A 3D Hand Pose Estimation Framework for Crewed Space Mission under Limited Sensing Conditions

Ruiqi Xun*†, Xingyu Zhu‡, Haoshuai Mu*†, Ziang Qu*†, Liang Chang*†§
* Innovation Academy for Microsatellites, Chinese Academy of Sciences, Shanghai, China
† University of Chinese Academy of Sciences, Beijing, China
‡ School of Artificial Intelligence, Jilin University, Changchun, China
Emails: xunruiqi22@mails.ucas.ac.cn, xingyu24@mails.jlu.edu.cn, {muhaoshuai, quziang}@microsate.com
§ Corresponding author: changl_researchlab@163.com

***Abstract*—During on-orbit missions, astronauts interact with cargo packages, control panels, and experimental payloads primarily through hand motions. 3D hand pose estimation is essential for automated operation monitoring and human–machine interaction. However, spacecraft impose strict constraints on sensing systems, where depth sensors are impractical, and perception must rely on limited RGB cameras. This sensing limitation poses a fundamental challenge to the accurate estimation of 3D hand pose. To address this challenge, we propose SpaceRGB-HandPose Net(SHPN), an RGB-based framework tailored to the sensing constraints of crewed spacecraft. SHPN enhances geometric perception from RGB images by injecting a pseudo-depth modality. It explicitly addresses the mismatch between pseudo-depth and metric depth via DTOR and ADRM alignment modules. Furthermore, SHPN introduces a cross-modal fusion network to integrate appearance and relative depth information for robust pose estimation. Experiments on the DexYCB dataset show that SHPN significantly outperforms RGB-based baselines and achieves performance comparable to RGB-D methods using only RGB input, demonstrating its applicability to realistic crewed spaceflight scenarios.**

***Index Terms*—Crewed space missions, 3D Hand Pose Estimation, Astronauts, Human-machine interaction**

I. Introduction

With the rapid advancement of modern crewed spaceflight, astronaut operations are increasingly constrained by limited time resources, high operational costs, and stringent safety requirements. Motivated by recent progress in artificial intelligence, researchers are exploring the integration of AI technologies into crewed space engineering to support astronauts during on-orbit operations, thereby reducing human-induced errors and enhancing overall mission efficiency.

China has been exploring low-cost and intelligent technologies for its space station, proposing concepts for intelligent spacecraft such as Haolong and Qingzhou Fig. 1. Astronauts interact with cargo packages, control panels, and experimental payloads primarily through hand motions. Hand pose estimation is essential for automated operation monitoring and task management. It further facilitates the detection of abnormal or constrained motion patterns, enabling fatigue assessment, cognitive workload monitoring, and operational risk warning.

Most state-of-the-art 3D hand pose estimation methods rely on RGB-D images and laser point clouds as input sensing modalities [1]–[7]. In contrast, crewed spacecraft are subject to strict constraints on sensing resources, including volume, power consumption, field of view, redundancy, and maintenance complexity. In such spacecraft, only RGB cameras are available as the perception system (Fig. 1). The inherent limitations of RGB imagery, such as appearance ambiguity, color similarity, and the lack of explicit geometric cues, pose significant challenges for accurate hand pose estimation.

This motivates a key research question: how can RGB-only approaches achieve performance comparable to RGB-D-based methods under realistic spaceflight constraints? A straightforward solution is to generate a pseudo-depth map from an RGB image using a pretrained depth estimation model, and then combine the pseudo-depth map with the RGB image for hand pose estimation. However, this strategy faces two major challenges. (1) There exist significant mismatches in depth distribution and physical scale between pseudo-depth estimates and metric depth captured by depth sensors. (2) It remains non-trivial to establish a unified representation across modalities and to effectively fuse heterogeneous features for joint pose prediction.

To tackle the above challenges, we propose SpaceRGB-HandPoseNet, composed of a pseudo-depth modality injection mechanism and a cross-modal fusion network. The pseudo-depth modality injection mechanism enhances the geometric representation capability of RGB-only perception systems in crewed spacecraft. To address the mismatch between pseudo-depth and RGB modalities, DTOR and ADRM are introduced to achieve pixel-level alignment, allowing pseudo-depth to be effectively injected as an auxiliary geometric cue. After that, the cross-modal fusion network integrates RGB features with pseudo-depth features through feature alignment and complementary learning, enabling robust 3D hand pose prediction under limited sensing conditions.

We evaluate the proposed method on the challenging DexYCB dataset and investigate the impact of pseudo-depth quality on RGB–depth alignment. Experimental results demonstrate that, using only a single RGB image as input, SHPN outperforms other methods that rely solely on RGB modality and achieves performance comparable to state-of-the-art approaches that use RGBD inputs, thereby validating the effectiveness of the proposed method and its potential for practical engineering applications.

