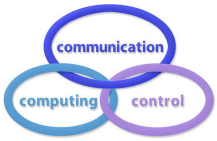
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 6, Month: December, Year: 2026
Article Number: 7424, https://doi.org/10.15837/ijccc.2026.6.7424



CCC Publications



Real-Time Dual-Modal 6D Pose Estimation for Virtual Reality Interaction Using Deep Learning

R. Hu, R. Li, P. Huang

Rong Hu[1,2,*]
[1] School of Intelligent Technology,
Geely University, Chengdu 641423, Sichuan, China
[2] Beihang University,
State Key Laboratory of Virtual Reality Technology and Systems
*Corresponding author: hurong@guc.edu.cn

Runlin Li
[1] School of Intelligent Technology,
Geely University, Chengdu 641423, Sichuan, China
[2] Beihang University,
State Key Laboratory of Virtual Reality Technology and Systems
lirunlin@guc.edu.cn

Ping Huang
College of Mathematics and Computer,
Panzhihua University, Panzhihua 617000, Sichuan, China
1269228282@qq.com

Abstract

Accurate and low-latency 6D pose estimation is a key requirement for natural interaction in virtual reality systems, yet existing methods often struggle to simultaneously satisfy precision, real-time performance, and robustness in multi-user or interference-prone environments. To address these challenges, this study proposes a deep learning-based dual-modal 6D pose estimation framework that integrates RGB and depth sensing for real-time virtual reality interaction. The proposed method employs a two-stream feature extraction architecture, where two-dimensional texture features are extracted from RGB images using a ResNet-18 backbone enhanced with efficient channel attention, while three-dimensional geometric features are obtained from depth data through an SO(3)-equivariant encoder combined with vector neuron linear layers and multilayer perceptrons. A cross-modal fusion strategy is introduced to achieve complementary enhancement between texture and geometric representations, followed by pose regression based on point pair feature matching. Experimental results on public benchmarks and multi-user virtual interaction scenarios demonstrate that the proposed approach achieves a median rotation error of 0.42° and a translation error of 0.38 cm, representing an error reduction of over 67% compared to models without SO(3) encoding. The system maintains a success rate of 95-98% under normal illumination and achieves a processing latency of 4.81 ms with 208 FPS on high-end hardware. Furthermore,

the framework exhibits strong robustness to lighting variation, occlusion, and cross-user interference. These results indicate that the proposed dual-modal, SO(3)-aware pose estimation framework provides an effective computational solution for real-time 6D pose tracking and control in virtual reality interaction systems.

Keywords: 6D pose estimation; RGB-D fusion; SO(3)-equivariant representation; real-time systems; virtual reality interaction; deep learning

1 Introduction

Virtual reality (VR) systems are increasingly deployed in applications that require real-time perception, low-latency interaction, and precise motion control, including immersive training, human-machine collaboration, and intelligent simulation environments [25]. In such systems, 6D pose estimation, which jointly estimates three-dimensional (3D) position and orientation, constitutes a fundamental computational task that directly influences interaction accuracy, control stability, and user immersion. From the perspective of intelligent systems and control, inaccurate or delayed pose estimation may lead to perceptual drift, control instability, and degraded system responsiveness, thereby limiting the practical deployment of VR interaction technologies.

Despite recent advances in deep learning-based perception, achieving simultaneously high accuracy, real-time performance, and robustness in 6D pose estimation remains challenging, particularly in dynamic and multi-user environments. Single-modality approaches based solely on RGB images are sensitive to illumination changes and occlusion, while point cloud-based methods relying on depth data often suffer from high computational overhead and insufficient temporal performance [19]. Traditional multi-modal fusion strategies, although capable of leveraging complementary information, frequently employ shallow fusion mechanisms that fail to fully exploit the structural consistency between two-dimensional texture features and 3D geometric representations [2]. From a computational intelligence standpoint, an effective 6D pose estimation framework for VR interaction must satisfy three key requirements. First, it should provide rotation-consistent and geometry-aware feature representations to ensure stable pose estimation under arbitrary viewpoint changes. Second, it must support real-time inference with bounded latency, enabling closed-loop interaction and control. Third, it should exhibit robustness to environmental disturbances, including illumination variation, partial occlusion, and inter-user interference, which are common in practical VR scenarios.

Recent research has explored deep learning techniques for 6D pose estimation, including monocular self-supervised learning, point cloud registration, and RGB-D feature fusion. While these methods have demonstrated promising accuracy on benchmark datasets, limitations remain in terms of cross-modal feature consistency, computational efficiency, and robustness in complex interactive environments [7]. In particular, most existing approaches lack explicit mechanisms to enforce equivariance in 3D rotation space, which is critical for maintaining geometric consistency in pose estimation and control tasks.

To address these challenges, this study proposes a real-time dual-modal 6D pose estimation framework that integrates RGB and depth sensing within a deep learning architecture designed for rotation consistency and computational efficiency. The core contribution lies in the introduction of an SO(3)-equivariant geometric feature encoder, which enables rotation-aware representation of depth-based spatial information, and its integration with a two-dimensional texture feature extraction network enhanced by efficient channel attention. By coupling these representations through cross-modal fusion and point pair feature matching, the proposed framework achieves accurate and stable pose estimation suitable for real-time VR interaction [8].

Unlike existing methods that treat pose estimation primarily as an offline perception task, this study emphasizes its role as a real-time computational component within an interactive control system. Extensive experiments conducted on public datasets and multi-user VR scenarios demonstrate that the proposed approach achieves sub-degree rotational accuracy, sub-centimeter translational error, and millisecond-level latency, while maintaining robustness against lighting variation, occlusion, and cross-user interference.

The remainder of this study is organized as follows. Section 2 reviews related work on 6D pose estimation and multi-modal feature fusion. Section 3 presents the proposed dual-modal pose es-

timation framework, including feature extraction, SO(3)-equivariant encoding, and pose regression. Section 4 reports experimental results and performance evaluations under various environmental and system configurations. Section 5 concludes the paper and discusses future directions for optimizing computational efficiency and robustness in real-time interactive systems.

2 Literature review

6D pose estimation is the core support for natural VR interaction. Currently, many experts are conducting research on 6D pose estimation. To overcome the limitations of convolutional neural networks that are highly data-dependent and time-consuming to obtain annotations in monocular 6D object pose estimation, Wang G et al. proposed a self-supervised learning monocular 6D pose estimation method that does not require real annotations. This method used synthetic RGB data to fully supervised train the proposed network. Results showed that this method outperformed methods relying on synthetic data [21]. To overcome the difficulties of high label data collection cost and lack of large-scale datasets in underwater multi-robot 6D attitude estimation, Tang J Y et al. proposed the benchmark dataset ROV6D. The training subset of this dataset contained a large number of synthetic images with 6D pose real data, and the realism was enhanced by an underwater neural rendering strategy. The results showed that among the various methods evaluated on this dataset, the method based on dense coordinates performed the best [20]. Dang Z et al. pointed out two major challenges and proposed solutions to the problem that the 6D pose estimation method of point cloud objects based on learning fails in real-world data. The first solution was to introduce a direct supervision loss function to solve the sensitivity problem based on the singular value decomposition loss function. The second solution was to propose a matching normalization strategy to solve the problem of feature distribution differences. The results showed that this method could achieve meaningful results on real scene datasets [6].

To solve the problem of missing depth information and inefficient fusion of RGB and depth features in 6D pose estimation based on RGB images, Liu P L et al. proposed a method of combining a bidirectional depth residual fusion network with a depth prediction network, and constructed a bidirectional depth residual structure. In addition, this study also proposed a two-dimensional key point selection method that considered object appearance and geometric information. The results showed that the method outperformed the most advanced methods on three benchmark dataset [14]. To solve the problem that existing monocular 6D pose estimation methods ignore 3D geometric features and topological clues in differentiable n-point perspective algorithms, Cao T et al. proposed an improved end-to-end method. This algorithm utilized the estimated depth information to guide the extraction process of 2D-3D correspondences. The results showed that the network outperformed existing methods in both performance and efficiency [3]. To enhance the complementarity between multi-modalities and solve the problems of single-modal 6D pose estimation due to modal limitations, Lin S F et al. proposed a neighbor-based cross-modal attention mechanism. In addition, this study also designed some features between the rendered image and the original image to obtain confidence in the pose estimation results. The results showed that this method had good performance on different datasets [11].

