IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Optimized Design and Calibration of a Human-Eye-Sized Active Binocular Vision System Based on Spherical Parallel Mechanism

Kaifang Wang, DongDong Yang, Li Zhang, Jun Liu, and Xiaolin Zhang

***Abstract*—The Active Binocular Vision System (ABVS), resembling the human eye, demonstrates potential for improving visual perception in robotic systems, especially in dynamic and complex environments. In this letter, we present an optimized design of a three degree-of-freedom (DoF) Active Monocular Vision System (AMVS) based on a Spherical Parallel Manipulator (SPM). By combining two identical AMVS units, we form an ABVS, which has been successfully integrated into a humanoid robotic head. Due to the highly nonlinear kinematics of SPM and complex error coupling in its multi-link structure, traditional end-to-end neural network training methods are insufficient in accuracy and require large datasets. To address these challenges, we propose a two-branch optimization network that significantly improves calibration accuracy. Furthermore, we introduce a four-branch fine-tuning strategy that enables accurate kinematic models to be obtained with only a small amount of data from new AMVS devices. Experimental results demonstrate that the two-branch optimization network reduces rotational prediction error by 16% and translational error by 5% compared to a single-branch network. Furthermore, the four-branch fine-tuning network achieves comparable accuracy to a fully trained single-branch network using only 343 data points. Finally, our ABVS shows the capability to perform 3D visual tasks, such as stereo reconstruction during movement.**

***Index Terms*—Humanoid robot systems, parallel robots, biologically-inspired robots.**

Received 8 November 2024; accepted 25 March 2025. Date of publication 28 April 2025; date of current version 23 May 2025. This article was recommended for publication by Associate Editor H. Liu and Editor X. Liu upon evaluation of the reviewers' comments. This work was supported by the Shanghai Municipal Science and Technology Major Project (ZHANGJIANG LAB) under Grant 2018SHZDZX01. *(Corresponding author: Xiaolin Zhang.)*

Kaifang Wang is with the Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China.

DongDong Yang and Li Zhang are with Anhui Eyevolution Technology Company Ltd., Shanghai 200050, China.

Jun Liu is with the Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China.

Xiaolin Zhang is with the Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, also with ShanghaiTech University, Shanghai 201210, China, and also with Anhui Eyevolution Technology Company Ltd., Shanghai 200050, China (e-mail: zhang.xiaolin@ilooktech.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3564757, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3564757

I. Introduction

In RECENT years, robotics has made significant strides, with humanoid robots that can adapt to complex, unknown, and dynamic environments garnering considerable attention. Among the key factors influencing the performance of these advanced robotic systems, environmental perception is critical [1]. For humans, approximately 80% of information is acquired through vision. Similarly, robots require ABVS, akin to human eyes, to dynamically adjust their gaze direction, thereby improving environmental perception. ABVS not only balances high-resolution and wide-field views but also integrates with actuators and inertial sensors to stabilize the line of sight in dynamic conditions, significantly enhancing the performance of vision algorithms [2], [3].

To effectively integrate ABVS into humanoid robots, addressing the system's size is crucial. Given that current humanoid robots are generally close to human dimensions, the ABVS should ideally be comparable in size to human eyes for seamless integration. The adult human eye has a diameter of approximately 24 mm, with an interpupillary distance ranging between 54 mm and 72 mm [4]. Furthermore, ABVS must possess stereoscopic vision capabilities similar to human eyes, which not only requires accurate kinematic calibration algorithms for different eye positions but also imposes certain requirements on overall system accuracy.

Existing research on ABVS can be classified into two categories based on the structure of the kinematic chain: serial and parallel. Serial-chain ABVS offers simpler kinematics and more straightforward error modeling, with established calibration systems in place [5]. However, due to the sequential structure of serial actuators, achieving a design close to the size of human eyes while maintaining accuracy remains a challenge. In contrast, parallel-chain ABVS, which employs multiple links to drive camera rotations, allows the actuators and encoders to be concentrated at the base, making miniaturization of the "eyeballs" feasible. However, the complexity of parallel structures, where any geometric error in a link is transmitted through coupling, complicates calibration and often necessitates the use of additional sensors [6] or external devices [7], increasing costs.

Inspired by the Agile Eye [8] and Orbita 3D [9], this letter presents an optimized design for a humanoid eye-sized AMVS based on SPM with 3-DoF (Pitch, Roll, Yaw), similar to the human eye. Through this optimized design, the diameter of the AMVS eye and the interpupillary distance are reduced to 30 mm and 65 mm, respectively, while maintaining high repeatability in motion accuracy. This system was successfully integrated into a humanoid head-sized robot (see Fig. 1(b)).

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

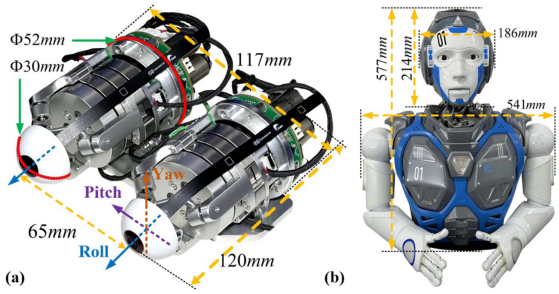

Fig. 1. **SPM-based 6-DoF ABVS (a) and a Torso-Humanoid Robot Equipped with the ABVS (b).** In (a), two red curves denote the maximum outer diameter circle and the eyeball outer diameter circle of the monocular structure. Each eye houses a fixed-focus color camera (1600 × 1200 resolution, 25 fps, 118.2° horizontal FOV), with the blue arrow indicating the optical axis. The ABVS can rotate within **±30° pitch**, **±30° yaw**, and **±15° roll**. In (b), the humanoid robot features the ABVS, a 3-DoF SPM-based neck, a 1-DoF mouth joint, stereo speakers (ears), a six-microphone array (head), and two 5-DoF robotic arms.

Given the nonlinear kinematics and complex error propagation in SPM, traditional kinematic modeling methods may face challenges in both computational complexity and calibration accuracy. Recently, neural network-based methods have made significant advances in robotic kinematic modeling, effectively learning the mapping between inputs and outputs from large datasets. This study proposes a novel two-branch optimization network that further improves the accuracy of forward kinematic calibration for the AMVS. Additionally, to address discrepancies in kinematic models across different AMVS units caused by factors such as machining, assembly, or wear, we propose a four-branch fine-tuning network, enabling calibration with minimal data for each new device.

The main contributions of this letter are as follows:

- Designed a 6-DoF ABVS based on SPM, achieving high repeatability in positioning while matching the human eye size.
- Proposed a two-branch optimization neural network that significantly enhances the calibration accuracy of the AMVS. Furthermore, the partitioned fine-tuning strategy reduces the data required for calibration while achieving satisfactory precision.
- Demonstrated accurate stereo reconstruction on the ABVS. To the best of our knowledge, this is the first instance of achieving accurate stereo reconstruction on a humanoid eye-sized ABVS.

II. Related Works

ABVS can be classified into series and parallel architectures based on differences in kinematic chain structures. Series ABVS offer different DoF, such as 3-DoF [10], 4-DoF [11], and 6-DoF [5], [12]. The 6-DoF version features precise kinematic calibration, enabling accurate depth maps and point clouds. In contrast, parallel ABVS, such as the Agile Eye introduced in [8], based on an SPM, achieve higher rotational speeds and ranges compared to the human eye. Furthermore, [13] presents a parallel-type bionic eye utilizing the Oculomotor Control Model and SPM principles. However, the aforementioned ABVS systems are not well-suited for integration into humanoid-sized robot heads due to their size and structural limitations.