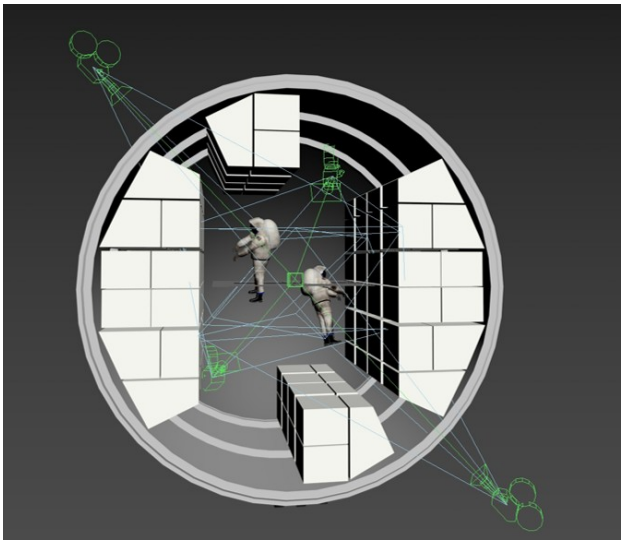

Fig. 1. Simulated view of the camera placement and field of view on the spacecraft. To ensure completeness of the captured field of view, the industrial-grade RGB cameras on the spacecraft are placed diagonally, and a gimbal camera is installed at the entrance of the spacecraft to supplement the information.

The main contributions of this work are summarized as follows:

1. We identify the key challenges of 3D hand pose estimation in crewed spacecraft, in contrast to traditional laboratory and ground-based environments.

2. To address these challenges, we propose a novel model termed SHPN, which consists of a pseudo-depth modality injection mechanism and a cross-modal fusion network. The injection mechanism mitigates modality mismatch by incorporating two alignment modules, DTOR and ADRM, while the fusion network effectively integrates relative depth information derived from the pseudo-depth map.

3. Extensive experiments on DexYCB demonstrate that SHPN outperforms RGB-based baseline methods. Furthermore, ablation studies and qualitative analyses validate the effectiveness of the proposed components.

II. Related Work

A. Pose Estimation in Crewed Space Missions

The International Space Station (ISS) has now been operating in orbit for nearly 30 years. The Chinese Space Station has been in service for over four years and is expected to remain operational for an extended period. Due to strict engineering constraints on onboard hardware, existing astronaut pose estimation studies have primarily focused on 2D human pose detection and 2D hand gesture recognition. Wu et al. [8] proposed ROpenPose, a lightweight single-branch network for on-orbit 2D human pose estimation. By replacing the VGG-19 backbone with MobileNets, reducing convolution kernel sizes for refined downsampling, and introducing residual connections to mitigate gradient vanishing, ROpenPose achieves efficient inference under limited computational resources. To monitor extravehicular activities, Liu et al. [9] proposed AstroPose, which performs remote calibration of a monocular camera using objects of known dimensions visible on the space station. Based on the estimated camera intrinsics and extrinsics, astronaut poses are initialized via a perspective projection-based PnP formulation and further refined through nonlinear optimization, reprojection error minimization, and Kalman filtering to improve accuracy and temporal smoothness. Ouyang et al. [10] introduced SpacesuitPose, which explicitly models inter-joint structural dependencies by vectorizing joint heatmaps, capturing cross-joint correlations, and applying global weighting via a relational matrix. A skeleton structure loss is further designed to constrain local limb connectivity, ensuring structurally consistent pose predictions. Xun et al. [11] proposed Astronaut Pose Net, which constructs a spatial feature volume from multi-view RGB images using voxel projection. The feature volume is decomposed into orthogonal feature planes, and 3D human poses are obtained by fusing inference results from different planes. To address posture variations under microgravity, a biomechanical encoding scheme is introduced to constrain the astronaut's occupancy within the spatial feature volume, mitigating neutral-pose expansion caused by root joint ambiguity. With respect to hand pose estimation in space applications, Gao et al. [12] designed a network architecture based on hand anatomical structure, employing six parallel branches to regress 2D poses of the palm and five fingers. A fully connected layer is then used to infer 3D joint coordinates from 2D predictions, enabling monocular teleoperation of dexterous robotic manipulators onboard space stations. However, lifting 2D hand poses to 3D inherently constitutes an ill-posed problem [5], [13], resulting in unstable predictions that are unsuitable for real-world engineering scenarios.

B. 3D Hand Pose Estimation

3D hand pose estimation has long been a fundamental yet challenging problem in the computer vision community. Early studies focused on detecting 2D hand keypoints from a single RGB image [14]. Subsequent works attempted to directly regress 3D hand poses from RGB images using deep learning approaches [15], [16]. Although these methods demonstrated partial effectiveness in alleviating 2D projection ambiguity, from a mathematical perspective, the inherent ambiguity of lifting 2D observations to 3D space cannot be fully resolved using a single image alone. As a result, such approaches are prone to overfitting to viewpoint distributions present in the training data, leading to limited generalization.

To address these limitations, later studies incorporated additional prior knowledge into RGB-based 3D hand pose estimation. Iqbal et al. [17] proposed preserving 2D pixel coordinates of keypoints while predicting their relative depths, thereby avoiding scale ambiguity and geometric irreversibility associated with directly regressing absolute 3D coordinates from monocular RGB images. Other researchers formulated 3D hand pose estimation as an optimization problem constrained by parametric hand models [18], [19]. Park et al. [20] designed a framework consisting of feature-injection transformers and self-enhancement transformers to explore the