In summary, existing methods have made good progress in modal fusion and geometric feature utilization. However, there are still certain problems: Firstly, most methods do not introduce SO(3) rotation invariant constraints, resulting in poor feature consistency from any perspective; Secondly, there is insufficient real-time optimization for multi-user VR interaction, making it difficult to meet millisecond level low latency requirements; Thirdly, cross-modal feature fusion often employs shallow mechanisms, resulting in insufficient complementarity between texture and geometric information; The fourth issue is insufficient robustness in scenarios with changes in lighting, occlusion, and user interference. Based on this, this study uses VR large screen, RGB and depth cameras in VR natural interaction technology to build a dual-modal acquisition network, and combines the deep learning technology ResNet-18 with the special orthogonal group (SO(3)) encoder to extract features. The purpose of this study is to improve the accuracy, real-time performance and anti-interference of 6D pose estimation in VR scenes, and to optimize the natural interaction experience.

3 Research methodology

To conduct accurate 6D pose estimation, this study first uses RGB cameras and depth cameras, and designed the deployment position of the cameras. Next, the study designs two-dimensional and 3D feature extraction methods. Finally, this study constructs an accurate 6D pose estimation method that integrates feature extraction, point pair matching, and pose regression.

3.1 Image acquisition design for accurate 6D pose estimation

To better predict the position and posture of VR users and thereby enhance their experience of VR natural interaction technology, this study uses RGB cameras and depth cameras to obtain user action image data. Then, based on the image data acquired by different cameras, the study designs the final accurate 6D pose estimation method. The core advantage of RGB cameras is that they can capture high-resolution, colorful texture information, have low hardware costs, and meet the low-latency requirements of real-time interaction [4, 17]. The core advantage of the depth camera is that it can directly obtain depth information in 3D space and accurately calculate the spatial coordinates and relative distances of joint points. It can provide key 3D geometric constraints for 6D pose estimation [12, 24]. This study deploys these two cameras around the VR large screen to form a collection network with multi-view coverage. Deployment details are shown in Figure 1.

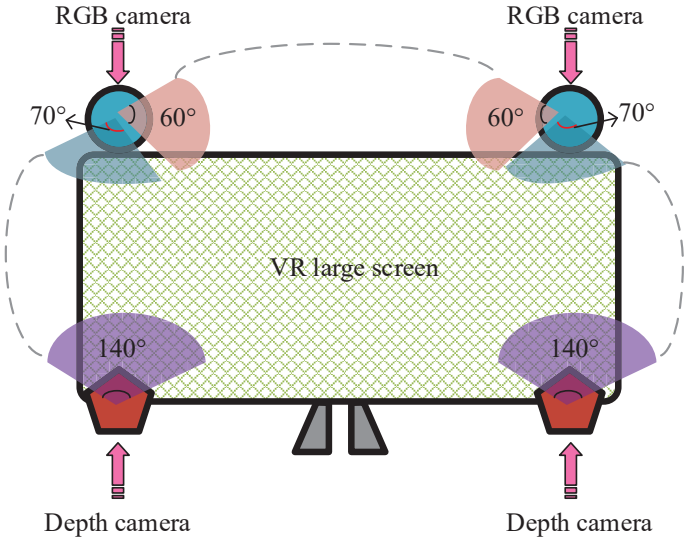


Figure 1: Deployment details of two types of cameras.

In Figure 1, around the VR large screen, the study adopts a streamlined deployment plan of "2 RGB cameras+2 depth cameras". Specifically, two RGB cameras, modeled Logitech C920s, are installed on the left and right ends of the top of the VR large screen. The monitoring range of each camera is 70° horizontally and 60° vertically, and can cover the area directly in front of the VR large screen to capture the texture details of the user's movements. Meanwhile, two Microsoft Azure Kinect DK depth cameras are deployed on the left and right edges of the bottom of the VR screen. The monitoring range is 140° horizontally, which can completely obtain the spatial geometric data of the user's limbs. The monitoring areas of these two types of cameras complement each other and can jointly cover the interactive core area of 1.5 meters to 3 meters directly in front of the large screen, meeting the multi-modal data requirements of 6D pose estimation.

In addition, in terms of camera calibration, the Zhang calibration method is used to obtain the intrinsic parameters and distortion coefficients of RGB cameras and depth cameras, and the extrinsic calibration is completed through a checkerboard calibration board to achieve spatial alignment between RGB and depth cameras; The RGB-D alignment stage projects the depth point cloud onto the RGB image plane based on the extrinsic matrix, achieving pixel level precise alignment; Time synchronization adopts a combination of hardware triggered synchronization and timestamp align-

ment to control the time deviation between multiple cameras within ±1ms; In terms of view strategy, multi-view fusion is not performed, but the optimal single-path view with unobstructed field of view and clearest features is selected for pose estimation.

3.2 Construction of accurate 6D pose estimation method considering different image types

3.2.1 Design of TFEN method for 6D pose estimation method

This study designs an image acquisition method for accurate 6D pose estimation, laying a good data foundation. To better perform 6D pose estimation, this study designs a 6D pose estimation algorithm starting from RGB images and depth images to better adapt to the dynamic capture needs of user body movements in VR natural interaction scenes. In the 6D pose estimation algorithm, three core modules are researched and designed. Among them, the first module is the feature extraction module, and this module includes TFEN and 3D geometric feature extraction.

On TFEN, the study adopts ResNet-18, which uses this network to capture the multi-scale texture features of RGB images. ResNet-18 is a classic residual convolutional neural network in the field of deep learning. It solves the vanishing gradient problem and has the advantages of moderate parameter quantity and high computational efficiency [5, 9, 16]. To better exert the effect of ResNet-18, this study introduces efficient channel attention into its structure. Efficient channel attention is a lightweight channel attention mechanism that can enhance attention to task-critical feature channels by adaptively learning the importance weights of feature channels [10, 13]. The structure of efficient channel attention is shown in Figure 2.

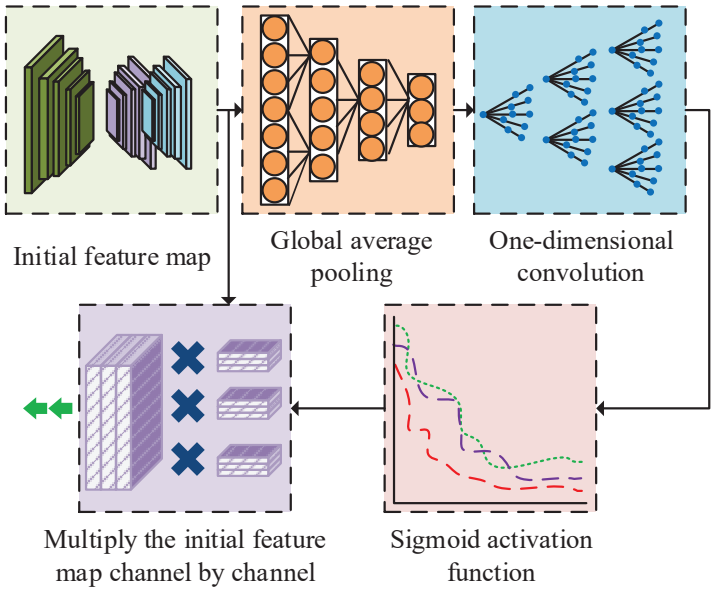


Figure 2: The structure of efficient channel attention.

