature		Cost							
FLA	3DCV	ACV	MCV						
X	✓	X	X	0.838	1.896	1.367	<u>1.117</u>	5965.7	<u>39.524</u>
✓	✓	X	X	0.708	1.707	1.208	1.129	5965.8	39.525
✓	X	✓	X	<u>0.548</u>	<u>1.164</u>	<u>0.856</u>	1.675	<u>5964.6</u>	39.484
✓	X	X	✓	0.521	0.952	0.737	1.085	5947.2	39.475

Table 4 evaluates the pose measurement pipeline under three training configurations. Group A is trained on 200 manually annotated images without augmentation. Group B uses an augmented set of 745 images produced by 3D Gaussian Splatting-based novel view augmentation, with manual keypoint annotation, and Group C uses the same images with only 10 of them annotated manually, the remaining 735 being labeled automatically by reprojection. Groups B and C share an identical image set and differ only in the annotation source; their comparison therefore isolates the effect of automatic annotation. Groups A and B differ in both the size and the composition of the training set, so their comparison quantifies the effect of the augmented training set as a whole rather than of 3DGS augmentation alone. Group B reduces the mean rotation error from 7.375° to 0.563° and the mean translation error from 0.694 m to 0.055 m relative to Group A. Group C maintains this accuracy, with all differences from Group B bounded by 0.8° and 0.04 m, while reducing manual labels from 745 to 10.

Table 4. Pose measurement on the physical satellite model. The best results are denoted in bold.

Group	Training Set (Images)	3DGS Augmentation	Manual Labels	Rotation	Translation
				ME/MaxAE/VarAE (°)	ME/MaxAE/VarAE (m)
A	200	X	200	7.375/41.957/110.334	0.694/4.268/1.017
B	745	✓	745	0.563/2.763/0.232	0.055/0.252/0.002
C	745	✓	10	0.573/2.659/0.258	0.053/0.220/0.002

The DINOv2-Base backbone is used as a frozen feature extractor and contributes no trainable parameters; it is adopted following MVSFormer++ [28] and is therefore shared with the baseline networks, so the comparison in Table 3 isolates the modules proposed in this work.

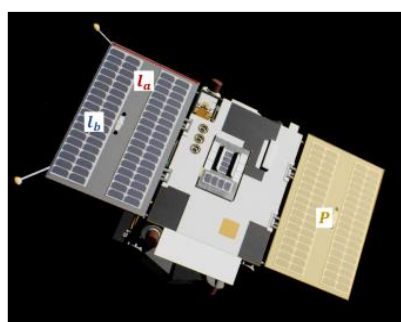
4.3. Dense 3D Reconstruction

The proposed method is evaluated against state-of-the-art MVS methods and 3DGS [10] on the CST dataset. Open-source pre-trained models of MVSFormer++ [28] and MVSFormer [24] are used without fine-tuning. For 3DGS, dense point clouds are generated after 30,000 training iterations. The best results are bolded and second-best underlined. All methods are evaluated on the same CST test split with identical input resolution (1152×1536), number of input views (five), camera parameters estimated by COLMAP and Gipuma-based fusion. The most recent state-space-based MVS methods [39,41] report results on DTU and Tanks and Temples but do not provide satellite benchmarks and are therefore discussed qualitatively rather than included in the quantitative comparison; extending the comparison to these methods on common satellite data is identified as future work. For simulated models with ground truth, we compare accuracy, completeness, and overall scores (Table 5). Our method achieves 0.521/0.952/0.737, demonstrating state-of-the-art performance.

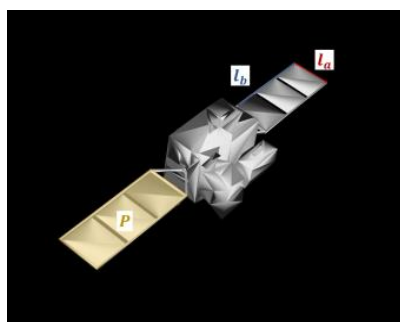
Table 5. Performance comparison on the simulated satellite model. The best results are denoted in bold, while the second-best results are underlined.

Method	Accuracy	Completeness	Overall
3DGS	1.555	1.138	1.247
MVSFormer	0.596	2.474	1.535
MVSFormer++	<u>0.751</u>	<u>1.060</u>	<u>0.906</u>
Ours	0.521	0.952	0.737

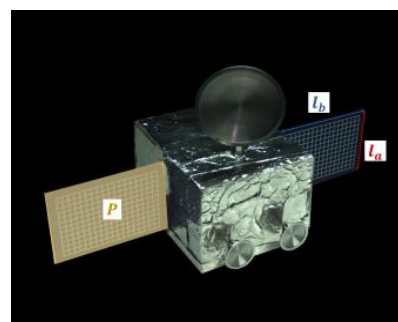
For the SPARK and real-world models, size error and plane fitting error are employed. The length l_a and width l_b of the solar panels were selected as key dimensions and the solar panel surface as the fitting plane P , as illustrated in Figure 9.



(a) Spark



(b) GF



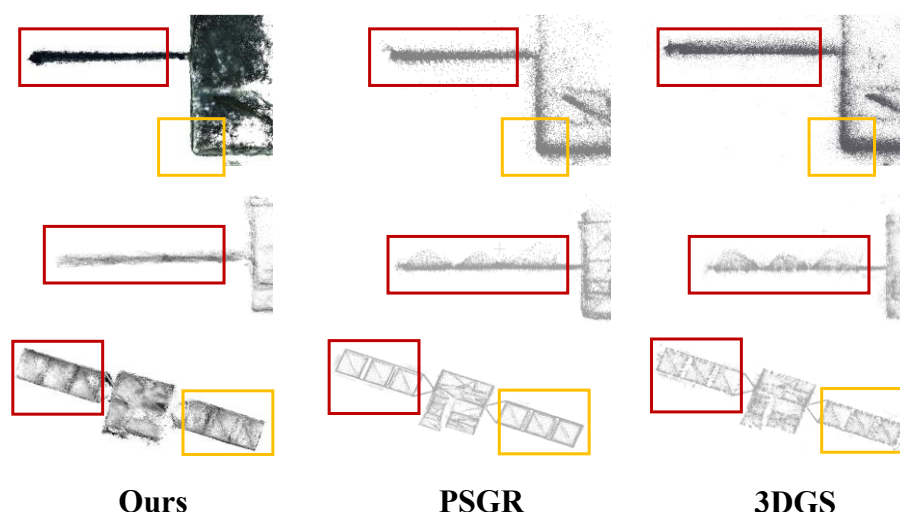
(c) Real

Figure 9. Schematic diagram of plane fitting and size measurement.

For size error evaluation, measurement points were selected at edge intersections, with distances averaged over multiple selections. For plane fitting error, planar point cloud selection and fitting were conducted using CloudCompare. Table 6 shows that the proposed method achieves the best scores for both SE and RMS. Furthermore, 3DGS performs significantly worse in plane fitting than the MVS methods. The visualization in Figure 10 shows that point clouds generated by 3DGS and PSGR contain more noise on planar surfaces.

Table 6. Performance comparison of size error and plane fitting error. The best results are denoted in bold, while the second-best results are underlined.