relationship between occluded regions and hand pose, leveraging occlusion cues as auxiliary information to enhance feature representations. Chen et al. [21] decomposed camera-space mesh recovery into two sub-tasks: extracting joint landmarks and contours from a single image to provide 2D cues, and reconstructing a root-relative 3D mesh using semantic relationships among joints. The root position is then registered to camera space by aligning the reconstructed mesh with the 2D cues. Lin et al. [22] employed a transformer encoder to jointly model vertex–vertex and vertex–joint interactions, enabling simultaneous prediction of 3D joint coordinates and mesh vertices. Recent mainstream research increasingly relies on richer sensing modalities to improve robustness and accuracy. Moon et al. [1] proposed a voxel-based 3D CNN that converts 2D depth images into 3D voxel grids and performs volumetric convolution to estimate 3D hand poses. Rezaei Rastgoo et al. [5] introduced TriHorn-Net, which decomposes 3D hand pose estimation into estimating 2D joint locations and their corresponding depth values in depth image space. The network employs two branches to compute joint attention maps and a third branch to estimate depth information, enabling accurate hand pose estimation from a single depth image. Jiang et al. [23] explored 3D hand pose estimation using event camera streams, addressing sparse annotation challenges through a sparse supervision framework and mitigating the domain gap between synthetic and real data via the EvRealHands dataset.

III. Methodology

As illustrated in Fig. 2, we propose a multi-modal hand pose estimation framework designed for RGB-only sensing conditions. The proposed network consists of three main components: an RGB branch for appearance-driven joint localization, a Pseudo-Depth branch for geometrically consistent joint initialization, and a cross-modal fusion module that progressively refines joint predictions through point-cloud–guided feature interaction and transformer-based modeling.

A. Pseudo-Depth Modality Injection

Due to the lack of explicit geometric perception in RGB-only sensing scenarios, inspired by recent advances in monocular depth estimation [24], we introduce a pseudo-depth modality injection mechanism to enhance geometric awareness under limited sensing conditions. Specifically, the proposed mechanism consists of two stages: pseudo-depth modality generation and pseudo-depth modality enhancement. Given an input RGB image $I_{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$, we first feed it into a pretrained [24] method $f_{\text{da}}(\cdot)$ to obtain an initial pseudo-depth estimation $D_{\text{init}} = f_{\text{da}}(I_{\text{rgb}})$. Although D_init produces high-quality relative depth maps, its output is represented in a normalized integer range, which lacks physical scale consistency with metric depth captured by depth sensors. This mismatch in depth distribution and scale renders the predicted depth unsuitable for direct integration into geometry-aware preprocessing steps, such as cube-based cropping and 3D pose regression Fig. 3(b). To overcome this limitation, we propose a two-stage depth preprocessing pipeline.

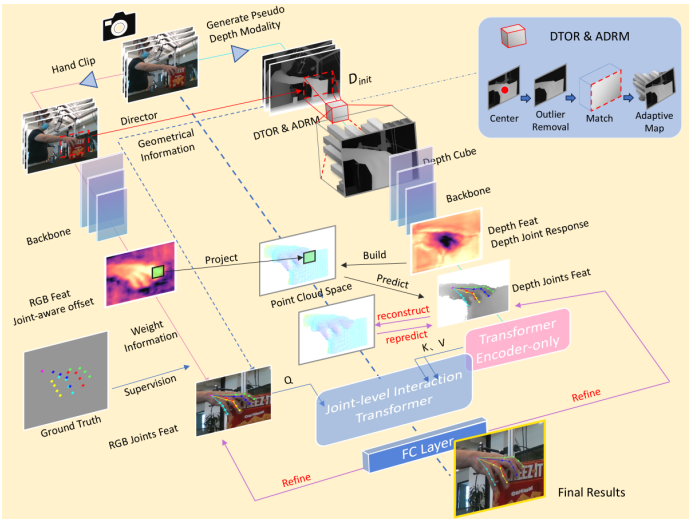

Fig. 2. Overview of the SHPN framework with an RGB branch, a pseudo-depth branch, and a single fusion stage. The fusion stage stacks two joint-level interaction modules to progressively refine joint representations.

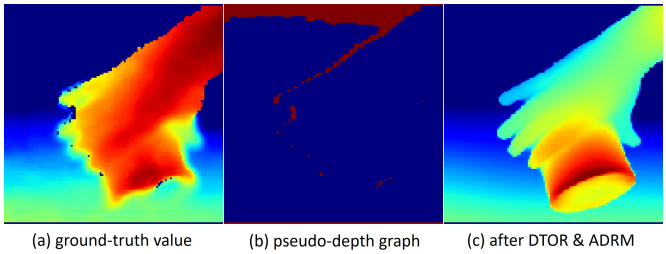

Fig. 3. Different initial depth information graphs.

Using the estimated hand center, we define a metric three-dimensional cube centered at $\mathbf{c}$:

$$\mathcal{C} = \{\mathbf{p} : |\mathbf{p} - \mathbf{c}| \leq \tfrac{\mathbf{L}}{2}\}. \tag{1}$$

where the inequality is applied element-wise. Here, $p = (x, y, z)$ denotes a point within the cube, and $L = (L_x, L_y, L_z)$ denotes the side lengths. In our experiments, we set $L = 250$ mm. The first stage, termed Distance-Threshold Outlier Removal (DTOR), takes the depth map value D_c on the same pixel of the hand center c as the reference and suppresses extreme depth outliers using a fixed threshold α, aiming to mitigate the heavy-tailed distribution commonly observed in the predicted depth maps:

$$D'(u,v) = \begin{cases} D_{\text{init}}, & |D_{\text{init}} - D_c| \leq \frac{\alpha}{2}, \\ D_c + \frac{\alpha}{2}, & D_{\text{init}} > D_c + \frac{\alpha}{2}, \\ D_c - \frac{\alpha}{2}, & D_{\text{init}} < D_c - \frac{\alpha}{2}. \end{cases} \tag{2}$$

In the experiment, we set the value of α to 40000. DTOR extracts a hand-centered 3D region with a fixed physical size, effectively decoupling hand geometry from background interference and camera distance variation. As a result, the dominant geometric structure of the hand is more clearly emphasized.