In Figure 2, the structure of efficient channel attention includes the initial feature map, global average pooling, one-dimensional convolution (calculating channel weight), Sigmoid activation function (output weight), and the multiplication of the processed feature map and the initial feature map channel by channel. The attention weight of the channel is calculated as shown in Equation (1).

$$\omega_c = \sigma\left(Conv_{1\times1}\left(\beta\left(F_c\right)\right)\right) \tag{1}$$

In Equation (1), ω_c represents the attention weight of the cth channel. σ is the Sigmoid activation function. $Conv_{1\times1}$ stands for 1×1 convolutional layer. β stands for global average pooling operation.

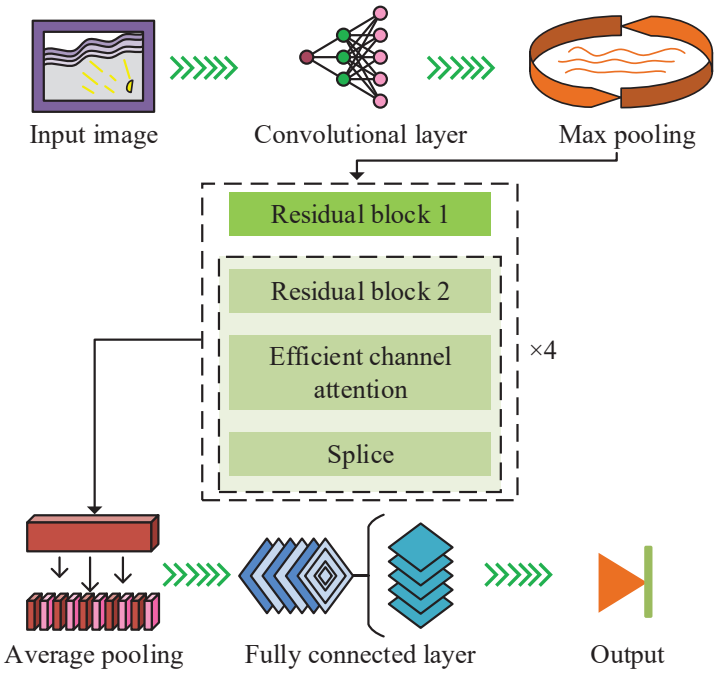


Figure 3: The structure of a TFEN designed for research.

F_c is the feature map of the cth channel in TFEN. The structure of the research-designed TFEN is shown in Figure 3.

Through Figure 3, the TFEN consists of convolutional layer, maximum pooling, efficient channel attention, residual block, average pooling, and fully connected layer. The specific process of TFEN is as follows: First, RGB image blocks are input, and preliminary dimensionality reduction is performed through the convolution layer and the maximum pooling layer. Second, feature enhancement is performed through 4 residual blocks containing efficient channel attention modules. Finally, after performing feature map space dimension compression, nonlinear transformation and dimensional mapping of features, a two-dimensional texture feature vector is output.

3.2.2 Design of three-dimensional geometric feature extraction method for 6D pose estimation method

This research has completed the design and construction of the TFEN method. To further improve the feature extraction system adapted to VR interactive scenes, this study designs a 3D geometric feature extraction method for depth images. In terms of 3D geometric feature extraction, this study adopts the SO(3) equivariant method and designs the final SO(3) encoder. The core advantage of the SO(3) equivariant method is that it can maintain the consistency of feature expression when the object posture changes. It can effectively avoid feature offset caused by data enhancement and greatly improve the robustness of pose estimation [18]. Therefore, this study provides spatial invariance guarantee through SO(3) equivariant method. Meanwhile, this study also introduces the vector neurons linear layer (VN-Linear) and multi-layer perceptron (MLP) methods. VN-Linear can directly process geometric data in the form of 3D vectors, and can provide support for extracting refined spatial features [1]. The MLP method can mine deep correlation information in geometric data, has high computational efficiency, and adapts to the real-time requirements of VR interaction [22, 23]. Therefore, this study uses VN-Linear to optimize the geometric data processing method, and uses MLP to strengthen feature depth mining. This research combines these three technologies to form the final SO(3) encoder. The structure of the SO(3) encoder is shown in Figure 4.

In Figure 4, the SO(3) encoder contains input (point cloud data converted from depth image), MLP, VN-Linear of three different sizes, splicing operation and geometric feature output. Through the SO(3) encoder, it can provide spatially constrained complements for 2D texture features. The

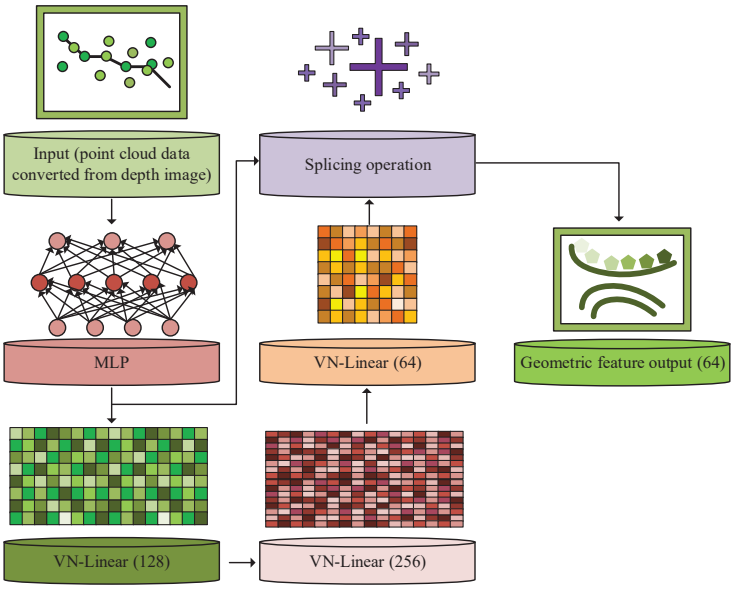


Figure 4: Structure of SO(3) encoder.

integration of two-dimensional features and geometric features is shown in Equation (2).

$$F_r = \alpha \cdot F_{2D} \otimes A_{2D \to 3D} + (1 - \alpha) \cdot F_{3D} \otimes A_{3D \to 2D} \tag{2}$$

In Equation (2), α is the fusion weight. $\otimes$ stands for element-wise multiplication. F_{2D} is a two-dimensional feature. F_{3D} represents three-dimensional features. $A_{2D \to 3D}$ represents the attention map of two-dimensional features to three-dimensional features. $A_{3D \to 2D}$ represents the attention map of three-dimensional features to two-dimensional features. Through the fused features, the texture detail advantages of RGB images and the spatial geometric characteristics of depth images can be fully integrated to achieve two-way complementary enhancement of two-dimensional texture information and three-dimensional structural information. This can provide core technical support for capturing user body movements in VR natural interaction scenes.

3.2.3 Design of point pair matching module and pose regression module

This study designs the first module in 6D pose estimation, which is the feature extraction module. The second module in 6D pose estimation is the point pair matching module, which is the key bridge connecting feature extraction and pose regression in 6D pose estimation. The core of point pair matching is to establish an accurate correspondence between the feature points of the collected image and the feature points of the preset target model to provide core data support for subsequent pose regression. Therefore, in point pair matching, the research uses the PPF algorithm. The PPF algorithm is a technology used for 3D object recognition and 6D pose estimation, which can estimate the pose of objects through clustering and other methods [15]. The core calculation of PPF is shown in Equation (3).

$$PPF\left(p_i, p_j, q_i, q_j\right) = \left[d\left(p_i, p_j\right), \angle\left(n_i, \overrightarrow{p_j p_i}\right), \angle\left(n_j, \overrightarrow{p_i p_j}\right), \angle\left(n_i, n_j\right)\right] \tag{3}$$

In Equation (3), p_i and p_j are any two points in the image feature point cloud. q_i and q_j are corresponding candidate points in the feature point cloud of the preset target model. $d\left(p_i, p_j\right)$ represents the Euclidean distance between two points. n_i and n_j are the surface normal vectors of two points respectively. $\overrightarrow{p_j p_i}$ and $\overrightarrow{p_i p_j}$ are direction vectors between two points. $\angle$ represents the angle, such as $\angle\left(n_i, n_j\right)$ represents the angle between the normal vectors of the surfaces of two points. After point pair matching, pose regression can be performed. The process from point pair matching module to pose regression is shown in Figure 5.