Method	SPARK		GF-6		Real	
	NSE	NRMS	SE	RMS	SE	RMS
3DGS	<u>0.565</u>	0.458	0.463	0.234	0.549	0.956
MVSFormer	0.746	0.216	0.458	<u>0.178</u>	0.488	0.730
MVSFormer++	0.574	<u>0.200</u>	<u>0.417</u>	0.205	<u>0.425</u>	<u>0.734</u>
Ours	0.523	0.173	0.402	0.151	0.412	0.709

**Figure 10.** Comparison of reconstruction with the Gaussian point cloud method in the real satellite model. Regions enclosed in boxes of the same color highlight areas with noticeable differences.

For qualitative evaluation, Figure 11 presents visual comparisons of dense point clouds, while Figure 12 compares predicted depth maps for selected frames. Benefiting from enhanced feature extraction and the MCV module, MambaMVS generates more reliable points in low-texture regions and on complex surfaces such as solar panels. The depth maps exhibit sharper edges, improved continuity, and higher accuracy. Table 7 compares inference time, memory, and trainable parameters. Our method is 1.5× faster than MVSFormer++ while reducing memory and parameters, effectively balancing speed and accuracy.

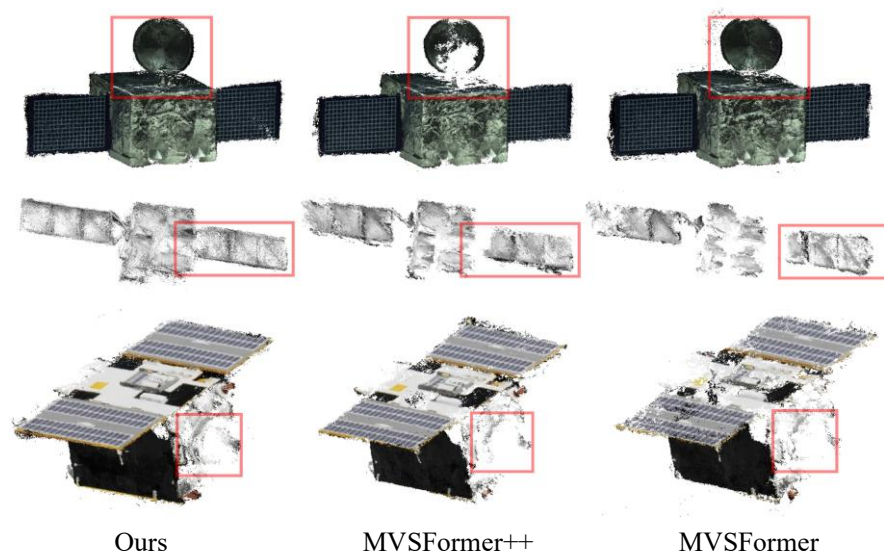


Figure 11. Qualitative comparisons of dense point clouds. Regions enclosed in boxes highlight areas with noticeable differences.

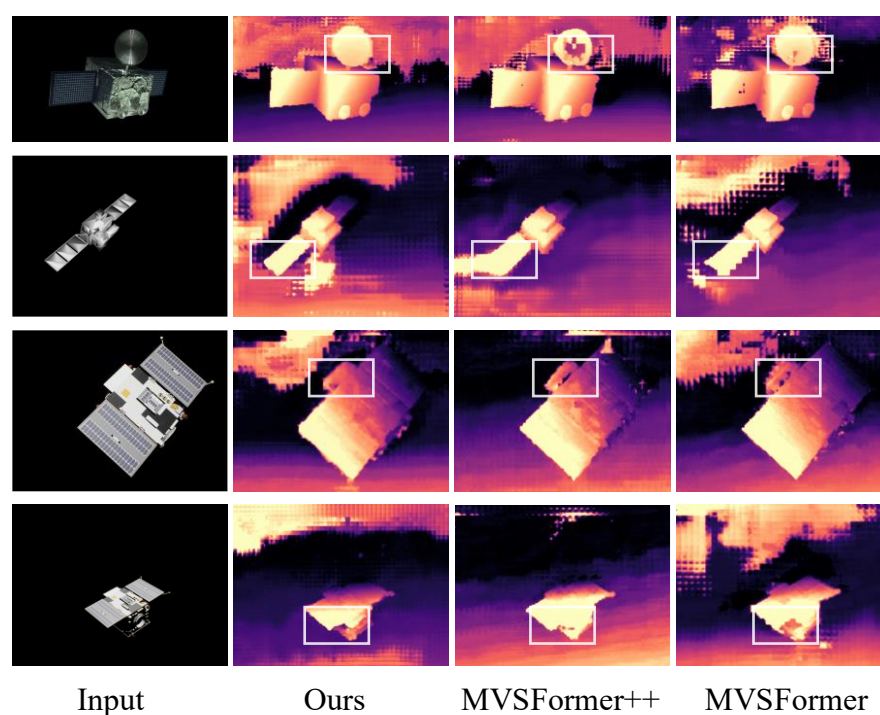


Figure 12. Qualitative comparisons of predicted depth map. Regions enclosed in boxes highlight areas with noticeable differences.

Table 7. Illustration of model memory and time costs. All timings and memory measurements were obtained on a single NVIDIA TITAN V GPU, with an input resolution of 1152×1536 and five input views, using the same fusion strategy for all methods. The best results are denoted in bold.

Method	Time (s/img)	Memory (MB)	Params. (Trainable)
MVSFormer	1.748	4970	28.013 M
MVSFormer++	1.656	5964	39.483 M
Ours	1.085	5947	39.475 M

4.4. Pose Measurement

4.4.1. Experiments Implementation Details

To validate the proposed data augmentation and keypoint self-labeling, three ablation settings (Groups A, B and C) were evaluated.

The relative pose is defined as the 6-DOF transformation from the satellite body frame to the camera frame (x-right, y-down, z-forward). Rotation is represented as Euler angles (roll, pitch, yaw) in degrees, and translation in meters. The PnP solver uses OpenCV's solvePnP Ransac with the iterative method.

- (1) Group A: The keypoint detection model was trained on 200 manually annotated images without data augmentation.
- (2) Group B: The model was trained on an augmented dataset of 745 images (25 real + 720 rendered) with manual keypoint annotation.
- (3) Group C: The same image set as Group B was used, with only 10 images annotated manually and the remaining 735 annotated automatically.

4.4.2. Performance of Pose Measurement

Figure 13 compares keypoint detection results across datasets. The model trained on the augmented dataset localizes keypoints accurately, while the non-augmented model identifies only approximate regions due to limited training data.