Although DTOR constrains extreme depth values by thresholding around the depth map center depth D_c, the resulting

depth map $D'(u,v)$ remains in the original metric space. To obtain a normalized and scale-consistent representation, we further apply Adaptive Depth Range Mapping (ADRM) to rescale the depth values into a canonical cube-centered range. Specifically, given the DTOR-processed depth map $D'(u,v)$, ADRM linearly maps the depth values within the predefined cube and recenters the cube at the estimated depth center:

$$\tilde{D}(u,v) = D'(u,v) \cdot \frac{L_z}{\alpha} + c_z \tag{3}$$

Here, L_z denotes the side length of the predefined cube along the depth axis. ADRM maps the relative depth representation into a canonical range, providing a scale-invariant and learning-friendly depth representation.

Through the proposed preprocessing and enhancement pipeline, the refined pseudo-depth map is accurately aligned with the original RGB image at the pixel level, forming a geometrically consistent and well-registered multi-modal input pair for subsequent cross-modal feature fusion.

B. SpaceRGB-HandPose Net

1) RGB Branch: The RGB branch aims to extract discriminative appearance features and provide reliable spatial cues for joint localization in the image plane. Given an input RGB image $\mathbf{I}^{rgb}$, we employ backbone network to extract RGB feature maps and pixel-wise joint-related predictions:

$$\{\mathbf{F}^{rgb}, \mathbf{O}^{rgb}\} = \text{Backbone}_{rgb}(\mathbf{I}^{rgb}) \tag{4}$$

where $\mathbf{F}^{rgb} \in \mathbb{R}^{C\times H\times W}$ denotes the RGB feature map, and $\mathbf{O}^{rgb}$ represents pixel-wise joint response maps, including joint response intensity, offset vectors, and confidence weights, since RGB images lack explicit depth cues, pixel-wise 3D offset voting is not conducted in the RGB branch, and is instead carried out only in the pseudo-depth branch where geometric information is available, the detailed definition of these components will be further explained in the depth branch description.

To supervise pixel-level predictions, we construct ground-truth feature maps $\mathbf{G}^{rgb}$ by projecting ground-truth 3D joint locations onto the image plane and generating corresponding response and offset representations. A pixel-wise regression loss is applied:

$$\mathcal{L}_{pix} = \|\mathbf{O}^{rgb} - \mathbf{G}^{rgb}\|_1 \tag{5}$$

which encourages the backbone to learn joint-aware spatial representations. Then the predicted joint response maps, RGB features are aggregated from the pixel space to the joint space. Specifically, joint-aligned RGB features are obtained via weighted spatial pooling, resulting in joint-level representations $\mathbf{J}^{rgb}$ which encode appearance cues around each joint. These features are later used as semantic guidance during cross-modal fusion.

2) Pseudo-Depth Branch: After scale alignment through DTOR and ADRM, the refined pseudo-depth map is treated as a single-channel input and fed into the depth backbone network to extract depth features and pixel-wise joint predictions:

$$\{\mathbf{F}^{d}, \mathbf{O}^{d}\} = \text{Backbone}_{d}(\mathbf{I}^{d}) \tag{6}$$

where $\mathbf{F}^d$ denotes the depth feature maps and $\mathbf{O}^d$ encodes joint response intensity, offsets, and confidence weights. Specifically, at each pixel location, the depth backbone predicts: (1) a 2D spatial offset from the pixel to the target joint; (2) a depth-direction geometric offset; (3) a joint response score indicating the spatial relevance of the pixel for joint voting; (4) a confidence weight representing the reliability of the pixel's contribution. Although both the response score and confidence weight act as weighting factors, they capture different aspects: the response emphasizes spatial relevance, while the confidence weight reflects geometric reliability. Unlike the RGB branch, the depth branch directly produces an initial 3D joint estimation, denoted as $\hat{\mathbf{J}}^{uvd}$. Specifically, joint coordinates are obtained by weighted aggregation over all pixel predictions:

$$\hat{\mathbf{J}}^{uvd} = \sum_{\mathbf{p}\in\Omega} w_{\mathbf{p}} \cdot (\mathbf{p} + \Delta\mathbf{p}) \tag{7}$$

where each pixel location $\mathbf{p} = (u,v,d)^\top \in \Omega$ consists of image-plane coordinates (u,v) and the associated depth value d. The network predicts a 3D offset vector $\Delta\mathbf{p} = (\Delta u, \Delta v, \Delta d)^\top$ pointing toward the target joint, along with a corresponding weight $w_{\mathbf{p}}$. The final joint position $\hat{\mathbf{J}}^{uvd}$ is obtained via weighted averaging over all pixel-level predictions. The depth branch is directly supervised using ground-truth joint coordinates, resulting in a coordinate regression loss:

$$\mathcal{L}_{coord} = \|\hat{\mathbf{J}}^{uvd} - \hat{\mathbf{J}}^{gt}\|_1. \tag{8}$$

In the following section, we introduce how cross-modal information from the RGB and pseudo-depth branches is effectively fused to further refine 3D hand pose estimation.