In Figure 5, the process of point pair matching module to pose regression involves feature similarity measurement, obtaining the initial matching point pair set, initial pose solution and pose

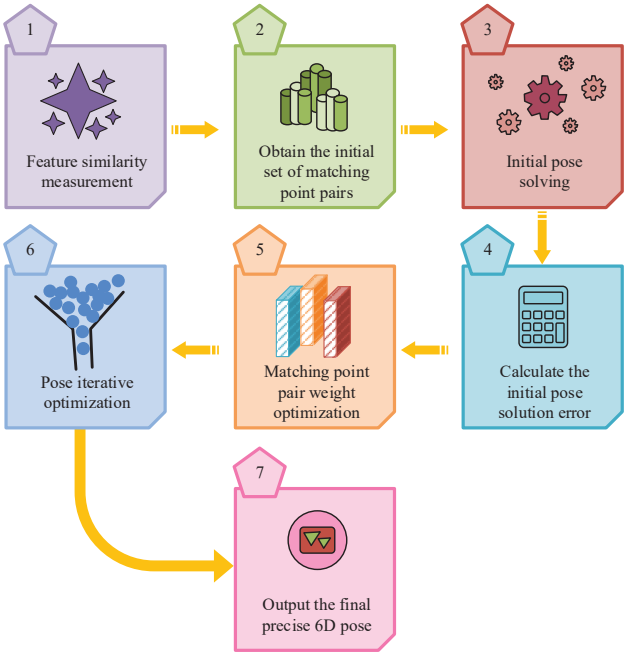


Figure 5: The process of pose regression using point-to-point matching module.

iterative optimization, etc. Through pose regression, this research can convert the feature point correspondence established by the point pair matching module into accurate 6D pose parameters. This provides core technical support for real-time spatial positioning and virtual-real fusion mapping of user body movements in VR natural interaction scenes. The structure of the 6D pose estimation method (TFEN+SO(3)+PPF) that integrates three major modules is shown in Figure 6.

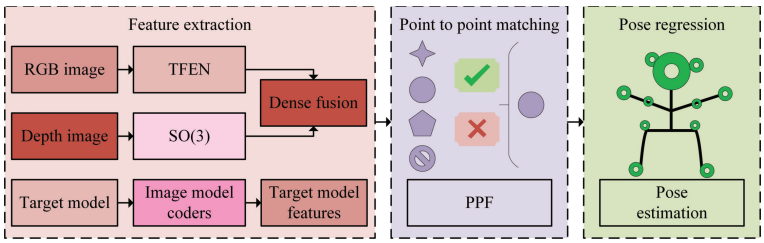


Figure 6: The structure of TFEN+SO(3)+PPF.

In Figure 6, the TFEN+SO(3)+PPF method uses RGB images and depth images as dual input sources. It first uses TFEN to extract the two-dimensional texture features of the RGB image, and then uses the SO(3) encoder to extract the 3D geometric features. Meanwhile, the two-way complementarity of texture and spatial information is achieved through feature fusion. Subsequently, the study uses the PPF point pair matching module to establish an accurate correspondence between the feature points of the image and the target model. Finally, accurate 6D pose parameters are output through pose regression. Through the 6D pose estimation method, this study can accurately capture the spatial position and posture of VR users' limbs, and provide high-quality pose data support for natural VR interaction to improve the immersion and accuracy of VR interaction.

4 Results and discussion

4.1 Experimental environment and parameter design

The operating system used in the experiment is Ubuntu20.04LTS, the central processor is Intel Core i9-12900K, and the graphics processor is NVIDIA GeForce RTX3090 (24GB video memory). These ensure efficient large-scale data processing and model inference. Based on the Human3.6M

human action dataset, Occluded LINEMOD occlusion dataset, and self-built multi-user interaction scenarios, the evaluation is ensured to be comprehensive and practical. The core parameters of the algorithm include: TFEN uses ResNet-18 as the backbone network, the input RGB image resolution is 640×480 pixels, and the depth map is synchronously sampled to this resolution. The SO(3) encoder uses Lie group Lie algebra theory to implement 3D rotation space mapping, with a coding dimension of 128 dimensions. The PPF threshold is set to 0.95, and the number of matching point pairs is no less than 50. The pose regression uses the iterative closest point algorithm, with a maximum number of iterations of 50 and a convergence threshold of 0.001. Multi-user scenario parameter settings: User distance is 0.5m-3.0m, action complexity is divided into three levels, and synchronization evaluation uses time offset measurement (the range of ±100ms is considered synchronized).

The core indicators of performance evaluation include: Rotation error (RE) is defined as the angular deviation between the estimated rotation matrix and the true rotation matrix. Translation error (TE) represents the Euclidean distance between the estimated translation vector and the true translation vector. Success rate (SR) refers to the proportion of tests in which the pose estimation error is within the acceptable threshold. The threshold is set to RE<5° and TE<5cm. Cross-interference rate (CR) evaluates the degree of mutual influence between users in a multi-user scenario. Frames per second (FPS) reflects the system's real-time processing capabilities. Processing latency (PL) includes three stages: feature extraction delay, matching delay and regression delay. The network core architecture parameters are shown in Table 1.

Table 1: Network core architecture parameters

Module	Backbone network	Attention module	Feature dimension	Calculate cost (per frame)
TFEN (RGB feature extraction)	ResNet-18	Efficient channel attention	256 dimensions	2.13ms
SO(3) encoder (deep feature extraction)	VN-Linear+MLP	Nothing	128 dimensions	1.68ms
Cross-modal fusion	/	Bidirectional cross attention	384 dimensions	0.52ms
PPF matching+pose regression	/	Nothing	/	0.48ms
Overall end-to-end	/	/	/	4.81ms

From Table 1, the computational overhead of each module in this method is balanced, and the overall end-to-end inference time is only 4.81ms, which can meet the real-time requirements. Among them, the TFEN module uses ResNet-18 and efficient channel attention to extract RGB features. The SO(3) encoder encodes deep features with a lightweight structure, without additional attention overhead. The cross modal fusion and pose regression modules both have a latency of less than 0.6ms. Overall, while ensuring feature expression and modal fusion capabilities, they achieve extremely low inference latency.

4.2 Training and implementation details

The key details of model implementation are as follows:

Input normalization: RGB images are normalized using ImageNet statistical mean and variance normalization, and depth maps are normalized to the 0-1 interval and invalid depth values are removed. Depth map to point cloud conversion: Based on camera interpolation, depth pixels are back projected into 3D space to generate unordered point clouds, with a fixed sampling of 8192 point clouds. Sampling strategy: Random uniform sampling is used to ensure balanced distribution of point clouds and avoid local feature loss. Fusion weight α: Adopting an adaptive learning strategy, α is dynamically updated during training, with an initial value of 0.5 and a convergence range of 0.4~0.6. Training loss: The pose regression loss adopts a weighted fusion of rotation smoothing L1 loss and translation L2 loss, and the corresponding loss adopts feature matching cross entropy loss. Training strategy: A phased training strategy is adopted, firstly pre-training the feature extraction module, and then jointly training the

fusion and regression modules end-to-end. In addition, the key parameters for model training and implementation are shown in Table 2.

Table 2: Key parameters for model training and implementation

Parameter category	Specific settings
Input resolution	RGB/depth map: 640×480
Number of point cloud samples	Eight thousand one hundred and ninety-two
Fusion weight	Adaptive learning, initial 0.5
PPF threshold	0.95, number of matching points≥50
ICP iteration setting	Maximum 50 iterations, convergence threshold 0.001
Optimizer	AdamW
Learning rate	Initial 1e-4, cosine annealing scheduling
Batch size	Sixteen
Feature dimension	256 dimensions (ResNet-18 final layer output)
Feature dimension	128 dimensional (SO (3) encoder output)
Feature dimension	384 dimensions (reduced after splicing)
Training round	100 rounds
Loss weight	Rotation loss: translation loss: corresponding loss=1:1:0.5

The 6D pose estimation target of this study is the rigid body segments of key upper/lower limbs of VR users, including 3D position and 3D rotation, used for limb motion tracking in VR natural interaction. The dataset and pose target description are shown in Table 3.