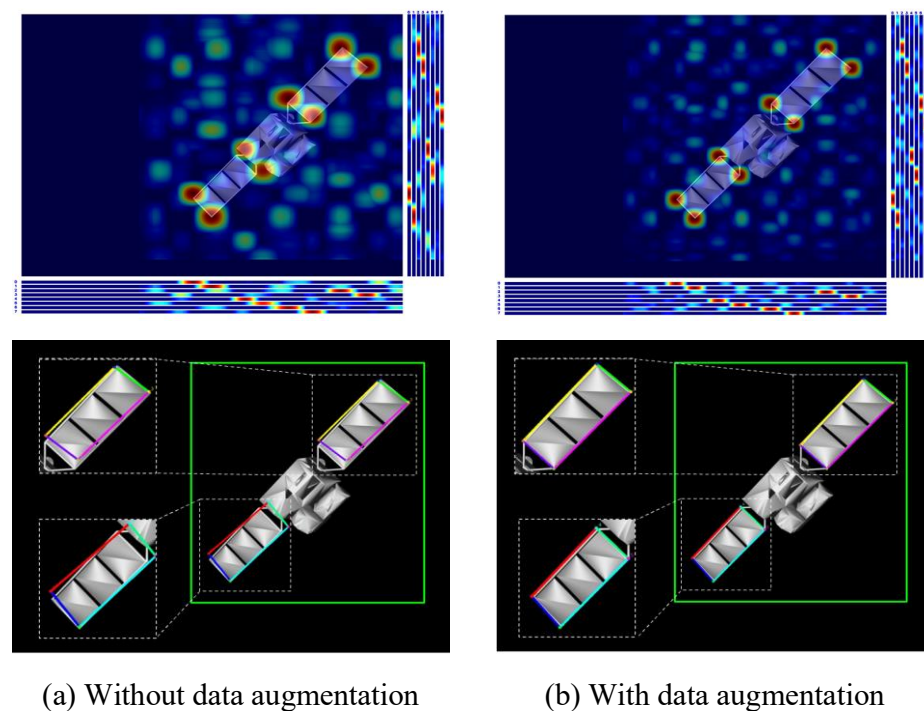


Figure 13. Comparison of keypoint detection performance with and without data augmentation. The green box highlights the target, and the inset shows the recognition differences.

Pose measurement curves under normal conditions are compared in Figure 14. Table 8 reports mean error (ME), maximum absolute error (MaxAE), and Variance of Absolute Error (VarAE) across training strategies. Group A exhibits the largest errors: rotation $MaxAE \approx 40^\circ$, $ME \approx 10^\circ$; translation $MaxAE \approx 40$ m, $ME \approx 1$ m. After data augmentation (Group B), rotation $MaxAE$ drops to $\approx 5^\circ$ (ME within 2.8°) and translation $MaxAE < 0.3$ m ($ME < 0.1$ m). With self-labeling (Group C), rotation $MaxAE$ further reduces to below 2.7° ($ME < 0.8^\circ$) and translation $MaxAE < 0.25$ m ($ME < 0.08$ m).

The experiments demonstrate that the proposed method accurately estimates the 6-DOF pose of non-cooperative satellites. Data augmentation and self-labeling achieve accuracy comparable to fully manual annotation while reducing the number of manually annotated images. Their effect under adverse imaging conditions is analyzed in Section 4.5.

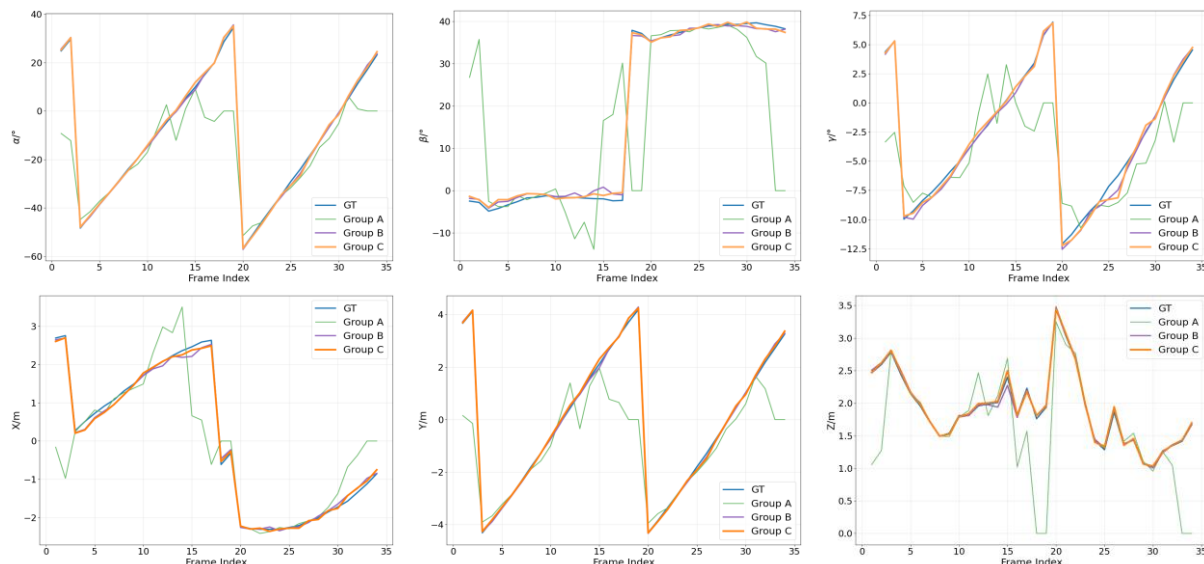


Figure 14. Pose curves under normal imaging conditions.

Table 8. Performance comparison on pose estimation under normal imaging conditions. The best results are denoted in bold, while the second-best results are underlined.

Group	α°			β°			γ°		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
A	8.934	41.957	130.578	10.502	38.839	195.582	2.690	7.818	4.842
B	0.623	1.860	0.242	0.744	2.763	0.366	0.321	1.239	0.088
C	0.713	2.659	0.359	0.687	2.075	0.290	0.320	1.905	0.124

Group	X/m			Y/m			Z/m		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
A	0.677	3.727	0.926	1.007	4.268	1.782	0.397	1.940	0.344
B	0.080	0.252	0.004	0.056	0.167	0.002	0.029	0.126	0.001
C	0.073	0.220	0.003	0.063	0.194	0.002	0.023	0.098	0.000

4.5. Robustness Experiments

4.5.1. Reconstruction Under Adverse Conditions

To evaluate robustness under low-light, high-brightness, and highly reflective conditions with suboptimal input sequences, we tested on the dataset in Figure 8 using the same metrics. The results are presented in Figure 15. The proposed method achieves more complete reconstructions with richer details than MVSFormer++ and MVSFormer across all brightness levels, particularly in highly reflective regions (real model) and complex front/low-texture side surfaces (SPARK model). Under motion blur and image subsampling, the method maintains superior reconstruction quality with minimal sensitivity to input variations.