3) Cross-modal Fusion: Due to the inherent limitations of representing depth data as 2D images, significant 3D geometric structure information is often lost during feature extraction. Based on [25], we introduce a 2D–3D projection module that maps the initial joint predictions produced by the depth branch into a shared point cloud space. Specifically, this module takes predicted joints as geometric anchors and establishes explicit indexing and neighborhood relationships among image pixels, point cloud coordinates, and joints. Through this process, 2D image features are explicitly associated with 3D point cloud geometry, providing a unified geometric intermediate space that facilitates subsequent cross-modal feature interaction.

Building upon this representation, we design a two-step keypoint-aware fusion stage inspired by [2]. Each step alternates between intra-modal interaction and inter-modal fusion, enabling progressive refinement of joint representations. The intra-modal Transformer encoder performs self-

attention on joint features derived from point cloud aggregation, where each joint serves as both query and key to capture long-range kinematic dependencies. The learned attention weights establish anatomical correspondences between connected joints (e.g., MCP-PIP-DIP finger chains) and enforce hand kinematic consistency. The inter-modal Transformer decoder performs cross-attention where RGB appearance features (queries) attend to depth geometric features (keys and values), learning joint-level correspondences that enhance appearance–geometry complementarity. This multi-level attention design reduces ambiguities by establishing explicit correspondences between RGB appearance patterns and depth geometric structures.

Specifically, after constructing local point cloud neighborhoods around the initial joint estimates predicted by the backbone, the network performs the first refinement step through joint-wise intra-modal interaction and cross-modal fusion. By jointly leveraging geometric and appearance cues, this step produces intermediate joint predictions. In the second step, the intermediate joint predictions obtained in Step 1 are treated as updated geometric anchors. Using the same fusion process as in the first step, joint-centered point cloud neighborhoods are reconstructed, enabling the network to further enhance geometric consistency in 3D space and refine the joint estimates. Finally, a fully connected layer aggregates the joint-level features and performs linear mapping to output the final 3D joint predictions.

For each refinement step, we supervise the predicted spatial attention maps using joint heatmap-based signals. An L1 loss is applied to enforce consistency between the learned attention distribution and the ground-truth joint response maps, encouraging the network to focus on joint-relevant spatial regions during feature fusion. Unlike the joint heatmap $\mathbf{H}$, which represents spatial likelihood of joint locations and serves as a supervisory signal, the attention map $\mathbf{A}$ is a learnable weighting mechanism that dynamically modulates RGB features during the fusion process. This attention is supervised using a heatmap-based loss:

$$\mathcal{L}_{spatial} = \|\mathbf{A} - \mathbf{H}^{gt}\|_1, \tag{9}$$

which encourages consistent spatial focus during feature fusion. Finally, the entire network is trained end-to-end with a weighted combination of losses:

$$\mathcal{L} = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{coord}\mathcal{L}_{coord} + \lambda_{spatial}\mathcal{L}_{spatial}, \tag{10}$$

where λ_{pix}, λ_{coord}, and $\lambda_{spatial}$ are hyperparameters that balance the contributions of different loss terms.

We will present our experimental setup and results analysis in the following sections.

IV. Experiments

A. Experimental Setup

DexYCB Dataset DexYCB dataset [26] is designed for capturing hand–object interaction scenarios. It contains 500000 samples distributed across 1,000 sequences, involving 10 subjects interacting with 20 different objects. DexYCB provides four official dataset split protocols, denoted as S0, S1, S2, and S3. In this work, performance comparisons are conducted under all four split settings, while the default S0 split is adopted for ablation studies.

DexYCB Single In practical space station scenarios, camera viewpoints are fixed after installation. To evaluate the applicability of the proposed method under a fixed field-of-view condition, we further construct a single-view subset from DexYCB.

Implementation Details Experiments are conducted on two NVIDIA RTX 4090 GPUs. The model is trained using the Adam optimizer with an initial learning rate of 1.6×10^{-3} and a batch size of 128. By default, the depth estimation model with 335.3M parameters is employed to generate the initial pseudo-depth maps D_{init}. Training is performed on the DexYCB dataset for 20 epochs, with each training run taking approximately 16 hours.

Evaluation Metrics Following the standard evaluation protocol in hand pose estimation [2], we use ground-truth hand centers for cropping and evaluate keypoint prediction accuracy using Mean Per Joint Position Error (MPJPE) and its Procrustes-aligned variant (PA-MPJPE). MPJPE measures the average Euclidean distance between the predicted 3D joint locations and the ground-truth annotations, and PA-MPJPE eliminates the influence of global pose and scale variations, focusing on evaluating the consistency of the relative hand pose structure.

B. Experimental Results

1) Qualitative Results: Fig. 4 presents qualitative results of our method across a variety of real-world scenarios. As shown, the proposed approach is able to stably predict structurally consistent, coherent, and accurate hand keypoints under diverse hand poses, occlusion patterns, and hand–object interaction conditions. Even in the presence of rapid motion, severe self-occlusion, or complex interactions with objects, our model preserves the integrity of hand joint topology and produces predictions that are well aligned with visual evidence in the input images, demonstrating strong robustness and generalization capability.

2) Quantitative Results: Table I reports quantitative comparisons on the DexYCB dataset. Among current state-of-the-art RGB-based methods, our approach achieves the best performance on both MPJPE and PA-MPJPE metrics, significantly outperforming existing monocular RGB baselines. Notably, despite relying solely on RGB input without access to real depth measurements, our method attains performance comparable to several depth-based or RGB-D approaches, highlighting its strong ability to recover 3D hand structures and its robustness under limited sensing conditions.