Calibration methods for parallel mechanisms, such as SPM, can be categorized into geometric error modeling and neural network-based fitting approaches. Geometric error modeling methods [14], [15], [16] often involve complex modeling and solving processes, and while hand-eye calibration [16], [17] can correct offsets between the mobile platform and the camera, it cannot address inherent errors within the mechanism itself. In contrast, neural network-based methods [18], whether for kinematic modeling or error compensation [19], are more suitable for highly nonlinear systems like SPM but require substantial calibration data for each device. Additionally, some methods utilize external devices [20], [7] or additional sensors [6] for kinematic estimation, which increase deployment and operational costs.

III. Proposed ABVS Based on SPM

The proposed ABVS achieves 6-DoF rotational motion with dimensions comparable to human eyes, enabling integration into a humanoid head. This compact design demands precise kinematic calibration, high-precision encoders, and excellent positioning repeatability. Utilizing a coaxial SPM—a 3-DOF pure rotational parallel mechanism with a fixed rotation center—the system offers advantages over serial structures [5], [12] in terms of low power consumption, compact size, lightweight design and high dynamic performance. Moreover, the Orbita 3D system-based design simplifies hinge structures relative to other SPMs [8], [13], yielding a more compact, cylindrical form that supports even smaller dimensions and baselines. The optimized AMVS was successfully integrated into a humanoid robot head (see Fig. 1(b)) as one of two identical units forming the ABVS.

A. Optimized Design of the AMVS

The AMVS consists of a mobile platform, a camera, upper and lower bases, a motor drive board, and three assemblies of "motor, gearbox, link gear, proximal link, distal link" (see Fig. 2(a)). The three distal links are connected to the mobile platform, which houses MIPI camera modules measuring just 7.8 mm × 7.8 mm × 5.12 mm, equipped with OV8856 sensors that support frame synchronization for dual-camera image capture. The selected motor is the MOONS' DCU13020 brushed coreless motor, with a diameter of only 13 mm, a maximum speed of 784 rpm, and a continuous maximum torque of 22.4 mNm. Speed reduction is achieved through reduction gearboxes and link gears with ratios of 16:1 and 62:21, respectively. To enhance repeatability in positioning and ensure accurate kinematic calibration of the AMVS, optimizations were made in both the mechanical structure and the position encoding system.

First, in terms of mechanical structure, the key components of the AMVS were manufactured using aluminum alloy, with shafts and holes machined to high precision standards of Grade 5 and Grade 6, respectively. Detailed tolerance analysis was conducted for the bearings and other parts in the hinge mechanism, with careful distribution of tolerances across components. This ensures the rotational accuracy of the hinges, thereby minimizing mechanical errors and backlash.

Second, in terms of position encoding, the AMVS utilizes multi-stage gear transmission involving reduction gearboxes and link gears, which introduces gear backlash. If the encoder is installed on the motor end [23] or at the link gear [9], the backlash would prevent the encoder from accurately reflecting the movement of the links, thus affecting the repeatability of positioning. Furthermore, incremental encoders require finding a reference point upon startup to determine absolute position. If the encoder is mounted on a shaft with a reduction ratio greater

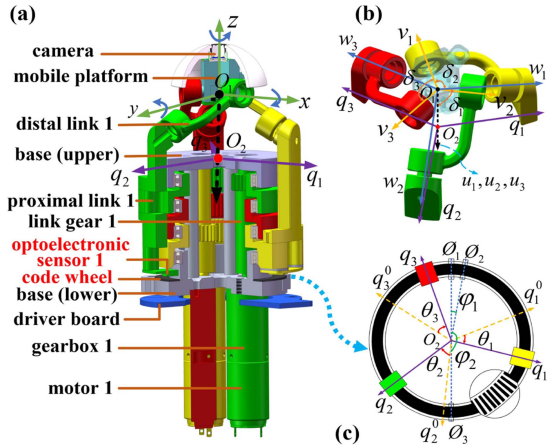

Fig. 2. **Partial Cutaway View (a), Kinematic Structure (b), and Encoder System (c) of the AMVS.** In (a), the yellow, green, and red represent the three gear transmission and linkage systems, with parts ending in "1" belonging to the green system. Three optoelectronic sensors and a black-marked shared code wheel form the position acquisition system, with O_1xyz representing the end-effector coordinate system. In (b), yellow dashed lines show the **home configuration** of the three proximal links forming 120° angles, rotating around O_2, with their axes $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3$ and deviations θ_1, θ_2 and θ_3. Unit vectors $\boldsymbol{u}_1, \boldsymbol{u}_2$, and $\boldsymbol{u}_3$ denote the directions of the proximal links' lower rotation axes, while $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, $\boldsymbol{w}_3$ and $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ denote those of the distal links' lower and upper rotation axes, respectively. All intersect at point O_1. Angles δ_1 and δ_2 define the link curvature. In (c), yellow, green, and red blocks show the optoelectronic sensors. The red arrows indicate positive motion. And three zero markers on the code wheel are labeled $\emptyset_1$, $\emptyset_2$, $\emptyset_3$, with angles φ_1 and φ_2 representing the angles between $\emptyset_1$ and $\emptyset_2$, and between $\emptyset_1$ and $\emptyset_3$. For this AMVS, $\delta_1 = 60°$, $\delta_2 = 90°$, $\delta_3 = 120°$, $\varphi_1 = 9.944°$, $\varphi_2 = 180°$ and $\boldsymbol{u}_1 = \boldsymbol{u}_2 = \boldsymbol{u}_3 = [0, 0, -1]$.

than 1, it cannot directly capture the pose of the mobile platform. Operating the motor without precise platform pose information may lead to singular configurations, where the angle between the z-axis of the AMVS mobile platform and the z-axis in the home configuration is 90°, potentially causing mechanical lockup or damage.

To address these issues, a new position acquisition system was designed, where three optoelectronic sensors and a shared code wheel were installed on the three proximal links and the lower base, respectively (see Fig. 2). This allows the encoder to directly measure the motion of the proximal links, eliminating the effects of gear backlash. Incremental optoelectronic encoders (Broadcom AEDR-871x) were used in combination with a custom 1448-line radial grating code wheel. These encoders feature dual-channel quadrature digital output and built-in 16x interpolation, providing a theoretical minimum resolution of 0.0039°. Since the three sensors share a single code wheel, the system can control all three motors to rotate in the same direction and at the same speed after startup. Once the sensors pass the zero position on the code wheel, the absolute positions of the proximal links can be determined. Three zero positions were set on the code wheel (labeled $\emptyset_1$, $\emptyset_2$, $\emptyset_3$ in Fig. 2(c)), allowing each sensor to determine its absolute position with a rotation of at most 180°. This design improves the speed and convenience of determining the pose of the mobile platform and eliminates the risk of entering singular positions.