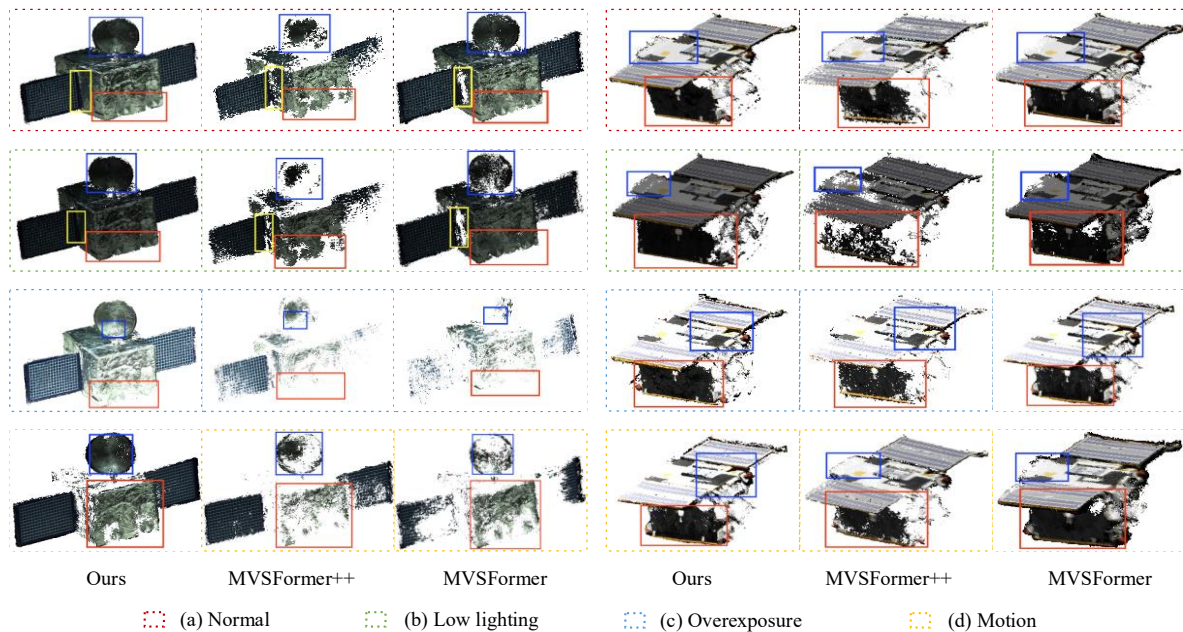


Figure 15. Reconstruction results under adverse imaging conditions. Regions enclosed in boxes of the same color highlight areas with noticeable contrast differences.

4.5.2. Pose Measurement Under Adverse Conditions

To evaluate stability under adverse conditions, two additional test groups were introduced: a low-quality imaging group with motion blur and Gaussian noise, emulating degraded acquisition, and an occlusion group with keypoints removed at increasing fractions, emulating self-occlusion. The corresponding curves and metrics are reported in Figures 16 and 17 and in Tables 9 and 10.

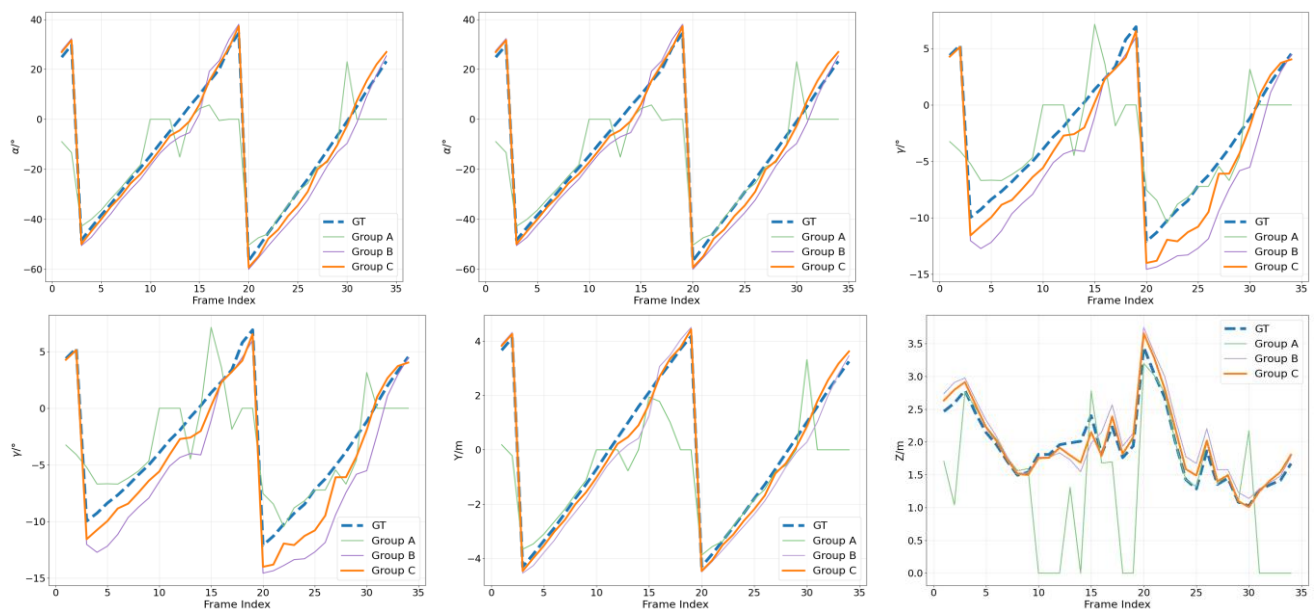


Figure 16. Pose curves under low imaging conditions (LowQuality).

Table 9. Performance comparison on pose estimation (LowQuality). The best results are denoted in bold, while the second-best results are underlined.

Group	$\alpha/^\circ$			$\beta/^\circ$			$\gamma/^\circ$		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
A	10.144	43.056	127.577	12.410	39.674	260.985	2.801	9.378	5.737
B	4.929	10.597	5.017	2.499	13.206	7.175	2.660	5.614	2.442
C	2.830	6.016	1.991	1.910	7.246	3.399	0.265	0.686	0.019
Group	X/m			Y/m			Z/m		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
A	0.947	3.794	1.272	1.163	4.343	1.772	0.683	2.010	0.557
B	0.329	1.359	0.069	0.455	1.168	0.047	0.206	0.464	0.016
C	0.231	0.703	0.037	0.265	0.686	0.019	0.115	0.323	0.006

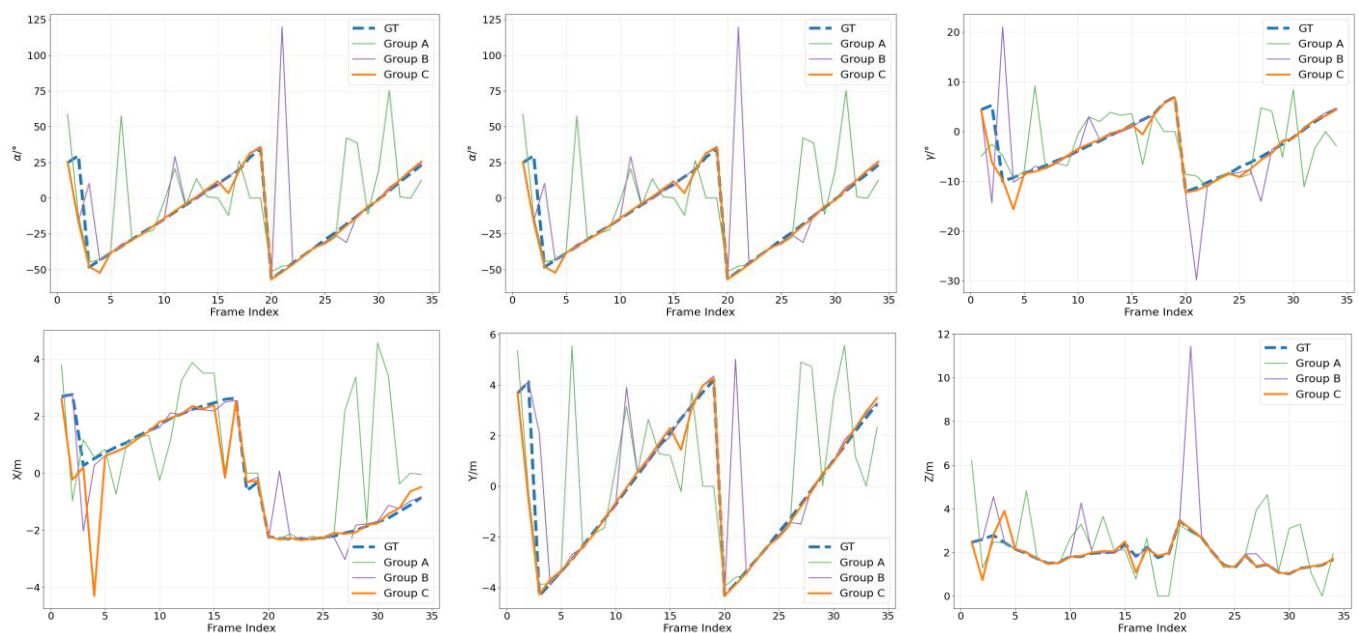


Figure 17. Pose curves under low imaging conditions (occlusion and missing).