Table II reports the performance of the model under batch inference. The results indicate that, during batch inference, the overall speed bottleneck is mainly constrained by the generation of pseudo-depth information. SHPN-None denotes

TABLE I
COMPARISON OF HAND POSE ESTIMATION METHODS ON DEXYCB DATASET.

Method	MPJPE (mm)					PA-MPJPE	Input Modality
	S0	S1	S2	S3	Avg		
Spurr et al. [27]	17.34	22.26	25.49	18.44	20.88	6.83	RGB
Liu et al. [28]	15.28	–	–	–	–	6.58	RGB
Tse et al. [29]	16.05	21.22	27.01	17.93	20.55	–	RGB
METRO [22]	15.24	–	–	–	–	6.99	RGB
HandOccNet [20]	14.04	–	–	–	–	5.80	RGB
A2J [30]	23.93	25.57	27.65	24.92	25.52	–	Depth
AWR [31]	11.23	–	–	–	–	–	Depth
IPNet [25]	8.03	9.01	8.60	7.80	8.36	–	Depth-PC
DiffHand [32]	12.10	–	–	–	–	4.98	RGB-D
Keypoint-Fusion [2]	6.94	8.64	7.56	7.02	7.54	4.79	RGB-D
Ours	**8.61**	**10.81**	**10.13**	**9.65**	**9.80**	**5.42**	RGB

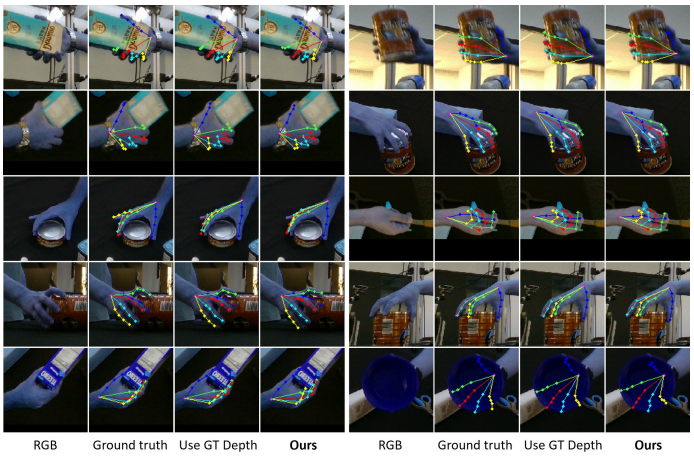

Fig. 4. Qualitative results on the DexYCB dataset.

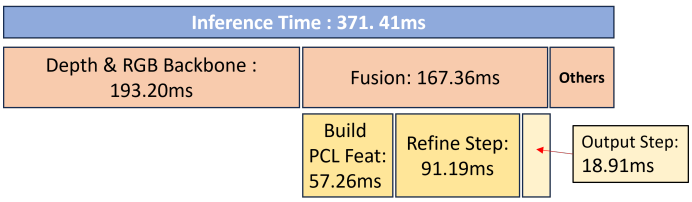

Fig. 5. Time taken by each module when inferring a single image.

the inference speed when precomputed pseudo-depth maps are directly used without online pseudo-depth generation.

In Fig. 5, we further present a detailed breakdown of the inference time for a single image after the pseudo-depth information has been obtained. The results show that the inference time of the backbone networks is comparable to that of the fusion stage. Within the fusion stage, the majority of computation time is consumed by the first refinement step, where the initial cross-modal alignment among RGB features, pseudo-depth features, and point cloud representations is established. The second step repeats the same fusion procedure using the refined joint estimates as updated anchors, focusing on local geometric refinement and therefore introducing significantly less computational overhead.

TABLE II
BATCH INFERENCE RESULTS.

Method	FPS	PA-MPJPE(mm)
SHPN-None	80.45	-
SHPN-Vits	28.71	6.06
SHPN-Vitl	15.17	5.42

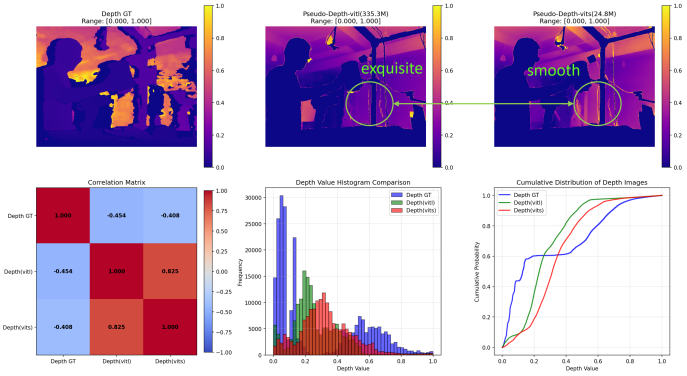

Fig. 6. Comparison of the quality of initial pseudo-depth images with different parameter models and the ground truth of depth values.

3) Ablation Study: Fig. 6 provides a detailed comparison between pseudo-depth information and depth maps captured by depth sensors with different statistical laws. Further experiments analyze the impact of pseudo-depth quality on overall model performance, evaluate computational efficiency, and discuss the limitations of the proposed method.