B. Kinematics of the AMVS

In this letter, the kinematic model of the AMVS is divided into the ideal and calibrated models. The ideal forward and inverse kinematic models, B_{ideal} and B^{-1}_{ideal}, can be derived directly from the AMVS design documentation, and their derivation and solutions have been thoroughly discussed in previous studies [22], [23]. The calibrated kinematic model B_{calib}, is obtained by collecting data and applying various calibration methods. For simplicity, B_{ideal} and B^{-1}_{ideal} denote the ideal models, while B_{calib} and B^{-1}_{calib} represent the calibrated models, defined as follows:

$$\mathbf{P}_{\text{ideal}} = B_{\text{ideal}}(\Theta), \quad \Theta_{\text{ideal}} = B^{-1}_{\text{ideal}}(\mathbf{P}_{\text{ideal}}) \tag{1}$$

$$\mathbf{P}_{\text{calib}} = B_{\text{calib}}(\Theta), \quad \Theta_{\text{calib}} = B^{-1}_{\text{calib}}(\mathbf{P}_{\text{calib}}) \tag{2}$$

Here, $\Theta = [\theta_1, \theta_2, \theta_3]$ represents the position vector of the proximal links in the AMVS, and $\mathbf{P} = [\alpha, \beta, \gamma, x, y, z]$ represents the end-effector pose vector of the camera, including Euler angles $[\alpha, \beta, \gamma]$ in the pitch, roll, and yaw directions, and the translation vector $[x, y, z]$ relative to the home configuration. $\mathbf{P}_{\text{ideal}}$ and $\mathbf{P}_{\text{calib}}$ represent the ideal and actual end-effector poses of the AMVS camera, respectively, while Θ_{ideal} and Θ_{calib} represent the ideal and actual proximal link positions required.

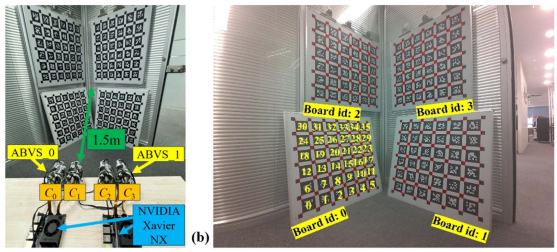

Fig. 3. **Data Collection Setup (a) and AprilGrid [21] Image Captured by AMVS (b).** In (a), the distance between the four AprilGrid boards and the two ABVSs (comprising four AMVSs: $\boldsymbol{C}_0$, $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\boldsymbol{C}_3$) is approximately 1.5m, ensuring sufficient detection of corner points throughout the camera's motion within the entire workspace. In (b), the red crosses indicate the detected corners of the boards, while the yellow labels show the board identifiers and the tag IDs. The tag ID ranges for board IDs 1, 2, and 3 are 36 to 71, 72 to 107, and 108 to 143, respectively.

IV. Calibration Data Collection and Analysis

A. Data Collection and Processing

As shown in Fig. 3(a), the calibration data was collected using two ABVS systems (a total of four AMVS units: $\boldsymbol{C}_0$, $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$). As previously described, the theoretical end-effector workspace of the AMVS is defined by the pitch (α), roll (β), and yaw (γ) angles. These angles are sampled at 1° intervals: $\alpha \in [-30°, 30°]$, $\beta \in [-15°, 15°]$, and $\gamma \in [-30°, 30°]$. The Cartesian product of these sampled values forms the complete set of theoretical poses:

$$\mathbf{P}_{\text{ideal},k} \in \{\alpha\} \times \{\beta\} \times \{\gamma\}, \quad k = 1, 2, \ldots, 115351 \tag{3}$$

This results in 115351 unique combinations that comprehensively represent the theoretical workspace. Due to variations in the reachable limits of different AMVS units, the total number of data points collected varies slightly. A total of $N_0 = 109625$, $N_1 = 110462$, $N_2 = 110624$, and $N_3 = 110729$ samples were collected from the devices $\boldsymbol{C}_0$, $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$, respectively, forming DatasetC0, DatasetC1, DatasetC2, and DatasetC3. During data collection, for the j-th ideal sampling pose of the AMVS, $\mathbf{P}_j = [\alpha_j, \beta_j, \gamma_j, 0, 0, 0]$, the corresponding proximal link positions $\Theta_j = [\theta_{1j}, \theta_{2j}, \theta_{3j}]$ were calculated using the ideal inverse kinematic model B^{-1}_{ideal}, and the AMVS motors were driven to these positions to perform sampling. Subsequently, the AprilGrid images captured by the camera were processed

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

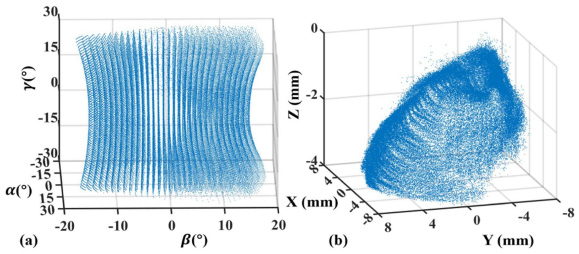

Fig. 4. **Data Distribution of Dataset_C0**: (a) Distribution of three-axis rotation data, and (b) Distribution of three-axis translation data, both obtained from camera pose estimation using AprilGrid.

using a combination of corner detection and bundle adjustment to optimize the estimation of the camera pose relative to the home configuration, yielding the actual end-effector pose $\mathbf{P}_{\text{actual},j} = [\alpha_j', \beta_j', \gamma_j', x_j', y_j', z_j']$, representing the actual rotation $[\alpha_j', \beta_j', \gamma_j']$ and translation $[x_j', y_j', z_j']$ of the camera relative to the home configuration.

To improve the accuracy of the camera pose estimation, four AprilGrid boards were used (see Fig. 3), ensuring that enough corner points could still be detected even after large-angle rotations of the AMVS. Each AprilGrid board has a unique identifier, and every tag on the boards also carries a distinct identifier. Using this information, the AMVS camera can reliably match the detected corner points in the current pose to those from the home configuration. With the matched corner points, the actual end-effector pose of the AMVS is precisely solved by minimizing the reprojection error.

B. Calibration Data Analysis

By processing the collected calibration images, the actual camera poses for each AMVS unit $\boldsymbol{C}_i$ ($i \in \{0,1,2,3\}$) were obtained as $\mathbf{P}_{\text{actual},j}$, where $j \in [1,N_i]$. An analysis of the actual pose data $\mathbf{P}_{\text{actual}}$ revealed significant differences in the data distribution between rotational and translational components. Specifically, the rotational data $[\alpha_j', \beta_j', \gamma_j']$ exhibited regularity and consistency, while the translational data $[x_j', y_j', z_j']$ showed greater randomness and dispersion. This is shown in Fig. 4, which displays the distribution of the camera poses for $\boldsymbol{C}_0$ across both rotational and translational dimensions.

From a theoretical perspective, rotation and translation jointly characterize the end pose, yet they exhibit significant differences in both physical meaning and numerical scale, as evidenced by the distribution of these data in Fig. 4. Specifically, even minor variations in rotation data can lead to pronounced global geometric transformations on the image plane, whereas translation data—affected by factors such as scene depth—display different sensitivity during optimization. This phenomenon aligns with common distribution optimization strategies in SLAM, such as sequentially solving for rotation and translation parameters in ORB-SLAM3 or optimizing them separately as described in [26]. Based on this observation, the next section introduces our branch network optimization strategy, which is designed to more precisely model the distinct characteristics of rotation and translation data.