Table 10. Performance comparison on pose estimation (occlusion and missing). The best results are denoted in bold, while the second-best results are underlined.

Group	$\alpha/^\circ$			$\beta/^\circ$			$\gamma/^\circ$		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
A	17.841	91.436	494.846	15.670	71.568	353.765	4.610	16.800	15.107
B	10.410	171.102	965.681	5.314	92.803	260.394	2.825	31.003	45.324
C	2.841	46.154	62.247	3.953	47.584	102.177	0.951	11.509	4.696
Group	X/m			Y/m			Z/m		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
A	1.282	6.289	2.796	1.697	8.412	4.039	0.899	3.736	1.120
B	0.266	2.357	0.295	0.664	8.859	3.591	0.419	8.403	2.182
C	0.401	4.813	1.024	0.259	4.614	0.616	0.143	1.862	0.161

Under occlusion, Group C attains the lowest mean error, maximum absolute error and variance on five of the six pose axes, the only exception being the x translation axis, where Group B is marginally more accurate. The contrast is largest on the α rotation axis, where the maximum absolute error of Group C (46.15°) is about half that of Group A (91.44°) and less than one third of that of Group B (171.10°).

Two features of these results are worth noting. First, occlusion degrades the worst-case rather than the mean accuracy. For Group C, the mean rotation error on the α axis is almost unchanged relative to the low-quality imaging condition (2.83° versus 2.84°), whereas the maximum error and the variance rise from 6.02° and 1.99 to 46.15° and 62.25 . This pattern is consistent with occasional ill-conditioned sets of correspondences rather than systematic bias, although the maximum errors still exceed those under normal imaging by an order of magnitude. Second, the ranking of Groups B and C depends on the imaging condition: the two are comparable under normal imaging (Table 8), whereas Group C is clearly more accurate under occlusion, indicating that the proposed data-efficiency strategy contributes mainly to robustness under incomplete observation rather than to nominal accuracy.

The conditions considered above cover the dominant degradation mechanisms for this application: specular reflection and extreme illumination (Section 4.5.1), incomplete observation (Sections 4.5.2 and 4.6) and relative angular motion, represented by its two measurable consequences, motion blur and a reduced number of usable views. Earth background interference is not treated as an experimental variable, since background rejection is a preprocessing stage of the pipeline (Section 4.1.1) and any residual background affects the camera pose estimation rather than the depth regression; extending the evaluation to unmasked Earth background imagery is identified as future work.

4.6. Real Experiment

A monocular industrial camera (Allied Vision AVT GC2000C, 12 mm focal length) and a computer were employed for data acquisition and processing. Figure 18 illustrates the setup. The camera calibration results (Table 11) show reprojection errors of 0.05 and 0.07 pixels. The computer processes and saves images in real time for subsequent reconstruction and pose measurement.

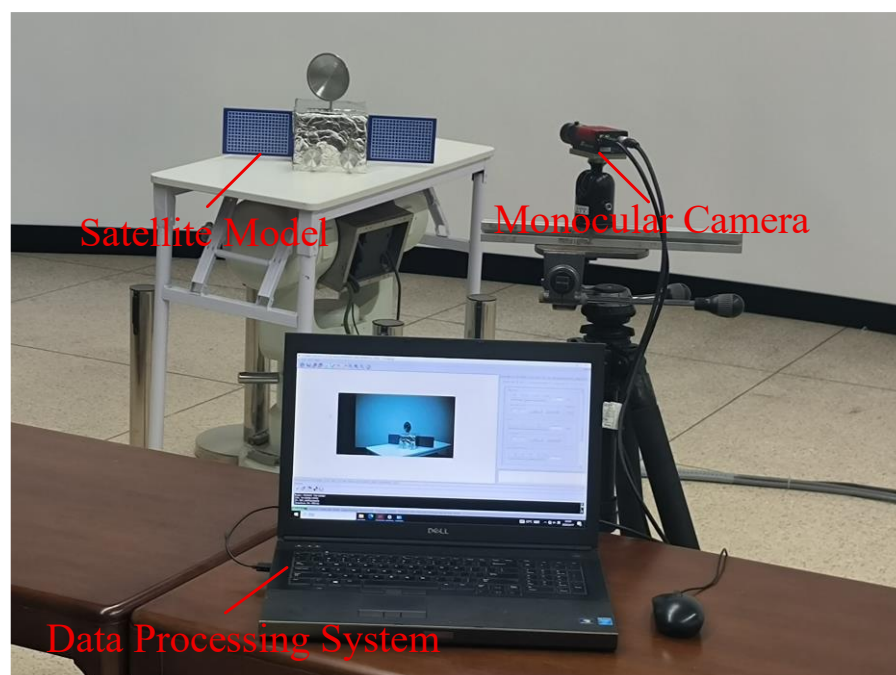


Figure 18. Configuration for measuring real models of satellites.

Table 11. Camera calibration results.

Parameter	f_x	f_y	u_0	v_0	k_1	k_2
Result	2311.550	2311.880	1013.160	544.040	−0.126	0.198

An image sequence dataset was constructed from a physical satellite model to verify reconstruction in real-world scenarios. A limited viewing angle strategy from one side was employed to simulate incomplete observation conditions caused by orbital constraints, time windows, or equipment limitations in actual missions. The calibrated camera captured 64 close-range, multi-angle images, simulating on-orbit measurement by a service spacecraft. The model's main body is wrapped in aluminum foil to mimic metallic thermal insulation. The solar panels feature repetitive textures, and the antenna is a textured metal disc. Exposure was reduced, and segmentation masks set non-target areas to black to simulate the high-contrast space environment. Representative examples are shown in Figure 19.



Figure 19. Image sequence of the real satellite model.

Relative size error and plane fitting error serve as quantitative metrics; depth maps and point clouds provide qualitative comparisons. Ground truth dimensions were measured with a vernier caliper (accuracy ± 0.02 mm); point cloud analysis used CloudCompare.

COLMAP was employed to estimate camera poses, depth hypothesis ranges, and neighboring image pairs. The results are visualized in Figures 20–22, which present depth maps and dense point cloud reconstruction on the physical model.