Table 3: Dataset and pose target description

Dataset	Positioning target	True value acquisition method	Rigid body composition
Human3.6M	3D human joints and limb segments	Optical motion capture system	Rigid body segments of human upper limbs, lower limbs, and trunk
Occluded LINEMOD	Rigid object	Annotation tool+3D scanning	Standard rigid object (without human limbs)
Self built multi-user VR interactive dataset	VR user limb rigid body segment	VR high-precision positioning tracker+optical marker	Rigid body segments of the user's upper limbs, forearms, upper arms, thighs, and lower legs

This study adopts two types of delay indicators. End-to-end inference delay refers to the pure computational delay (excluding data preprocessing) of the model from input RGB-D data to output 6D pose, which is divided into four stages: TFEN feature extraction, SO(3) encoder, cross-modal fusion, PPF matching, and pose regression. The system processing time refers to the entire process time including image acquisition, denoising, alignment, invalid value removal, and model inference preprocessing. All experiments have clear hardware configurations: lighting robustness experiments, occlusion experiments, multi-user experiments, and hardware adaptation experiments all use flagship hardware (Intel Core i9-12900K+NVIDIA GeForce RTX 3090).

4.3 Core performance and anti-interference capability verification

To verify the contribution of the SO(3) encoder in 6D pose estimation, this study compares the performance differences between models that contain and do not contain the SO(3) encoder. Using a cross-validation method, each model is subjected to 5 independent training-test cycles. It records RE and TE, calculates the mean and standard deviation, and evaluates the contribution of the SO(3) encoder to the robustness of pose estimation. Figure 7 shows the results of the ablation experiment.

Figure 7(a) shows that the TFEN+SO(3) encoder+PPF model has the most concentrated error distribution, with a median of 0.42° and an interval of 0.35°-0.50°. The distribution of the model without SO(3) becomes wider, with a median of 1.28°, and an interval of 0.9°-1.6°. The model with only SO(3) is in the middle, with a median of 0.89°. Figure 7(b) shows that the median of the research

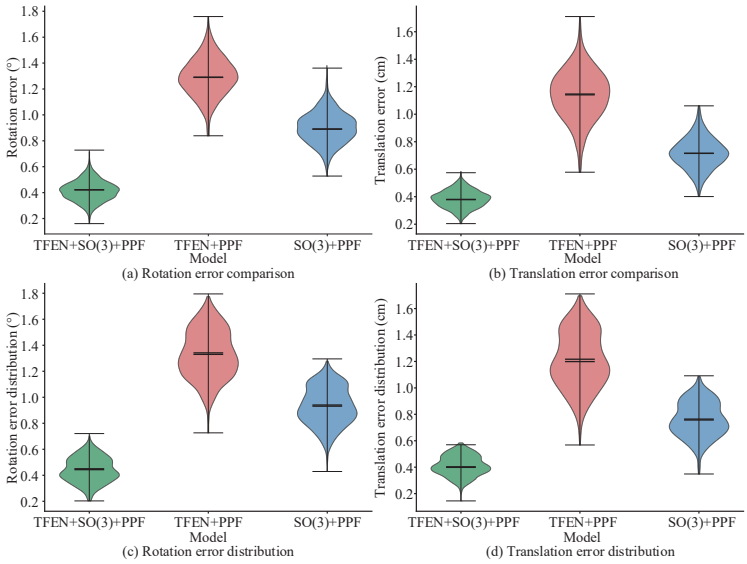


Figure 7: Ablation experiment results.

model is 0.38cm, with an interval of 0.30cm-0.45cm. The median value of the model without SO(3) is 1.15cm, with an interval of 0.8cm-1.5cm. Only the SO(3) model has a median value of 0.72cm and a scattered distribution. Figure 7(c) shows that the research model has a normal distribution with a 95% error of 0.25°-0.60°. There is no SO(3) model with right skew, and the 15% error exceeds 1.5°. Only the SO(3) model is symmetric but discrete. Figure 7(d) shows that the 90% error of the research model is 0.25cm-0.55cm. The model without SO(3) has a double peak, and only the SO(3) model has a heavier tail. In order to evaluate the robustness of the 6D pose estimation method under different lighting conditions, the experiment set up 5 lighting conditions in a controllable lighting environment, collected user action data, and analyzed the stability of feature extraction. Then this study conducts 100 pose estimation tests under each lighting condition, and records the feature extraction SR and pose estimation error. Figure 8 shows the performance under different lighting conditions.

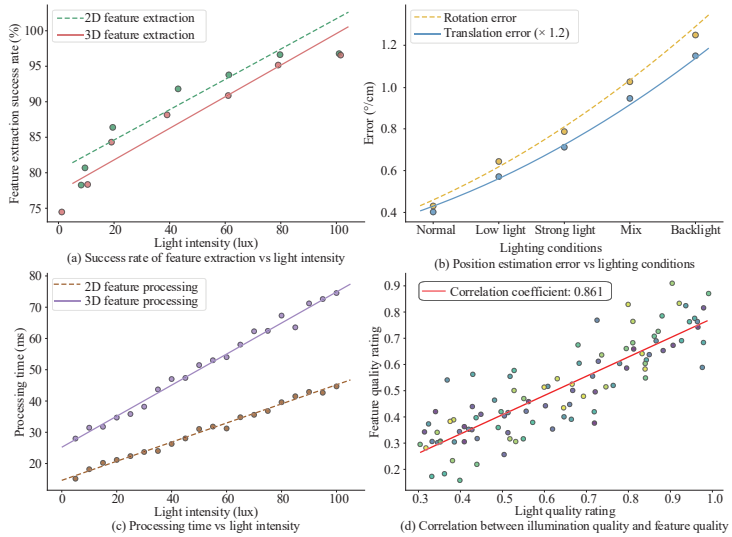


Figure 8: Performance under different lighting conditions.

Figure 8(a) shows that when the normal illumination is 600-1000lux, the SR is stable at 95%-98% and the data points are dense. When the light is below 200lux, the SR drops significantly, with a minimum of 72% and scattered points. When the light intensity exceeds 1200lux, the SR slightly

decreases to 88%~92%, which may be due to overexposure of the image causing distortion of texture information, thereby affecting the accuracy of feature extraction. The processing time under strong light has been shortened to 22-28ms, possibly due to the reduction of the algorithm's feature search space caused by changes in lighting conditions. Figure 8(b) shows that under normal lighting, RE is 0.4°-0.5° and TE is 0.35cm-0.45cm, and the data is concentrated. Under dark light, RE is 0.8°-1.4°, TE is 0.7cm-1.2cm, and the data shifts to the high error area. The error is uniform under mixed lighting, RE0.6°-0.8°, TE 0.5cm-0.7cm. In Figure 8(c), normal lighting processing time is 25ms-30ms, data is concentrated. Under low light, preprocessing is required and the time is extended to 35-42ms. The processing time under strong light is shortened by 22ms-28ms, or the algorithm search space is reduced due to lighting conditions. In Figure 8(d), there is a strong positive correlation between lighting quality and feature extraction quality, with a correlation coefficient of 0.87. When the lighting quality score is greater than 0.8, the feature quality score is 0.85-0.95. When 0.5-0.8, it corresponds to 0.6-0.8. When it is less than 0.5, it drops to 0.4-0.6, with scattered points and increased performance uncertainty under extreme lighting. To analyze the influence of occlusion degree on the accuracy of 6D pose estimation, the experiment simulates different occlusion degrees through artificial occlusion, and tests the pose estimation accuracy on the Occluded LINEMOD dataset. 200 tests are conducted at each level of occlusion, with RE and TE recorded. Figure 9 shows the relationship between occlusion degree and pose estimation error.