V. Proposed Neural Network-Based Forward Kinematic Calibration of AMVS

Due to the highly nonlinear kinematic characteristics of SPM, its multi-link coupling structure causes errors to propagate and amplify progressively during the end-effector's movement. Traditional kinematic modeling methods face challenges in terms of both accuracy and computational complexity. To address these challenges, this letter proposes a two-branch **Multilayer Perceptron (MLP)** network to improve calibration accuracy based on the distinct characteristics of rotational and translational data. For practical application, the network is further refined into a four-branch structure for fine-tuning, improving calibration precision and substantially reducing training data requirements. For clarity, the three network structures (see Fig. 5) involved in this letter are labeled as follows: the baseline MLP model ($\boldsymbol{M}_1$) [18], the two-branch optimized network ($\boldsymbol{M}_2$), and the four-branch fine-tuning network ($\boldsymbol{M}_4$).

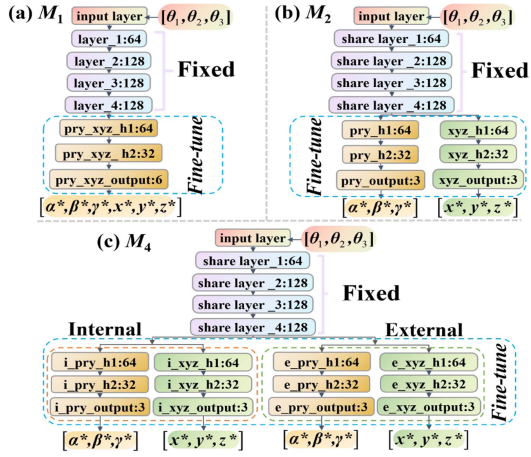

Fig. 5. **Neural network architectures for AMVS forward kinematic calibration**: (a) Baseline MLP model $\boldsymbol{M}_1$, (b) Proposed two-branch optimization network $\boldsymbol{M}_2$, (c) Proposed four-branch fine-tuning network $\boldsymbol{M}_4$. The blue dashed boxes highlight the layers subject to fine-tuning during training in each network.

A. MLP

The MLP is capable of learning complex nonlinear relationships and accurately fitting the mapping between inputs and outputs, making it suitable for handling the intricate error relationships in SPM. In this letter, the MLP directly maps the proximal link positions $\theta = [\theta_1, \theta_2, \theta_3]$ to the actual end-effector pose $\mathbf{P}_{\text{actual}}$. The network is trained by minimizing the error between the predicted pose and the measured pose. The loss function is defined as:

$$\mathcal{L}_{M1} = \sum_{i=1}^{N_{\text{train}}} \|\mathbf{P}_{\text{actual},i} - B_{\text{calib}}(\theta_i)\|^2 \Big/ N_{\text{train}} \tag{4}$$

where N_{train} is the number of training samples, $\|\cdot\|$denotes the Euclidean norm, $\mathbf{P}_{\text{actual},i}$ is the actual measured pose of the i-th sample, and $B_{\text{calib}}(\theta_i)$ is the model's predicted output.

B. Two-Branch Optimization Network

Through the analysis of actual camera pose data, it was observed that the distribution characteristics of rotational and translational data differ significantly. To accommodate this difference and enhance model robustness, this letter proposes treating the estimation of rotation and translation as two related but relatively independent tasks. The network adopts a backbone-branch structure (see Fig. 5(b)), where a shared backbone is responsible for feature extraction, while independent branches handle rotation and translation. The backbone extracts high-dimensional

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

features that benefit both tasks, while the branches handle the distinct spatial distribution and error characteristics of each.

Additionally, a weighted loss function is introduced to handle the inherent differences between rotational and translational data. Since rotational data tends to follow more regular patterns, the loss function assigns different weights to the rotational and translational terms. This approach ensures precise rotational accuracy while maintaining good translational accuracy. The total loss function is defined as:

$$\mathcal{L}_{M2} = w_{rot} \cdot \mathrm{MSE}_{rot} + w_{trans} \cdot \mathrm{MSE}_{trans} \tag{5}$$

where MSE_{rot} and MSE_{trans} represent the loss functions for rotation and translation, respectively, formulated as:

$$\mathrm{MSE}_{rot} = \frac{1}{N}\sum_{i=1}^{N}\left[(\alpha'_i - \alpha_i^*)^2 + (\beta'_i - \beta_i^*)^2 + (\gamma'_i - \gamma_i^*)^2\right] \tag{6}$$

$$\mathrm{MSE}_{trans} = \frac{1}{N}\sum_{i=1}^{N}\left[(x'_i - x_i^*)^2 + (y'_i - y_i^*)^2 + (z'_i - z_i^*)^2\right] \tag{7}$$

Here, w_{rot} and w_{trans} are the weights for the rotational and translational losses in the total loss function. $[\alpha_i^*, \beta_i^*, \gamma_i^*]$ and $[x_i^*, y_i^*, z_i^*]$ denote the predicted rotation and translation values for the i-th sample.

C. Partitioned Fine-Tuning Strategy

Although the AMVS exhibits high structural precision and repeatability, manufacturing and assembly errors lead to subtle variations in the kinematic models across different units. Based on our tests, accurate calibration typically requires large datasets, with approximately 4500 samples needed to achieve acceptable reprojection errors. However, collecting such extensive data for each unit decreases ease of use and deployment efficiency. To reduce the required data, a pre-trained AMVS forward kinematic model is used. For each new unit, only the last three layers are fine-tuned with a small dataset, while the first four layers remain fixed (see Fig. 5(c)). A partitioned fine-tuning strategy based on pose sensitivity coefficients further improves fine-tuning accuracy.

In AMVS, three motors jointly control the three-axis rotation of the mobile platform. Due to the nonlinear characteristics of the SPM's forward kinematics, identical changes in motor angles can result in varying degrees of rotational pose changes at different initial positions. To quantify these rotational changes, we represent the platform's rotation using a rotation vector. The forward kinematic model of the SPM is adjusted as: $B_{rot}(\boldsymbol{\theta}) = \mathbf{r}$, where $\mathbf{r} = [r_x, r_y, r_z]^T$ is the rotation vector of the camera mounted on the AMVS. To measure the sensitivity of the mobile platform's pose to changes in motor angles, we define a **Pose Sensitivity Coefficient** (**PSC**), denoted as S. By applying a small increment $\Delta\theta$ to each proximal link position θ_i, new proximal link position vectors can be obtained:

$$\boldsymbol{\theta}'_i = \boldsymbol{\theta} + \Delta\theta \cdot \mathbf{e}_i, \quad i = 1, 2, 3 \tag{8}$$

where $\mathbf{e}_1 = [1, 0, 0]^T$, $\mathbf{e}_2 = [0, 1, 0]^T$, and $\mathbf{e}_3 = [0, 0, 1]^T$. Substituting the new proximal link positions $\boldsymbol{\theta}_i'$ into the forward kinematic model $B_{rot}(\boldsymbol{\theta})$, the updated rotation vectors are obtained:

$$\mathbf{r}_i = B_{rot}(\boldsymbol{\theta}'_i) \tag{9}$$

The change in the rotation vector is given by $\Delta\mathbf{r}_i = \mathbf{r}_i - \mathbf{r}$. The 2-norm of each $\Delta\mathbf{r}_i$ is then computed as $\|\Delta\mathbf{r}_i\|$, for i = 1,2,3.

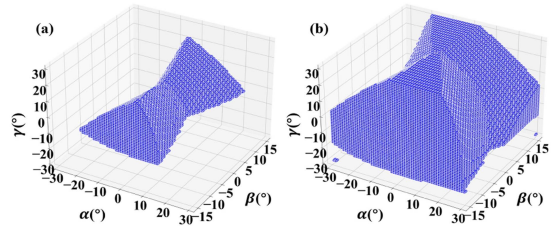

Fig. 6. 3D Voxel Distribution of PSCs for $S \leq 0.035$ (a) and $S \leq 0.041$ (b), Representing 12.3% and 59.2% of the Workspace.