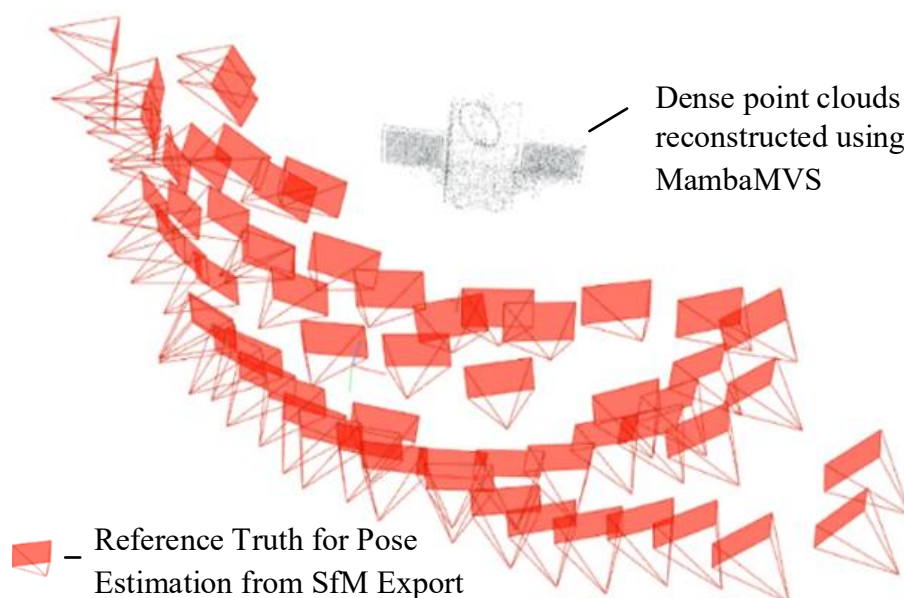


Figure 20. Visualization of SfM results for the real satellite model. Regions enclosed in boxes highlight areas with noticeable contrast differences.

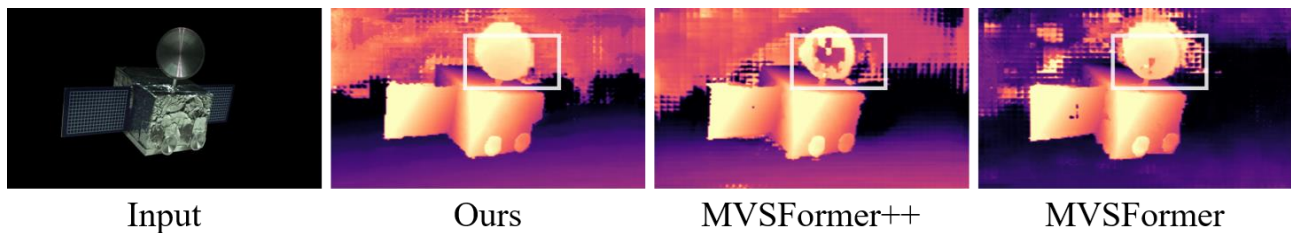


Figure 21. Comparison of depth maps of satellite physical models. Regions enclosed in boxes highlight areas with noticeable contrast differences.

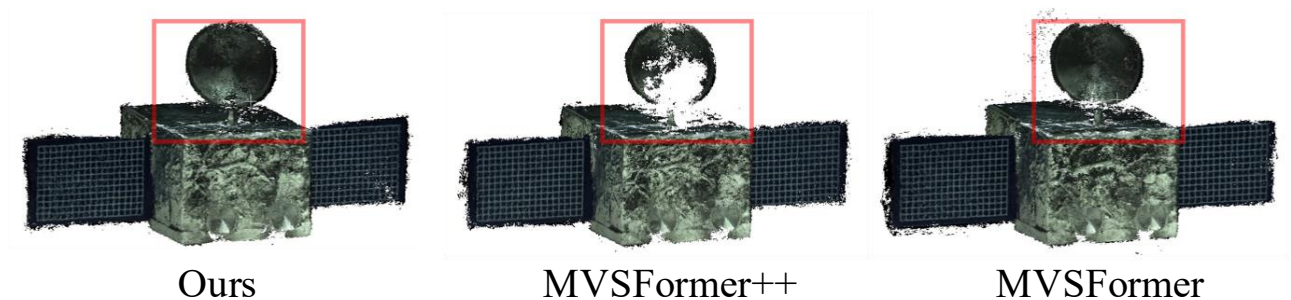


Figure 22. Reconstruction results of the real satellite model.

All algorithms share the same training parameters and take five images as input per depth map estimation, with the same depth fusion strategy. The proposed method generates depth maps with better continuity and point clouds with greater completeness, particularly in low-texture regions such as solar panels. Real satellite pose measurement and image back-projection are shown in Figure 23.

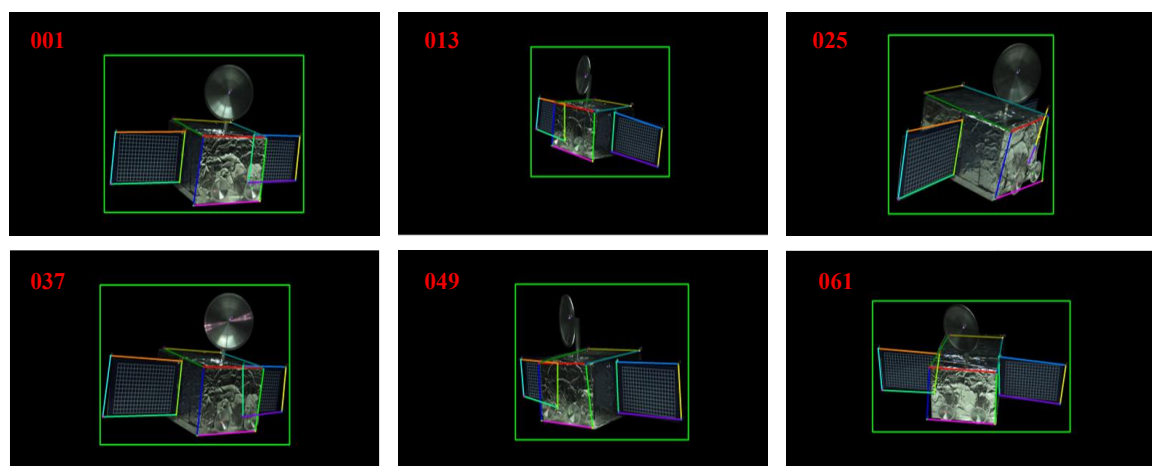


Figure 23. Real satellite pose measurement and image back-projection. The green box highlights the target, and the inset shows the recognition differences.

Figure 24 compares real and measured poses. As shown in Table 12, for rotation, all three Euler angles achieve MaxAE within 0.6° and ME within 0.2° . For translation, ME is 0.012 m (X), 0.032 m (Y), and 0.008 m (Z), with MaxAE below 0.05 m in all directions. The online pose measurement operates at approximately 30 Hz, satisfying real-time requirements for on-orbit applications. These results demonstrate accurate 6-DOF pose estimation for non-cooperative satellites.