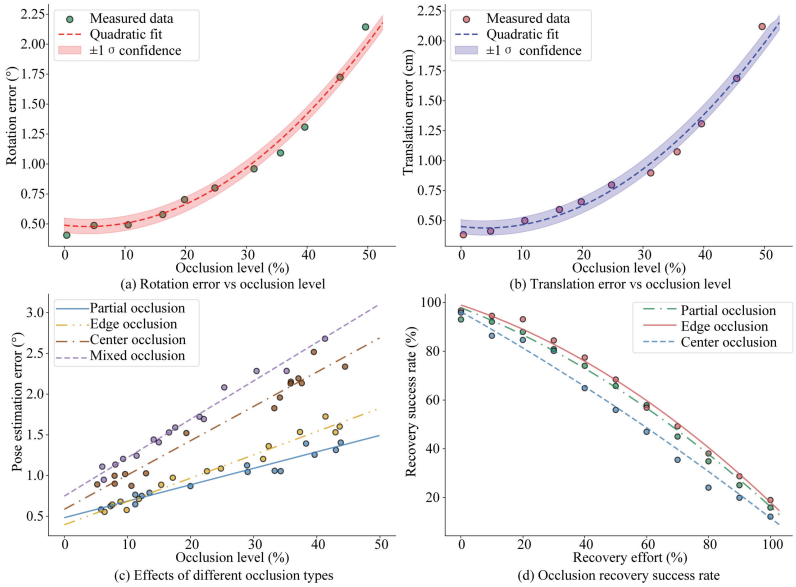


Figure 9: Relationship between occlusion level and pose estimation error.

In Figure 9(a), when there is no occlusion, RE0.35°-0.50°, the occlusion error within 10% is 0.4°-0.7°, and the occlusion error within 20%-30% is 0.6°-1.0°. The occlusion error increases exponentially above 40%, and is 1.2°-2.0° at 50% occlusion. In Figure 9(b), the error without occlusion is 0.30cm-0.45cm, 10% occlusion is 0.4cm-0.6cm, 20%-30% occlusion is 0.5cm-0.9cm, and 50% occlusion is 0.8cm-1.4cm. The scatter plot of different occlusion types in Figure 9(c) shows that edge occlusion has the smallest impact. RE is 0.4°-0.7°, TE is 0.35cm-0.6cm, center occlusion has the greatest impact. RE is 1.5°-2.2°, TE is 1.2cm-1.8cm, and the effects of partial occlusion and mixed occlusion are centered. The occlusion recovery SR scatter plot in Figure 9(d) shows that the two are negatively correlated, with a correlation coefficient of -0.79, and the pose estimation error increases in a stepwise manner as the degree of occlusion increases. The error index increases when more than 40% is blocked. When 50% blocked, RE reaches 1.2°-2.0° and TE is 0.8cm-1.4cm. At this time, center occlusion has the greatest impact and edge occlusion has the smallest impact. There is a strong negative correlation between the degree of occlusion and the recovered SR, and the SR is nearly 99% when there is no occlusion.

For the 6D pose estimation task, the RGB-D fusion method Bidirectional Deep Residual Fusion Network (BDR6D), the point cloud 6D pose estimation method Frequency Guided CNN-Transformer

Fusion Network (FGCT6D), and the pose tracking method Class-level 6D pose tracking (using soft correspondence matrix estimation Corr-Track as the baseline) are selected for comparative testing on flagship hardware configuration (i9-12900K+RTX3090). The results are detailed in Table 4.

Table 4: Comparison results with strong baseline model

Model	RE/°	TE/cm	SR%	FPS	End-to-end latency/ms
BDR6D	1.26	1.18	89.2	112.6	9.24
FGCT6D	1.03	0.97	91.5	98.3	10.12
Corr-Track	0.87	0.76	93.8	106.7	8.76
Proposed method (TFEN+SO (3)+PPF)	0.42	0.38	96.5	208.3	4.81

From Table 4, the proposed TFEN+SO(3)+PPF is significantly better than the selected baseline model in all core indicators, fully verifying its dual advantages in accuracy and real-time performance. In terms of accuracy, the RE of our method is only 0.42°, and the TE is as low as 0.38cm, which is 51.72% and 50.0% lower than the optimal baseline Corr Track, respectively. At the same time, the SR is improved to 96.5%, indicating that the SO(3) equivariant encoder and cross-modal fusion strategy effectively enhance feature robustness and significantly reduce pose estimation bias. In terms of real-time performance, the proposed method achieves a frame rate of 208.3FPS and an end-to-end latency of only 4.81ms, which is 1.95 times and 0.55 times higher than Corr Track, respectively. This is far superior to BDR6D and FGCT6D, indicating that TFEN's lightweight structure and efficient computing design can meet the millisecond level interaction requirements in VR scenes. Overall, this method achieves a better balance between accuracy and real-time performance, breaking through the accuracy bottleneck of existing RGB-D fusion and point cloud methods, while maintaining leading inference efficiency and better adapting to the demanding requirements of multi-user VR interaction scenarios.

4.4 Performance evaluation of actual application scenarios

To evaluate the performance of the 6D pose estimation method in multi-user collaborative interaction scenarios, the VR interaction area is equipped with a VR large screen of specified specifications. A 1.0m-4.0m 3D test space is constructed in front, divided into grids with a step size of 0.2m. The multi-user scenario uses 2-6 subjects, the action complexity is divided into three levels, and the synchronization is measured by the time offset (±100ms threshold). The hardware adaptation test covers 5 types of configurations, covering flagship/professional/mid-range levels. Each scenario ensures consistency in assessment through a standardized data collection process. The control variable method is used to test the impact of user number, spacing, and action synchronization on system performance, and record CR, pose estimation error, and PL. Figure 10 shows the performance analysis results in a multi-user scenario.

Figure 10 shows that CR increases linearly as the number of users increases, and decays exponentially as user distance increases. After the user distance exceeds 2.0m, the interference rate drops below 5%. The pose estimation error also increases as the number of users increases, but the growth trend is relatively flat. When there are 6 users, RE increases by approximately 50% and TE increases by approximately 45%. Comprehensive performance evaluation shows that the optimal working area is 2-4 users with a spacing of 1.5m-2.5m. The system can maintain good performance in this area. To analyze the coverage effect of RGB cameras and depth cameras at different distances, the experiment is carried out in the distance range of 1.0m-4.0m, with a step length of 0.2m, to test the pose estimation accuracy at different positions. This evaluates the effectiveness of camera coverage. A grid test method is used to establish a 3D test space in front of the VR large screen, and the pose estimation SR, accuracy, and delay of each position are recorded. Moreover, heat maps are used to visualize the differences in coverage effects in different areas. Figure 11 shows the analysis results of camera coverage effect.

Figure 11 shows that the RGB camera has the best coverage effect within the distance range of 1.5-2.5m, with SR>90%. The depth camera performs well in the 1.0-3.0m range, with SR>85%. The fused system achieves excellent performance within the range of 1.5-3.0m distance and ±45° an-

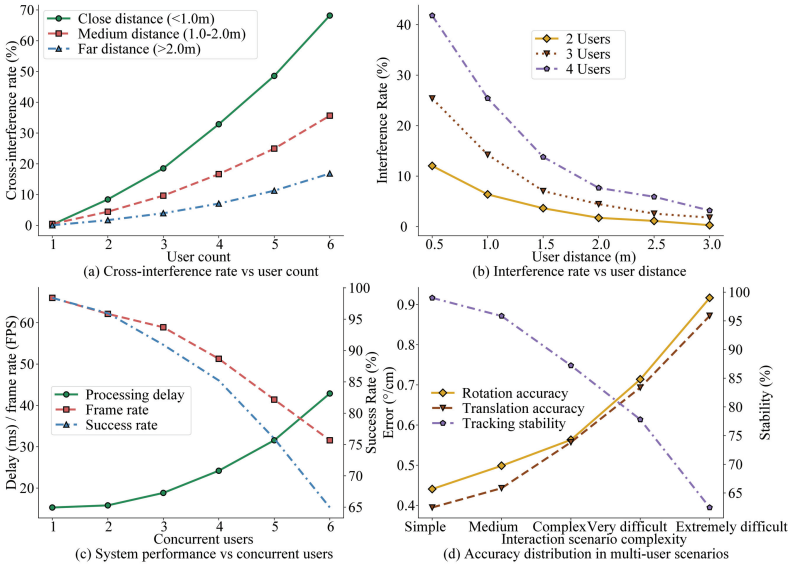


Figure 10: Performance analysis results in multi-user scenarios.