The **PSC**, S is defined as the maximum rate of change among the three rotation vectors:

$$S = \max_i \left(\|\Delta\mathbf{r}_i\|/\Delta\theta\right) \tag{10}$$

The **PSC** reflects how sensitive the mobile platform's rotation vector is to changes in proximal link positions, providing a quantitative description of the impact of different proximal link variations on pose adjustment. To compute the **PSC**s across the entire workspace, the theoretical end poses $\mathbf{P}_{\mathrm{ideal},k}$ (k = 12, …,115351) defined in (3) are used. For each $\mathbf{P}_{\mathrm{ideal},k}$, the corresponding proximal link positions $\boldsymbol{\theta}_k = [\theta_{1k}, \theta_{2k}, \theta_{3k}]$ are calculated using inverse kinematics. Using these proximal link positions, the corresponding **PSC**, S_k is computed (with $\Delta\theta$ set to 1° in this letter). This yields the **PSC**, S_k for each theoretical end pose $\mathbf{P}_{\mathrm{ideal},k}$.

Fig. 6 presents the 3D voxel distributions of the **PSC** for $S \leq 0.035$ and $S \leq 0.041$. It is evident that points near the center of the AMVS theoretical workspace exhibit lower S. This observation aligns with the physical characteristics of the SPM, where proximal link rotations near the initial position cause smaller changes in the platform's angle. Consequently, higher angular resolution can be achieved near the initial position when the encoder resolution remains constant.

Utilizing the properties of the **PSC**s, the theoretical workspace is partitioned into internal and external regions. A sensitivity coefficient threshold S_{init} is defined; points with $S \leq S_{\mathrm{init}}$ constitute the internal region, while those with $S > S_{\mathrm{init}}$ form the external region. To prevent potential underfitting due to the limited fine-tuning dataset, an incremental value ΔS is introduced. Data satisfying $S > S_{\mathrm{init}}$ - ΔS are used to train the external model, and data with $S \leq S_{\mathrm{init}} + \Delta S$ are used for the internal model, thereby increasing the data available for fine-tuning both branches. During fine-tuning, the shared layers of the pre-trained model $\boldsymbol{M}_2$ remain unchanged. The two branches for rotation and translation estimation are each duplicated and connected to the shared layers, forming a four-branch network $\boldsymbol{M}_4$ (see Fig. 5(c)). Fine-tuning is performed separately on the rotation and translation branches for both internal and external regions, while the rest of the network remains frozen. In the inference phase, the model selects the appropriate branch's prediction based on the **PSC**.

VI. Experiments and Results

A. AMVS Repeatability Positioning Accuracy Experiment

To achieve accurate kinematic calibration for the AMVS, ensuring good repeatability in positioning accuracy is essential. This experiment evaluates the repeatability of the optimized AMVS based on ISO 9283 standards, using the setup shown in Fig. 3(a) with $\boldsymbol{C}_0$. The camera poses calculated from Aprilgrid

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
REPEATABILITY POSITIONAL ACCURACY RESULTS FOR AMVS

$\mathbf{P}_{\text{ideal}}$	α (Pitch)[°]	β (Roll)[°]	γ (Yaw)[°]	D [mm]
(3,15,18)	7.119±0.010	15.558±0.004	16.847±0.019	4.932±0.067
(-21,0,25)	−21.274±0.006	−0.066±0.003	25.115±0.010	7.392±0.074
(-11,-7,-26)	−6.653±0.007	−8.149±0.002	−27.299±0.004	12.731±0.086
(25,15,-26)	17.785±0.013	15.696±0.006	−31.564±0.004	9.187±0.048

a. The result data is in the form of mean ± sample standard deviation.

images are taken as the true camera poses. Due to space constraints, four randomly chosen AMVS theoretical poses ($\mathbf{P}_{\text{ideal},i}$, $i = 1,2,3,4$) were selected as target positions. For each $\mathbf{P}_{\text{ideal},i}$, 30 random poses were chosen within the AMVS theoretical workspace, and in accordance with ISO 9283, the system was moved from these starting positions to $\mathbf{P}_{\text{ideal},i}$. Notably, for each target position, the encoder value fluctuations remained within ±0.01° over the 30 trials. Aprilgrid images were recorded, and the camera poses were calculated. The repeatability results are shown in Table I.

In Table I, D represents the translation distance, i.e., the linear displacement measured during the positioning accuracy tests. The results show that the repeatability in a given rotational direction is 0.019°, and the overall rotational repeatability is 0.022°. However, significant absolute angular deviations were observed, with some rotational directions deviating up to 7.2°, and translation distances varying up to 7.8 mm depending on $\mathbf{P}_{\text{ideal},i}$. These results highlight the considerable discrepancy between the actual and ideal kinematics of the AMVS. Furthermore, analysis of the end-effector's velocity curve indicates a maximum speed of 1561°/s, exceeding the approximately 1000°/s reported in [8], which highlights the improved dynamic performance of the proposed AMVS design.

B. Experiments on the Two-Branch Optimization Network

To conduct comparative experiments, the commonly used hand-eye calibration method $\boldsymbol{M}_h$ was introduced [16]. Based on the ideal forward kinematics of the AMVS, $\boldsymbol{M}_h$ incorporates the relationship between the camera coordinate system and the AMVS end-effector to establish a more accurate forward kinematic model from proximal link positions to camera pose. Experiments were performed using DatasetC0. Sampling at 3° intervals across the three rotational axes within the theoretical workspace yielded $21 \times 11 \times 21 = 4851$ points for the training set. The remaining data was divided into validation and test sets in a 1:10 ratio.

For $\boldsymbol{M}_{\text{h}}$, all 4851 points were used for calibration, and the same test set was employed for accuracy assessment. $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ were trained with identical hyperparameters (ADAM optimizer, batch size = 128, learning rate = 1e-4, ReLU activation). For $\boldsymbol{M}_2$, the loss function defined in (5) was used with an experimentally determined weight ratio of w_{rot}:w_{trans} = 2:1. All models were trained for 1000 epochs with early stopping triggered if validation loss did not decrease for 100 consecutive epochs. Following calibration with $\boldsymbol{M}_{\text{h}}$, $\boldsymbol{M}_1$, and $\boldsymbol{M}_2$, absolute errors (maximum, mean, and standard deviation for each pose parameter [α, β, γ, x, y, z]) were evaluated on the test set, with error defined as $|\Delta\alpha| = |\alpha'\text{-}\alpha^*|$ (and similarly for the other parameters). Note that the maximum value represents the 99.9th percentile, not the absolute maximum. As shown in Table II, compared to $\boldsymbol{M}_{\text{h}}$, $\boldsymbol{M}_2$ reduces rotational error—computed using the Euclidean norm of $|\Delta\alpha|$, $|\Delta\beta|$ and $|\Delta\gamma|$—by over 93%, and the translation error by more than 69%. Relative to $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ further reduces the Euclidean-norm-based rotational error by 17% and the translation error by 3.9%, thereby significantly improving the accuracy of the AMVS forward kinematic model.