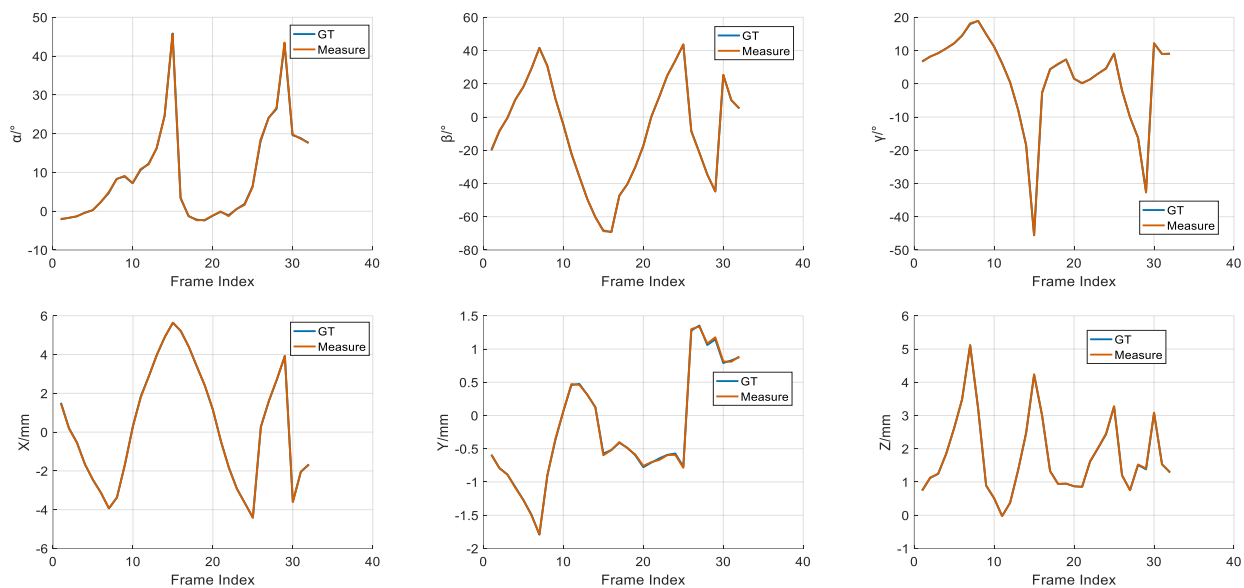


Figure 24. Pose curves of a real satellite.

Table 12. Evaluation results of attitude measurement of physical satellite models. The best results are denoted in bold, while the second-best results are underlined.

Euler Angle Evaluation Item	$\alpha/^\circ$			$\beta/^\circ$			$\gamma/^\circ$		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
Experimental result	0.165	0.568	0.013	0.154	0.518	0.017	0.081	0.278	0.005
Translation Evaluation item	X/m			Y/m			Z/m		
	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE	<u>ME</u>	MaxAE	VarAE
Experimental result	0.012	0.050	1.372×10^{-4}	0.013	0.032	7.123×10^{-5}	0.008	0.026	5.821×10^{-5}

5. Conclusions

This paper presents an on-orbit 3D reconstruction and pose measurement method for space non-cooperative satellites using monocular vision. The method couples an end-to-end MVS network for dense 3D reconstruction, named MambaMVS, with a data-efficient pose measurement strategy. The MCV Module leverages the linear complexity of structured state space models for efficient cost volume regularization, achieving fast inference with global context modeling. An improved focused linear attention module enhances intra-view and cross-view feature representation, mitigating the performance degradation inherent to linear attention. For pose measurement, the dense point cloud reconstructed by MambaMVS provides the 3D coordinates of the semantic keypoints. 3D Gaussian Splatting-based novel view augmentation combined with reprojection-based automatic annotation reduces the number of manually annotated images from 745 to 10. The resulting pose accuracy is comparable to that obtained with fully manual annotation and is further improved under the adverse conditions examined in this work. Simulation experiments across diverse challenging conditions, including varying illumination, reflective surfaces and motion blur, demonstrate that MambaMVS achieves state-of-the-art reconstruction quality in point cloud accuracy, dimensional measurement, and plane fitting error while operating $1.5\times$ faster than current leading methods. Physical satellite experiments validate the complete reconstruction-and-measurement pipeline, achieving 6-DOF pose estimation with rotation ME within 0.2° and translation ME below 0.01 m, confirming effectiveness for on-orbit applications. The pipeline separates computation between its two operating phases: reconstruction and keypoint model training are performed offline, whereas the online stage requires only the keypoint detector and the PnP solver. The

present implementation nevertheless runs on ground-based GPUs, while onboard deployment is constrained by the available processing capability, memory and power budget. Future work will therefore target pruning, quantization and distillation of the detector, together with benchmarks on space-grade embedded accelerators, so that the online stage can be executed on resource-constrained platforms.

Author Contributions: Conceptualization, X.P., Z.L. and B.F.; methodology, X.P. and X.Q.; software, X.P., X.Q. and Q.L.; validation, X.P. and X.Q.; formal analysis, X.Q., J.S. and Q.L.; investigation, X.P.; resources, X.P. and X.Q.; data curation, X.P. and X.Q.; writing—original draft preparation, X.Q. and X.P.; writing—review and editing, X.P., Q.L., Z.L. and B.F.; visualization, X.P.; supervision, X.P., Z.L. and B.F.; project administration, X.P.; funding acquisition, X.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC), under Grant Nos. 62105015, Aviation Science Foundation 2023, Project No. 20230046051003, and “the Fundamental Research Funds for the Central Universities”.

Data Availability Statement: The data used in this study consist of three parts. (1) The SPARK dataset is publicly available at <https://cvi2.uni.lu/spark-2021> (accessed on 31 January 2019). (2) The DTU dataset used for pre-training is publicly available at <https://roboimagedata.compute.dtu.dk> (accessed on 24 March 2025). (3) The Customized Space Target (CST) dataset generated in this study, including the image sequences of the simulated and physical satellite models and the corresponding camera parameters, is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bai, S.; Wang, Y.; Liu, H.; Sun, X. Finite-thrust Lambert transfer based on multistage constant-vector thrust control. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4947–4967.
2. Gu, Y.; Liu, H.; Pan, T.; Bai, S.; Liu, J.; Wu, Y.; Wu, G. Ensemble time freedom heuristic and intelligent optimization algorithm for relay satellites scheduling considering multi-type task requirements. *Expert Syst. Appl.* **2026**, *316*, 131815.
3. Peng, J.; Xu, W.; Liang, B.; Wu, A.G. Pose measurement and motion measurement of space non-cooperative targets based on laser radar and stereo-vision fusion. *IEEE Sens. J.* **2019**, *19*, 3008–3019.
4. Du, R.; Liao, W.; Zhang, X. Feasibility analysis of angles-only navigation algorithm with multisensor data fusion for spacecraft noncooperative rendezvous. *Astrodynamics* **2023**, *7*, 179–196.
5. Yu, S.U.; Zhang, Z.; Yuan, M.; Xu, T.; Deng, H.; Wang, J. A point cloud fusion method for space target 3D laser point cloud and visible light image reconstruction method. *J. Deep Space Explor.* **2021**, *8*, 534–540.
6. Xie, W.; Ren, X.; Wang, C.; Jiang, L.; Wang, Z. MVS-STRNet: Cross-view space target recognition from multi-view stereo. In *Proceedings of the 4th International Conference on Computer Vision and Data Mining (ICCVDM 2023)*; SPIE: Bellingham, WA, USA, 2024; Volume 13063, p. 130630B.
7. Gao, K.; Nouduri, K.; Bunyak, F.; Massaro, R.; AliAkbarpour, H.; Fraser, J.; Seetharaman, G.; Palaniappan, K. Local feature performance evaluation for structure-from-motion and multi-view stereo using simulated city-scale aerial imagery. *IEEE Sens. J.* **2021**, *21*, 11615–11627.
8. Tian, P.; Yao, M.; Xiao, X.; Xi, Y.; Liu, H.; Zheng, B.; Cao, T.; Cui, H. 3-D semantic terrain reconstruction of monocular close-up images of Martian terrains. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 46007161.
9. Fu, T.; Zhou, Y.; Wang, Y. Neural field-based space target 3D reconstruction with predicted depth priors. *Aerospace* **2024**, *11*, 997.
10. Kerbl, B.; Kopanas, G.; Leimkühler, T. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 139.
11. Deng, J.; Xie, P.; Zhang, L.; Cong, Y. ISAR-NeRF: Neural radiance fields for 3-D imaging of space target from multiview ISAR images. *IEEE Sens. J.* **2024**, *24*, 11705–11722.