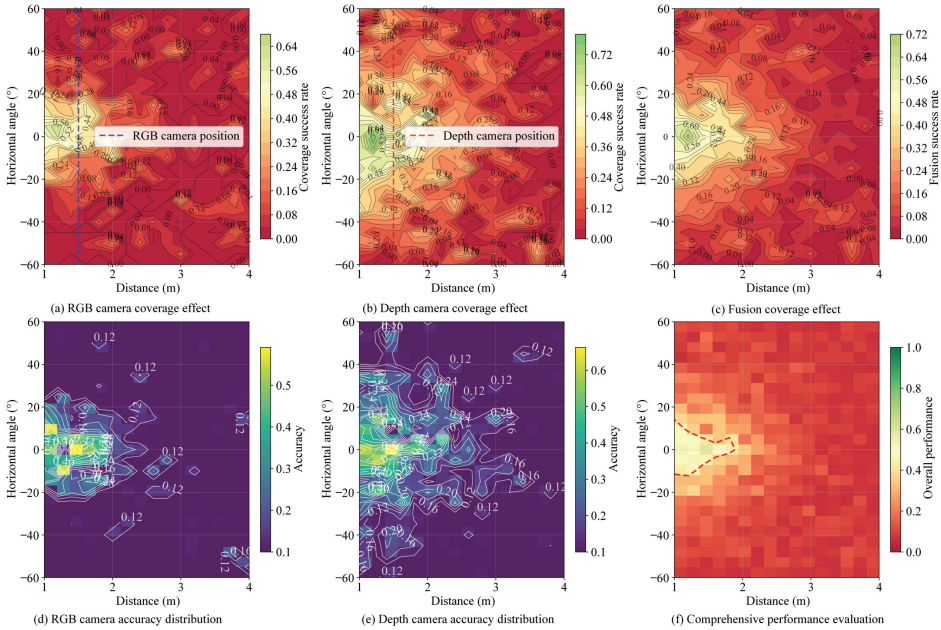


Figure 11: Camera coverage effect analysis results.

gle, with a comprehensive score >0.8. Accuracy distribution analysis shows that both cameras can maintain high accuracy in the central area, with a distance of 1.5-2.5m and an angle of ±30°. The comprehensive performance evaluation divides the space into excellent areas, good areas, and acceptable areas, verifying the effectiveness of the deployment plan. To evaluate the real-time performance of the 6D pose estimation method under different hardware configurations, five different hardware configurations are experimentally tested. Using the benchmark test method, 1,000 tests are performed under each configuration to record the PL and memory usage at each stage. Box plots are used to display the performance distribution characteristics under different hardware configurations to evaluate the hardware adaptability of the method. The results are shown in Figure 12.

Figure 12 shows that the flagship configuration performs best, with a total latency of 4.81±0.16ms and an FPS of 208.3±7.1FPS. This meets VR real-time requirements (>30FPS) and is well below the human eye perception threshold. The professional-grade configuration also provides good performance, with a total latency of 7.99±0.42ms and an FPS of 125.5±6.7FPS. Although the mid-range configuration has higher latency, with a total latency of 22.42±1.98ms, it can still reach 45.0±4.0FPS.

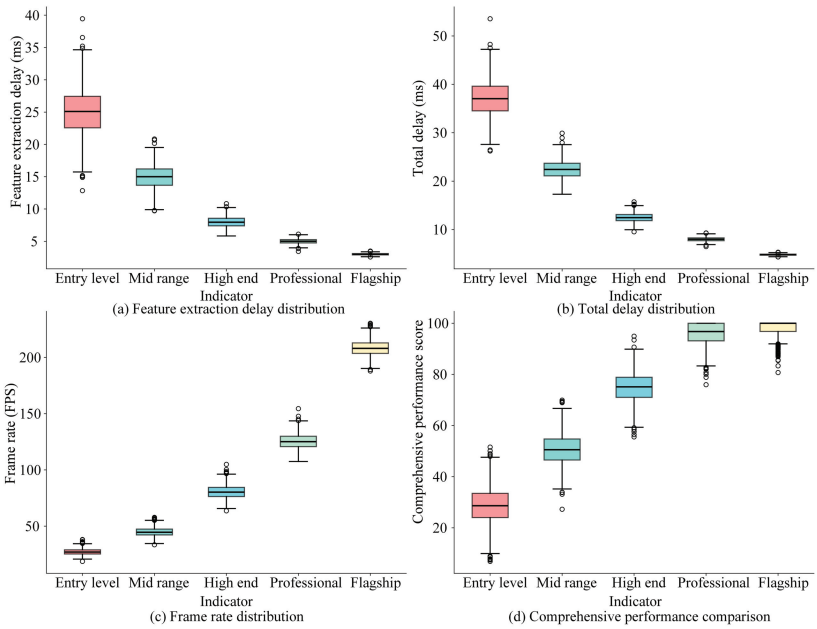


Figure 12: Hardware performance analysis results.

This meets basic VR application needs. Latency analysis at each stage shows that feature extraction is the main time-consuming link, accounting for 60%-70% of the total delay. The box plot distribution of all hardware configurations shows that the FPS distribution of high-end and above configurations is more concentrated and the performance is more stable.

5 Conclusion

This study develops a real-time dual-modal 6D pose estimation framework for VR interaction, addressing the key challenges of accuracy, latency, and robustness in dynamic and multi-user environments. By integrating RGB and depth sensing within a deep learning architecture, the proposed method effectively combines two-dimensional texture information with 3D geometric constraints.

The introduction of an SO(3)-equivariant encoder enables rotation-consistent geometric feature representation, significantly improving pose estimation accuracy and stability. Experimental results demonstrate that the proposed framework achieves sub-degree rotation accuracy and sub-centimeter TR while maintaining high frame rates and low end-to-end latency. The system further exhibits strong resistance to lighting variation, occlusion, and cross-user interference, making it suitable for real-time interactive applications.

From a computational and control perspective, the proposed approach provides a scalable solution for accurate pose tracking in complex environments, with clear advantages in real-time performance and robustness. Although performance degradation is observed under extreme illumination and heavy occlusion, these limitations primarily arise from feature extraction overhead and sensor constraints.

Future work will focus on optimizing the computational efficiency of the feature extraction stage through lightweight network architectures, as well as enhancing robustness in extreme environments by incorporating infrared sensing and occlusion recovery mechanisms. These extensions are expected to further improve the adaptability of the framework for large-scale multi-user interaction and real-time control applications in virtual and mixed reality systems.

Funding

Supported by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2025C17).

Author contributions

Rong Hu conceived the study, developed the methodology, designed and supervised the experimental framework, analyzed the results, and wrote the original draft. Runlin Li contributed to the software implementation and assisted with literature organization and manuscript formatting. Ping Huang supervised the research, refined the analytical framework, validated the findings, contributed to the interpretation of results, and reviewed and edited the manuscript. All authors read and approved the final version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Andonie, R.; Muşat, B. (2024). Signs and Supersigns in Deep Learning, *International Journal of Computers Communications & Control*, 19(1), 6392, 2024. DOI: 10.15837/ijccc.2024.1.6392.

[2] Cao, X.; Li, J.; Zhao, P.P.; Li, J.C.; Qin, X.Y. (2024). Corr-Track: Category-Level 6D Pose Tracking with Soft-Correspondence Matrix Estimation, *IEEE Transactions on Visualization and Computer Graphics*, 30(5), 2173–2183, 2024. DOI: 10.1109/TVCG.2024.3372111.