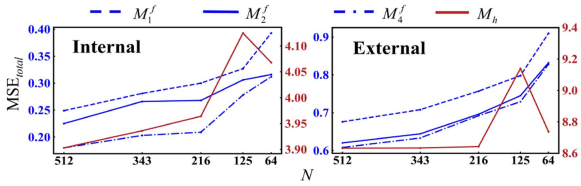

Fig. 7. Comparison of $\text{MSE}_{\text{total}}$ for Four Methods on Internal and External Test Sets with Different Fine-tuning Data Points. The red and blue data lines correspond to the red and blue vertical axes, respectively.

To further assess the impact of training data diversity on model generalization, the training range was restricted to ±20° pitch, ±20° yaw, and ±12° roll, with a 2° sampling interval, yielding a training set of 5072 points. $\boldsymbol{M}_1$' and $\boldsymbol{M}_2$' were trained under this constrained setting. Additionally, the test set was partitioned into a restricted region (within the reduced range) and an extrapolated region (beyond the reduced range), with generalization performance assessed using the test set's **total mean square error** (MSE_{total}):

$$\text{MSE}_{total} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \|\mathbf{P}_{\text{actual},i} - \mathbf{P}_{\text{calib},i}\|^2, \tag{11}$$

where N_{test} is the number of samples in the test set. On the restricted test set, $\boldsymbol{M}_1$' (0.081) and $\boldsymbol{M}_2$' (0.079) outperformed $\boldsymbol{M}_1$ (0.100) and $\boldsymbol{M}_2$ (0.088). However, for the extrapolated test set, MSE_{total} increased sharply for $\boldsymbol{M}_1$' (2.339) and $\boldsymbol{M}_2$' (1.904), compared to $\boldsymbol{M}_1$ (0.236) and $\boldsymbol{M}_2$ (0.222). Similarly, across the entire test set, MSE_{total} rose from 0.192 ($\boldsymbol{M}_1$) and 0.180 ($\boldsymbol{M}_2$) to 0.531 ($\boldsymbol{M}_1$') and 0.528 ($\boldsymbol{M}_2$'), highlighting the necessity of comprehensive training data coverage for robust generalization. Nonetheless, the two-branch optimization strategy remained effective across different training ranges.

C. Experiment on the Four-Branch Fine-Tuning Network

To verify the general effectiveness of $\boldsymbol{M}_4$ across different AMVS units, experiments were conducted using datasets collected from the $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_3$. To distinguish the fine-tuned models, a superscript $\boldsymbol{f}$ was added to the model symbols, where $\boldsymbol{M}_1^{\boldsymbol{f}}$, $\boldsymbol{M}_2^{\boldsymbol{f}}$, and $\boldsymbol{M}_4^{\boldsymbol{f}}$ represent the fine-tuned versions of $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, and $\boldsymbol{M}_4$, respectively. For the partitioned fine-tuning of $\boldsymbol{M}_4$, S_{init} from 0.036 to 0.043 (step 0.001) and ΔS of 0.001, 0.002, and 0.003 were evaluated. With a fixed ΔS, test set's MSE_{total} decreased and then increased as S_{init} rose, indicating an optimal threshold. The best performance was achieved with $S_{\text{init}} = 0.038$ and $\Delta S = 0.003$, a configuration adopted for all subsequent experiments. The corresponding **PSC**s of 0.035 and 0.041, along with their 3D voxel distributions, are shown in Fig. 6. For $\boldsymbol{M}_h$, calibration was performed directly using the fine-tuning datasets, while $\boldsymbol{M}_1^{\boldsymbol{f}}$, $\boldsymbol{M}_2^{\boldsymbol{f}}$, and $\boldsymbol{M}_4^{\boldsymbol{f}}$ were fine-tuned based on the full models $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, and $\boldsymbol{M}_4$ trained on DatasetC0 in Section VI-B. The test set consisted of the remaining data after excluding the fine-tuning samples.

First, the methods were evaluated under different fine-tuning data volumes. Based on DatasetC1, five datasets were created by sampling 8, 7, 6, 5, and 4 points along the three rotational axes, generating 512, 343, 216, 125, and 64 training samples. Fig. 7 shows the MSE_{total} results for camera pose predictions

TABLE II
RESULTS OF ROTATION AND TRANSLATION ERRORS FROM CALIBRATIONS USING $\mathbf{M}_h$, $\mathbf{M}_1$, AND $\mathbf{M}_2$, COMPUTED ON THE DATASETC0 TEST SET

Method	$\lvert\Delta\alpha\rvert$(°)			$\lvert\Delta\beta\rvert$(°)			$\lvert\Delta\gamma\rvert$(°)			$\lvert\Delta x\rvert$(mm)			$\lvert\Delta y\rvert$(mm)			$\lvert\Delta z\rvert$(mm)		
	Max	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*
M_h	10.455	2.053	1.822	2.566	0.443	0.450	8.111	2.059	1.718	5.741	1.714	0.986	12.377	0.960	0.909	4.504	1.047	0.652
M_1	0.723	0.138	0.116	**1.013**	0.152	0.158	0.779	0.136	0.114	**3.350**	0.412	0.427	4.600	0.375	0.442	2.718	0.428	0.374
M_2	**0.653**	**0.116**	**0.097**	1.014	**0.122**	**0.155**	**0.606**	**0.115**	**0.094**	3.378	**0.398**	**0.422**	**4.532**	**0.364**	**0.441**	**2.499**	**0.405**	**0.372**

TABLE III
COMPARISON OF ROTATIONAL AND TRANSLATIONAL ERRORS FOR $\mathbf{C}_0$, $\mathbf{C}_1$, AND $\mathbf{C}_2$ WITH **343** FIXED FINE-TUNING SAMPLES

AMVS	Method	$\lvert\Delta\alpha\rvert$(°)			$\lvert\Delta\beta\rvert$(°)			$\lvert\Delta\gamma\rvert$(°)			$\lvert\Delta x\rvert$(mm)			$\lvert\Delta y\rvert$(mm)			$\lvert\Delta z\rvert$(mm)		
		Max	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*	*Max*	*Mean*	*Std.*
C_1	M_h	11.899	2.099	1.925	3.024	0.600	0.445	8.729	1.063	1.800	5.938	2.065	0.904	6.978	0.907	0.678	3.626	1.209	0.693
	M_2^{C0}	1.526	0.362	0.277	1.412	0.434	0.295	2.124	0.552	0.416	4.684	0.767	0.592	4.645	0.723	0.675	6.831	2.139	1.519
	M_1^f	1.235	0.247	0.223	1.331	0.253	0.203	1.172	0.247	0.200	3.937	0.456	0.435	3.425	0.412	0.387	3.011	0.650	0.503
	M_2^f	**0.898**	0.211	0.170	1.086	0.227	0.171	1.148	0.237	0.202	3.945	0.428	**0.428**	3.528	0.399	0.407	2.895	0.625	0.483
	M_4^f	0.916	**0.206**	**0.170**	**1.052**	**0.208**	**0.166**	**1.130**	**0.235**	**0.201**	**3.931**	**0.418**	0.430	**3.511**	**0.390**	**0.406**	**2.773**	**0.584**	**0.470**
C_2	M_h	11.282	2.258	2.023	3.262	0.471	0.465	8.408	1.960	1.712	6.863	0.848	0.860	7.016	0.816	0.569	2.760	1.103	0.421
	M_2^{C0}	4.324	0.974	0.865	2.088	0.409	0.308	1.181	0.259	0.195	6.820	1.023	1.051	4.609	0.592	0.502	8.102	2.742	1.795
	M_1^f	1.030	0.222	0.179	1.101	0.206	0.181	0.939	0.215	0.166	4.022	0.400	0.415	3.482	0.353	0.336	1.540	0.347	0.255
	M_2^f	0.921	0.196	0.161	0.972	0.157	0.151	0.946	0.206	0.165	3.813	0.287	0.354	**3.355**	0.252	**0.288**	1.391	0.218	0.192
	M_4^f	**0.887**	**0.186**	**0.155**	**0.918**	**0.147**	**0.141**	**0.946**	**0.195**	**0.158**	**3.812**	**0.277**	**0.351**	3.387	**0.245**	0.292	**1.345**	**0.211**	**0.178**
C_3	M_h	10.297	2.124	1.903	2.675	0.417	0.427	7.891	1.989	1.727	5.900	0.872	0.742	7.026	0.794	0.635	2.509	0.820	0.449
	M_2^{C0}	2.087	0.532	0.371	1.422	0.307	0.231	1.387	0.325	0.246	5.132	0.797	0.745	4.380	0.832	0.547	5.964	1.744	1.209
	M_1^f	1.017	0.217	0.173	1.230	0.194	0.181	1.012	0.218	0.169	3.399	0.354	0.355	3.298	0.332	0.310	1.366	0.325	0.258
	M_2^f	0.841	0.187	0.150	0.928	0.148	0.142	1.001	0.212	0.172	3.151	**0.242**	0.292	**3.057**	0.248	0.280	**1.226**	0.199	0.163
	M_4^f	**0.825**	**0.178**	**0.146**	**0.887**	**0.140**	**0.135**	**0.984**	**0.204**	**0.167**	**3.095**	0.247	**0.289**	3.114	**0.236**	**0.280**	1.238	**0.199**	**0.162**