Statistical Analysis between Depth Images We compare the statistical differences between the ground-truth depth and pseudo-depth maps of varying quality in Fig. 6. By comparing the second and third images, it can be observed that the initial pseudo-depth estimated by a large-scale pretrained model exhibits exquisiter and more detailed textures, whereas the pseudo-depth generated by a lightweight model appears significantly smoother. However, as revealed by the correlation matrices, there remains a substantial discrepancy between the pseudo-depth and the ground-truth depth. For example, as shown in the fifth and sixth images, the pseudo-depth maps are inconsistent with the ground-truth depth in terms of physical

TABLE III
RESULTS OF THE SINGLE-VIEW SUBSET EXPERIMENT.

Subset	MPJPE (mm)	PA-MPJPE (mm)
View-1	9.38	5.65
View-2	9.40	5.98
View-3	14.7	7.35
View-4	8.10	5.19
Full View	**8.61**	**5.42**

distribution and scale. In particular, the value distribution of pseudo-depth is overly uniform, which is a key factor that makes it difficult to align pseudo-depth with the original depth information.

Single-view Subset Experimental Results We selected 4 representative viewpoints (view 1-4 corresponding to 836212060125, 839512060362, 932122061900, and 932122060857) from the complete eight-viewpoint dataset as the sub-dataset, which was used to simulate the actual engineering camera scene in the cargo spacecraft. As shown in Table III, we observe comparable performance when training the proposed method on either the full multi-view dataset or a single-view subset. From a system engineering perspective, this result demonstrates the feasibility of deploying the proposed method under a fixed field-of-view camera configuration, which is consistent with practical space station scenarios.

Impact of Depth Map Quality on Model Performance Table IV presents a detailed comparison of model performance under different backbone configurations (ResNet-18 vs. ResNet-50), as well as the impact of varying pseudo-depth quality in the proposed pseudo-depth injection mechanism. Specifically, we evaluate how different initial pseudo-depth estimates D_{init} affect inference accuracy. Taking the pseudo-depth generated by the ViT-L model as the reference, we introduce Gaussian random noise into D_{init} in the ViT-Mask ablation study to simulate invalid depth maps. As illustrated in Fig. 6, pseudo-depth maps generated by the lightweight ViT-S model are used to represent low-quality depth estimates.

TABLE IV
ABLATION STUDY OF DIFFERENT BACKBONE AND PSEUDO-DEPTH QUALITY SETTINGS

Setting	MPJPE (mm)	PA-MPJPE (mm)
ResNet-18	8.61	5.42
ResNet-50	8.63	5.41
ViT-Mask	14.77	7.93
ViT-S (24.8M)	9.92	6.06
ViT-L (335.3M)	8.61	5.42

As shown in Table IV, the quality of pseudo-depth information has a more pronounced impact on 3D hand pose estimation performance than the choice of backbone architecture. When the quality of pseudo-depth severely degrades, the fusion module tends to rely more on RGB appearance cues. Compared with using RGB information alone, incorporating depth priors learned from large-scale external depth datasets [24] indeed helps improve the model's reasoning and detection performance. Low-quality depth maps, which typically suffer from limited resolution, weak structural details, and poor global consistency compared to higher-capacity depth models, introduce noise, discontinuities, and local geometric errors. During the fusion stage, these artifacts may be mistakenly treated as valid geometric cues, thereby disrupting feature alignment and degrading geometric reasoning. In contrast, higher-quality pseudo-depth representations provide more complete and coherent local geometric structures, enabling the network to better infer the true spatial relationships among hand joints in 3D space. When the depth maps are coarse or lack fine-grained details, the model's ability to perceive local geometric structure is significantly weakened.

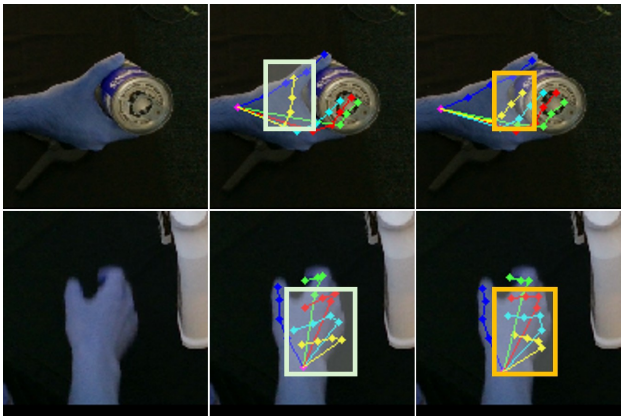

Fig. 7. A case of error detection.

Overall, stable and fine-grained pseudo-depth information plays a critical role in enhancing the model's 3D geometric reasoning capability and is one of the key factors driving performance improvements.

Case analysis of model limitations This subsection discusses a representative failure case of our model. We also discuss an inherent limitation of the proposed method. Although pseudo-depth modality injection enhances geometric perception for 3D hand pose estimation, certain cases remain fundamentally ill-posed. As illustrated in Figure 7, consider a can-grasping action as an example: individual grasping habits vary across subjects (e.g., a raised little finger as shown in the figure). Under viewpoints with complete self-occlusion, the 2D projections of some fingers are entirely hidden, making it impossible to recover the underlying hand configuration from visual evidence alone. In such cases, even with pseudo-depth enhancement, the model cannot accurately infer the hand pose behind the occluded regions.