12. Lei, T.; Liu, X.F.; Cai, G.P.; Liu, Y.M.; Liu, P. Pose measurement of a noncooperative target based on monocular visual SLAM. *Int. J. Aerosp. Eng.* **2019**, *2019*, 9086891.
13. Zhao, H.; Yao, M.; Zhao, Y.; Jiang, Y.; Zhang, H.; Xiao, X.; Gao, K. M2CS: A multimodal and campus-scapes dataset for dynamic SLAM and moving object perception. *J. Field Robot.* **2024**, *41*, 1733–1756.
14. Peng, J.; Xu, W.; Yuan, H. An efficient pose measurement method of a space non-cooperative target based on stereo vision. *IEEE Access* **2017**, *5*, 22344–22362.
15. Huang, K.; Luo, Z.; Quan, J.; Liang, B.; Lin, J. Non-cooperative target recognition technology based on line structured light. In *Chinese Intelligent Systems Conference*; Springer: Singapore, 2024; pp. 95–110.
16. Zhang, P.; Wu, D.; Baoyin, H. Real-time hybrid method for maneuver detection and estimation of non-cooperative space targets. *Astrodynamics* **2024**, *8*, 437–453.
17. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
18. Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y. TransMVSNet: Global context-aware multi-view stereo network with transformers. In Proceedings of the 39th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8575–8584.
19. Yang, H.; Xiao, X.; Yao, M.; Xiong, Y.; Cui, H.; Fu, Y. PVSPE: A pyramid vision multitask transformer network for spacecraft pose estimation. *Adv. Space Res.* **2024**, *74*, 1327–1342.
20. Yao, Y.; Luo, Z.X.; Li, S.W.; Fang, T.; Quan, L. MVSNet: Depth inference for unstructured multi-view stereo. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 785–801.
21. Yang, J.; Mao, W.; Alvarez, J.M.; Liu, M. Cost volume pyramid based depth inference for multi-view stereo. In Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4877–4886.
22. Liu, Y.; Wu, H.; Wang, C.; Wei, Y.; Ren, M.; Feng, T. Real-time dense construction with deep multiview stereo using camera and IMU sensors. *IEEE Sens. J.* **2023**, *23*, 19648–19659.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
24. Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; Wang, X. MVSTER: Epipolar transformer for efficient multi-view stereo. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 573–591.
25. Zhang, K.; Liu, M.; Zhang, J.; Dong, Z. PA-MVSNet: Sparse-to-dense multi-view stereo with pyramid attention. *IEEE Access* **2021**, *9*, 27908–27915.
26. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
27. Cao, C.J.; Ren, X.L.; Fu, Y.W. MVSFormer: Multi-view stereo by learning robust image features and temperature-based depth. *arXiv* **2022**, arXiv:2208.02541.
28. Cao, C.; Ren, X.; Fu, Y. MVSFormer+: Revealing the devil in transformer’s details for multi-view stereo. *arXiv* **2024**, arXiv:2401.11673.
29. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In Proceedings of the 36th Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 16344–16359.
30. Katharopoulos, A.; Vyas, A.; Pappas, N. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 5156–5165.
31. Yao, Y.; Luo, Z.X.; Li, S.W.; Shen, T.W.; Fang, T.; Quan, L. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5520–5529.
32. Chen, W.; Xu, H.; Zhou, Z.; Liu, Y.; Sun, B.; Kang, W.; Xie, X. CostFormer: Cost transformer for cost aggregation in multi-view stereo. *arXiv* **2023**, arXiv:2305.10320.
33. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
34. Hong, R.; Cai, Z.; Yang, J.; Da, F. LV2DMOT: Language and visual multimodal feature learning for multiobject tracking. *IEEE Sens. J.* **2025**, *25*, 7482–7495.

35. Wang, Z.; Zheng, J.Q.; Zhang, Y.; Cui, G.; Li, L. Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation. *arXiv* **2024**, arXiv:2402.05079.
36. Liu, J.; Yu, R.; Wang, Y.; Zheng, Y.; Deng, T.; Ye, W.; Wang, H. Point Mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv* **2024**, arXiv:2403.06467.
37. Li, Y.; Yang, W.; Fei, B. 3DMambaComplete: Exploring structured state space model for point cloud completion. *arXiv* **2024**, arXiv:2404.07106.
38. Shen, Q.; Wu, Z.; Yi, X.; Zhou, P.; Zhang, H.; Yan, S.; Wang, X. Gamba: Marry Gaussian splatting with Mamba for single view 3D reconstruction. *arXiv* **2024**, arXiv:2403.18795.
39. Jiang, J.; Liu, Q.; Liu, H.; Yu, H.; Wang, L.; Chen, J.; Ma, H. MVSMamba: Multi-view stereo with state space model. In Proceedings of the 39th Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; pp. 160679–160706.
40. Qu, X.; Liu, Z.; Song, J. Efficient multi-view stereo for space target with Mamba-based cost aggregation. In Proceedings of the 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), 2025; pp. 871–877.
41. Li, L.; Liu, G.; Xu, F.; Liao, W.; Zhao, W. Mamba-enhanced multi-view stereo: Geometry-aware feature fusion and 3D cost volume regularization. *Appl. Soft Comput.* **2026**, *195*, 115074.
42. Volpe, R.; Palmerini, G.B.; Sabatini, M. A passive camera based determination of a non-cooperative and unknown satellite's pose and shape. *Acta Astronaut.* **2018**, *151*, 805–817.
43. Volpe, R.; Circi, C.; Sabatini, M.; Palmerini, G.B. GNC architecture for an optimal rendezvous to an uncooperative tumbling target using passive monocular camera. *Acta Astronaut.* **2022**, *196*, 380–393.
44. Atallah, M.; Kassem, A.H.; Sharif, M.A. Twistor-based unscented Kalman filter for robust relative pose estimation of spacecraft. *Acta Astronaut.* **2025**, *228*, 113–128.
45. Gui, Y.; Li, Z.; Chen, X.; Wang, H. SPTN: Space object tracking transformer network for long-range space target tracking. *IEEE Sens. J.* **2025**, *25*, 4521–4533.
46. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
47. Han, D.; Pan, X.; Han, Y.; Song, S.; Huang, G. Flatten Transformer: Vision Transformer using focused linear attention. In Proceedings of the 19th IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 5938–5948.
48. Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; Chen, K. RTMPose: Real-time multi-person pose measurement based on MMPose. *arXiv* **2023**, arXiv:2303.07399.
49. Musallam, M.A.; Ismaeil, K.A.; Oyedotu, O. SPARK: Spacecraft recognition leveraging knowledge of space environment. *arXiv* **2021**, arXiv:2104.05978.
50. Schonberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4104–4113.
51. Galliani, S.; Lasinger, K.; Schindler, K. Massively parallel multi-view stereopsis by surface normal diffusion. In Proceedings of the 15th IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 873–881.

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.