on internal and external data. Both M_2^f, and M_4^f significantly outperformed M_h and M_1^f, with M_4^f showing the best performance, particularly in the internal region.

Next, to validate the proposed method across different devices, fine-tuning experiments were conducted on datasets from C_1, C_2, and C_3. Due to space limitations, only results with a fixed fine-tuning dataset of 343 samples per network, based on test set data, are presented. In addition to the four previously mentioned methods, a new approach, M_2^{C0}, was introduced, which reuses the fully trained M_2 model from DatasetC0 to infer the camera poses for other AMVS units. Table III summarizes the maximum (99.9th percentile), mean, and standard deviation of angular errors ($|\Delta\alpha|$, $|\Delta\beta|$, $|\Delta\gamma|$) and positional errors ($|\Delta x|$, $|\Delta y|$, $|\Delta z|$) for C_1, C_2, and C_3, using 343 fine-tuning points. The results in Table III show that M_2^f and M_4^f performed exceptionally well on C_1, C_2, and C_3, maintaining the lowest error means and standard deviations. Notably, M_4^f consistently achieved the best results, significantly reducing angular errors and moderately improving positional errors, demonstrating superior accuracy and stability. These findings confirm that the proposed partitioned fine-tuning strategy effectively enhances calibration accuracy with small fine-tuning datasets, validating its generality and practical value.

Finally, to visually demonstrate the improvement in calibration accuracy brought by the proposed M_2 and M_4^f, the mean projection error (**MPE**) across the entire AMVS workspace was evaluated. Specifically, the Aprilgrid corner points detected at the AMVS's home configuration were projected to the current pose using different methods, and the **MPE** was calculated as the average pixel distance of all corresponding Aprilgrid corner points.

Fig. 8 shows the **MPE** distribution of the theoretical end poses $\mathbf{P}_{\text{ideal}}$ for C_0 and C_1 using various methods. The results indicate that M_h results in larger **MPE**, with errors gradually increasing from the center to the peripheral regions. This demonstrates that as the AMVS mobile platform's Pitch/Yaw angles increase, the discrepancy between the ideal kinematic model and the actual kinematic model grows. For the full-data network models, the proposed M_2 improved accuracy across the entire workspace,

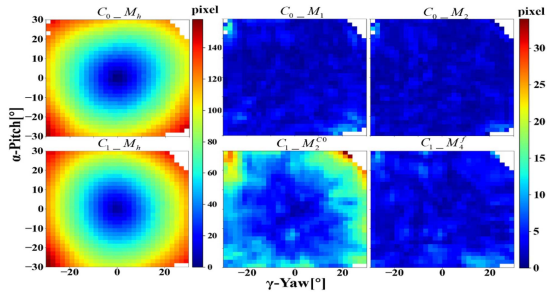

Fig. 8. **Mean Reprojection Error Heatmaps on Test Datasets.** M_4^f for C_1 was fine-tuned using 343 data points. The workspace is divided into $2° \times 2°$ grids for α and γ, with each grid showing the average **MPE** at that pose. Blank grids indicate unreachable positions for the AMVS. The first column uses the left colorbar and the others use the right.

especially in regions with large rotations, reducing the average **MPE** from 2.66 pixels in M_1 to 2.06 pixels. Additionally, the M_4^f, fine-tuned using only 343 data points from C_1, achieved results close to the full-data model M_1, reducing the average **MPE** from 9.61 pixels in M_2^{C0} to 3.47 pixels.

D. Experiment With ABVS

Stereoscopic reconstruction is sensitive to external parameter errors between cameras, especially in ABVS, where binocular positions change over time. This experiment evaluates ABVS reconstruction accuracy after binocular movement using the proposed calibration method. ABVS_0 (Fig. 3(a)) was mounted on the humanoid robot (Fig. 1(b)), with C_0 using the M_2 calibration model and C_1 using the M_4^f model. The homogeneous transformation from the left to the right camera $\mathbf{T}_l^r = \mathbf{T}_{rH}^r \mathbf{T}_{lH}^{rH} \mathbf{T}_l^{lH}$, where $\mathbf{T}_{rH}^r$ and $\mathbf{T}_l^{lH}$ are the transformations from home to current positions for the right and left cameras, estimated using their respective calibration models. $\mathbf{T}_{lH}^{rH}$, the transformation between the cameras' home configurations, is obtained through stereo calibration.

To evaluate stereoscopic accuracy during ABVS motion, 2999 image and motor data sets were collected from various

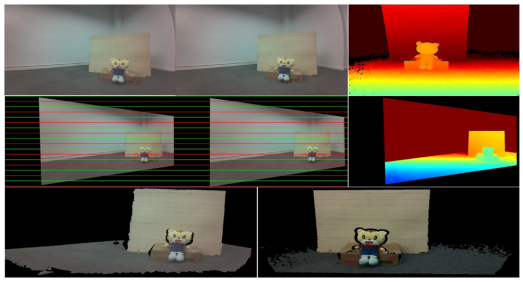

Fig. 9. **Stereo Reconstruction**. The top-left image corresponds to the camera pose (0.939°, 0.501°, 14.972°) for $\boldsymbol{C}_0$ and (0.947°, 0.504°, 14.957°) for $\boldsymbol{C}_1$. Using the compensated poses of the two cameras, the **middle-right** image depicts the depth map computed by HITNet [24], the **middle-left** image depicts the alignment of $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$ images along epipolar lines, the depth map (**middle-right**) is converted to point cloud (**bottom-left**) with depth between **0.2m to 1.6m**. The **top-right** and **bottom-right** images depict the depth map and point cloud computed by Intel RealSense L515 LiDAR**,** respectively.

angles, and the corresponding $\mathbf{T}_{lH}^{rH}$ was computed. Epipolar lines were aligned, depth maps generated with HITNet [24], and point clouds reconstructed. The results were compared with RealSense L515 LiDAR. Fig. 9 shows that the proposed two-branch and four-branch fine-tuned models produced accurate depth maps and point clouds, confirming the effectiveness of the proposed calibration in achieving accurate forward kinematic calibration of AMVS based on SPM.