V. CONCLUSION

To address the problem of hand pose perception in sensor-constrained scenarios for crewed space missions, we propose a 3D hand pose estimation method from an input RGB image. This method enhances geometric features in the data through pseudo-depth modality injection and introduces the SpaceRGB-HandPose Net for multimodal information fusion.

Experimental results show that our approach outperforms pure RGB-based solutions and approaches the performance of RGB-D methods that rely on hardware-captured depth information fused with RGB data. The annotation of 3D hand models remains a challenging and costly task. Future work will therefore focus on collecting data from real-world engineering scenarios for training, as well as exploring algorithmic and technological innovations to alleviate this fundamental limitation, thereby facilitating more practical and efficient 3D hand interaction in human spaceflight engineering.

REFERENCES

[1] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018, pp. 5079–5088.

[2] X. Liu, P. Ren, Y. Gao, J. Wang, H. Sun, Q. Qi, Z. Zhuang, and J. Liao, "Keypoint fusion for rgb-d based 3d hand pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3756–3764.

[3] J. Ma, Y. Zhou, Z. Wang, H. Sang, R. Jiang, and B. He, "Geometric-aware rgb-d representation learning for hand–object reconstruction," *Expert Systems with Applications*, vol. 257, p. 124995, 2024.

[4] D.-C. Hoang, A.-N. Nguyen, T.-U. Nguyen, N.-A. Hoang, V.-D. Vu, D.-Q. Vu, P.-Q. Ngo, K.-T. Phan, D.-T. Tran, V.-T. Nguyen *et al.*, "Attention-based hand pose estimation with voting and dual modalities," *Engineering Applications of Artificial Intelligence*, vol. 139, p. 109526, 2025.

[5] M. Rezaei, R. Rastgoo, and V. Athitsos, "Trihorn-net: A model for accurate depth-based 3d hand pose estimation," Jun 2022.

[6] W. Cheng, J. H. Park, and J. H. Ko, "Handfoldingnet: A 3d hand pose estimation network using multiscale-feature guided folding of a 2d hand skeleton," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 260–11 269.

[7] W. Cheng and J. H. Ko, "Handr2n2: Iterative 3d hand pose estimation using a residual recurrent neural network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 20 904–20 913.

[8] E. Q. Wu, Z.-R. Tang, P. Xiong, C.-F. Wei, A. Song, and L.-M. Zhu, "Ropenpose: a rapider openpose model for astronaut operation attitude detection," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 1, pp. 1043–1052, 2021.

[9] Z. Liu, Y. Li, C. Wang, L. Liu, B. Guan, Y. Shang, and Q. Yu, "Astropose: Astronaut pose estimation using a monocular camera during extravehicular activities," *Science China Technological Sciences*, vol. 67, no. 6, pp. 1933–1945, 2024.

[10] J. Ouyang, W. Xie, G. Xu, C. Li, S. Gan, X. Zhang, X. Yang, C. Jiang *et al.*, "Spacesuitpose: Deep learning-based spacesuit pose estimation in extravehicular activities from monocular images," *Acta Astronautica*, 2025.

[11] R. Xun, Z. Zhang, and L. Chang, "H3-astronaut pose net: A multiview 3d astronaut pose estimation system for spacecraft," *Neurocomputing*, vol. 650, p. 130877, 2025.

[12] Q. Gao, J. Li, Y. Zhu, S. Wang, J. Liufu, and J. Liu, "Hand gesture teleoperation for dexterous manipulators in space station by using monocular hand motion capture," *Acta Astronautica*, vol. 204, pp. 630–639, 2023.

[13] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2640–2649.

[14] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[15] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian conference on computer vision*. Springer, 2014, pp. 332–347.

[16] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 89–98.

[17] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz, "Hand pose estimation via latent 2.5 d heatmap regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 118–134.

[18] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[19] N. Qian, J. Wang, F. Mueller, F. Bernard, V. Golyanik, and C. Theobalt, "Html: A parametric hand texture model for 3d hand reconstruction and personalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 54–71.

[20] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1496–1505.

[21] X. Chen, Y. Liu, C. Ma, J. Chang, H. Wang, T. Chen, X. Guo, P. Wan, and W. Zheng, "Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 274–13 283.

[22] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1954–1963.

[23] J. Jiang, J. Li, B. Zhang, X. Deng, and B. Shi, "Evhandpose: Event-based 3d hand pose estimation with sparse supervision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6416–6430, 2024.

[24] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[25] P. Ren, Y. Chen, J. Hao, H. Sun, Q. Qi, J. Wang, and J. Liao, "Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2163–2171.

[26] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9044–9053.

[27] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, "Weakly supervised 3d hand pose estimation via biomechanical constraints," in *European conference on computer vision*. Springer, 2020, pp. 211–228.

[28] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang, "Semi-supervised 3d hand-object poses estimation with interactions in time," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 687–14 697.

[29] T. H. E. Tse, K. I. Kim, A. Leonardis, and H. J. Chang, "Collaborative learning for hand and object reconstruction with attention-guided graph convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1664–1674.

[30] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 793–802.

[31] W. Huang, P. Ren, J. Wang, Q. Qi, and H. Sun, "Awr: Adaptive weighting regression for 3d hand pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 061–11 068.

[32] L. Li, L. Zhuo, B. Zhang, L. Bo, and C. Chen, "Diffhand: end-to-end hand mesh reconstruction via diffusion models," *arXiv preprint arXiv:2305.13705*, 2023.