[3] Cao, T.; Zhang, W.X.; Fu, Y.P.; Zheng, S.J.; Luo, F.; Xiao, C.X. (2024). DGECN++: A Depth-Guided Edge Convolutional Network for End-to-End 6D Pose Estimation via Attention Mechanism, *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6), 4214–4228, 2024. DOI: 10.1109/TCSVT.2023.3327430.

[4] Chen, X.X.; Yang, F.; Duan, S.S.; Gao, X.J.; Ge, B.; Li, K.C.; Fang, X.J.; Zhang, Y.; Ke, Y. (2024). ECW-EGNet: Exploring Cross-Modal Weighting and Edge-Guided Decoder Network for RGB-D Salient Object Detection, *Computer Science and Information Systems*, 21(3), 947–969, 2024. DOI: 10.2298/CSIS231206022X.

[5] Chen, Z.; Yan, L.; Wang, C.; Ye, Z. (2024). Improved Discriminator Network Based on Residual Network for Person Image Synthesis, *Tehnicki vjesnik–Technical Gazette*, 31(6), 1815–1825, 2024. DOI: 10.17559/TV-20240416001467.

[6] Dang, Z.; Wang, L.Z.; Guo, Y.; Salzmann, M. (2024). Match Normalization: Learning-Based Point Cloud Registration for 6D Object Pose Estimation in the Real World, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6), 4489–4503, 2024. DOI: 10.1109/TPAMI.2024.3355198.

[7] Feng, G.K.; Xu, T.B.; Liu, F.L.; Liu, M.K.; Wei, Z.Z. (2024). NVR-Net: Normal Vector Guided Regression Network for Disentangled 6D Pose Estimation, *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2), 1098–1113, 2024. DOI: 10.1109/TCSVT.2023.3290617.

[8] Gupta, H.; Thalhammer, S.; Weibel, J.B.; Haberl, A.; Vincze, M. (2024). ReFlow6D: Refraction-Guided Transparent Object 6D Pose Estimation via Intermediate Representation Learning, *IEEE Robotics and Automation Letters*, 9(11), 9438–9445, 2024. DOI: 10.1109/LRA.2024.3455897.

[9] Hourri, S. (2024). Empowering Speaker Verification with Deep Convolutional Neural Network Vectors, *Studies in Informatics and Control*, 33(2), 97–107, 2024. DOI: 10.24846/v33i2y202409.

[10] Li, M.; Zhang, Z.; Lu, M.; Jia, X.; Liu, R.; Zhou, X.; Zhang, Y. (2023). Internet Financial Credit Risk Assessment with Sliding Window and Attention Mechanism LSTM Model, *Tehnicki vjesnik–Technical Gazette*, 30(1), 1–7, 2023. DOI: 10.17559/TV-20221110173532.

[11] Lin, S.F.; Wang, Z.R.; Zhang, S.H.; Ling, Y.G.; Yang, C.G. (2024). Deep Fusion for Multi-Modal 6D Pose Estimation, *IEEE Transactions on Automation Science and Engineering*, 21(4), 6540–6549, 2024. DOI: 10.1109/TASE.2023.3327772.

[12] Liu, J.; Sun, W.; Yang, H.; Liu, C.P.; Zhang, X.; Mian, A. (2024). Domain-Generalized Robotic Picking via Contrastive Learning-Based 6-D Pose Estimation, *IEEE Transactions on Industrial Informatics*, 20(6), 8650–8661, 2024. DOI: 10.1109/TII.2024.3366248.

[13] Liu, J.; Zhou, G. (2024). Learning Discriminative Representations through an Attention Mechanism for Image-based Person Re-identification, *Computer Science and Information Systems*, 21(4), 1483–1498, 2024. DOI: 10.2298/CSIS230829044L.

[14] Liu, P.L.; Zhang, Q.S.; Cheng, J. (2024). BDR6D: Bidirectional Deep Residual Fusion Network for 6D Pose Estimation, *IEEE Transactions on Automation Science and Engineering*, 21(2), 1793–1804, 2024. DOI: 10.1109/TASE.2023.3248843.

[15] Malega, P.; Daneshjo, N. (2024). Increasing the Production Capacity of Business Processes Using Plant Simulation, *International Journal of Simulation Modelling*, 23(1), 41–52, 2024. DOI: 10.2507/IJSIMM23-1-669.

[16] Motiejauskas, M.; Dzemyda, G. (2024). EfficientNet Convolutional Neural Network with Gram-Matrices Modules for Predicting Sadness Emotion, *International Journal of Computers Communications & Control*, 19(5), 6697, 2024. DOI: 10.15837/ijccc.2024.5.6697.

[17] Shen, S.Q.; Wymeersch, H. (2024). A Projective Geometric View for 6D Pose Estimation in mmWave MIMO Systems, *IEEE Transactions on Wireless Communications*, 23(8), 9144–9159, 2024. DOI: 10.1109/TWC.2024.3359253.

[18] Shi, D.; Li, L.; Shao, Y.X.; Zhang, W.X.; Ding, X.L. (2024). Multimode Control Strategy for Robotic Rehabilitation on Special Orthogonal Group SO(3), *IEEE Transactions on Industrial Electronics*, 71(2), 1749–1757, 2024. DOI: 10.1109/TIE.2023.3257378.

[19] Sun, H.; Zhou, Z.N.; Wang, Y.Z.; Zhang, Z.Z.; Cao, Q.X. (2024). FGCT6D: Frequency-Guided CNN-Transformer Fusion Network for Metal Parts' Robust 6D Pose Estimation, *IEEE Robotics and Automation Letters*, 9(5), 4385–4392, 2024. DOI: 10.1109/LRA.2024.3381016.

[20] Tang, J.Y.; Chen, Z.Y.; Fu, B.W.; Lu, W.J.; Li, S.Q.; Li, X.; Ji, X.Y. (2024). ROV6D: 6D Pose Estimation Benchmark Dataset for Underwater Remotely Operated Vehicles, *IEEE Robotics and Automation Letters*, 9(1), 65–72, 2024. DOI: 10.1109/LRA.2023.3331624.

[21] Wang, G.; Manhardt, F.; Liu, X.Y.; Ji, X.Y.; Tombari, F. (2024). Occlusion-Aware Self-Supervised Monocular 6D Object Pose Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3), 1788–1803, 2024. DOI: 10.1109/TPAMI.2021.3136301.

[22] Xing, H.R. (2024). Optimizing Human-Machine Systems in Automated Environments, *International Journal of Simulation Modelling*, 23(4), 716–727, 2024. DOI: 10.2507/IJSIMM23-4-CO19.

[23] Yang, L.B.; Nguyen-Thoi, T.; Tran, T.T. (2024). Predicting the friction angle of clays using a multi-layer perceptron neural network enhanced by Yeo-Johnson transformation and coral reefs optimization, *Journal of Rock Mechanics and Geotechnical Engineering*, 16(10), 3982–4003, 2024. DOI: 10.1016/j.jrmge.2024.02.025.

[24] Yu, S.; Zhai, D.H.; Xia, Y.Q.; Wang, W.; Zhang, C.Y.; Zhao, S.Q. (2024). Robotic Grasp Detection with 6-D Pose Estimation Based on Graph Convolution and Refinement, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(6), 3783–3795, 2024. DOI: 10.1109/TSMC.2024.3371580.

[25] Zhang, W. (2024). Investigation on the Use of Virtual Reality in the Teaching of Engineering Education Based on Functional Linked Neural Network, *Studies in Informatics and Control*, 33(3), 103–113, 2024. DOI: 10.24846/v33i3y202410.



Copyright ©2026 by the authors. Licensee Agora University, Oradea, Romania.
This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.
Journal's webpage: http://univagora.ro/jour/index.php/ijccc/



This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Hu, R.; Li, R.; Huang, P. (2026). Real-Time Dual-Modal 6D Pose Estimation for Virtual Reality Interaction Using Deep Learning, *International Journal of Computers Communications & Control*, 21(6), 7424, 2026.

https://doi.org/10.15837/ijccc.2026.6.7424