VII. Discussion and Conclusion

This letter presents an optimized 6-DoF ABVS based on SPM, with satisfactory repeatability, successfully integrated into a humanoid robotic head. To achieve accurate forward kinematics, we introduced a two-branch optimization network and a four-branch fine-tuning network, improving model precision while reducing training data requirements. Experiments VI-B and VI-C revealed significant discrepancies between actual and ideal kinematics, reflected in large hand-eye calibration errors. While the proposed neural network strategies improved calibration accuracy and stereo reconstruction, the average **MPE** in VI-C remains higher than that of serial ABVS [5] or fixed binocular systems, highlighting the challenge of achieving high-precision depth estimation in SPM-based ABVS. Preliminary tests using a Transformer-based baseline combined with a two-branch network indicate that the benefits of the multi-branch architecture are not dependent on a specific backbone, achieving accuracy comparable to or even exceeding that of an MLP. Nonetheless, given the limited data availability and real-time constraints in robotic systems, the lightweight MLP remains a attractive option.

In the future, to address the high costs stemming from the current system's need for high-precision mechanical processing, we will explore using lower-precision components (e.g., 3D printed parts [23], [25],) combined with deep learning for stereovision functionality. Moreover, while this work focuses on constructing an accurate forward kinematics model, we will also investigate an inverse kinematics model to enable precise saccade control.

References

[1] S. Chen, Y. Li, and N. M. Kwok, "Active vision in robotic systems: A survey of recent developments," *Int. J. Robot. Res.*, vol. 30, no. 11, pp. 1343–1377, Sep. 2011.

[2] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "Gaze stabilization for humanoid robots: A comprehensive framework," in *Proc. 2014 IEEE-RAS Int. Conf. Humanoid Robots*, 2014, pp. 259–264.

[3] S. Bazeille et al., "Active camera stabilization to enhance the vision of agile legged robots," *Robotica*, vol. 35, no. 4, pp. 942–960, 2017.

[4] I. Sanchez, R. Martin, F. Ussa, and I. Fernandez-Bueno, "The parameters of the porcine eyeball," *Graefe's Arch. Clin. Exp. Ophthalmol.*, vol. 249, no. 4, pp. 475–482, Apr. 2011.

[5] C. Zhou, Q. Sun, K. Wang, J. Li, and X. Zhang, "Simultaneous calibration of multiple revolute joints for articulated vision systems via SE(3) kinematic bundle adjustment," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 12161–12168, Oct. 2022.

[6] H. Saafi, M. A. Laribi, and S. Zeghloul, "Forward kinematic model improvement of a spherical parallel manipulator using an extra sensor," *Mechanism Mach. Theory*, vol. 91, pp. 102–119, Sep. 2015.

[7] P. Renaud, N. Andreff, P. Martinet, and G. Gogu, "Kinematic calibration of parallel mechanisms: A novel approach using legs observation," *IEEE Trans. Robot.*, vol. 21, no. 4, pp. 529–538, Aug. 2005.

[8] C. M. Gosselin, E. St. Pierre, and M. Gagne, "On the development of the agile eye," *IEEE Robot. Automat. Mag.*, vol. 3, no. 4, pp. 29–37, Dec. 1996.

[9] V. Leconte et al., "First-person teleoperation of humanoid robot Reachy by valid and arm-amputated participants with a novel movement-based prosthesis control," TechRxiv, Aug. 1, 2024. [Online]. Available: https://doi.org/10.36227/techrxiv.172253934.48844352/v1

[10] E. S. Maini, G. Teti, M. Rubino, C. Laschi, and P. Dario, "Bio-inspired control of eye-head coordination in a robotic anthropomorphic head," in *Proc. 1st IEEE/RAS-EMBS Int. Conf. Biomed. Robot. Biomechatron.*, 2006, pp. 549–554.

[11] Q. Wang, W. Zou, D. Xu, and Z. Zhu, "Motion control in saccade and smooth pursuit for bionic eye based on three-dimensional coordinates," *J. Bionic Eng.*, vol. 14, no. 2, pp. 336–347, Jun. 2017.

[12] D. Fan et al., "Eye gaze based 3D triangulation for robotic bionic eyes," *Sensors*, vol. 20, no. 18, Sep. 2020, Art. no. 5271.

[13] H. Li et al., "Design and control of 3-DoF spherical parallel mechanism robot eyes inspired by the binocular vestibule-ocular reflex," *J. Intell. Robot. Syst.*, vol. 78, no. 3–4, pp. 425–441, Jun. 2015.

[14] P. Bai et al., "Kinematic calibration of Delta robot using distance measurements," *Proc. Inst. Mech. Engineers, Part C, J. Mech. Eng. Sci.*, vol. 230, no. 3, pp. 414–424, Feb. 2016.

[15] D. Deblaise and P. Maurine, "Effective geometrical calibration of a delta parallel robot used in neurosurgery," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 1313–1318.

[16] G. Palmieri et al., "Vision-based kinematic calibration of a small-scale spherical parallel kinematic machine," *Robot. Comput.-Integr. Manuf.*, vol. 49, pp. 162–169, Feb. 2018.

[17] K. Daniilidis, "Hand-eye calibration using dual quaternions," *Int. J. Robot. Res.*, vol. 18, no. 3, pp. 286–298, 1999.

[18] L. Ghorbani et al., "Forward kinematics of a 6x6 UPU parallel mechanism by ANFIS method," in *Proc. IEEE 6th Int. Conf. Control Eng. Inf. Technol.*, 2018, pp. 1–6.

[19] H. Liu, Z. Yan, and J. Xiao, "Pose error prediction and real-time compensation of a 5-DOF hybrid robot," *Mechanism Mach. Theory*, vol. 170, Apr. 2022, Art. no. 104737.

[20] P. Renaud, N. Andreff, J.-M. Lavest, and M. Dhome, "Simplifying the kinematic calibration of parallel mechanisms using vision-based metrology," *IEEE Trans. Robot.*, vol. 22, no. 1, pp. 12–22, Feb. 2006.

[21] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. 2011 IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3400–3407.

[22] X. Kong and C. M. Gosselin, "A formula that produces a unique solution to the forward displacement analysis of a quadratic spherical parallel manipulator: The agile eye," *J. Mechanisms Robot.*, vol. 2, no. 4, Nov. 2010, Art. no. 044501.

[23] I. Tursynbek, A. Niyetkaliye, and A. Shintemirov, "Computation of unique kinematic solutions of a spherical parallel manipulator with coaxial input shafts," in *Proc. IEEE 15th Int. Conf. Automat. Sci. Eng.*, Vancouver, BC, Canada, 2019, pp. 1524–1531.

[24] V. Tankovich, C. Häne, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HITNet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2021, pp. 14357–14367.

[25] N. Saiki et al., "2-DOF spherical parallel mechanism capable of biaxial swing motion with active arc sliders," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4680–4687, Jul. 2021.

[26] I. Cvii, J. Esi, I. Markovi, and I. Petrovi, "SOFT-SLAM: Computationally efficient stereo visual SLAM for autonomous UAVs," *J. Field Robot.*, vol. 35, pp. 578–595, 2017.